Article

Not peer-reviewed version

Beyond Visual Generation: Frontiers, Challenges, and Future Directions in Agentic Visual Creation

Hefei Mei [†], Hangzhou He [†], Haiyi Qiu [†], Longrong Yang [†], Lunhao Duan [†], Shanshan Zhao [*], Pengxin Zhan, Qing-Guo Chen, Zhao Xu, Weihua Luo

Posted Date: 9 September 2026

doi: 10.20944/preprints202609.0682.v1

Keywords: agentic visual creation; multimodal agents; visual generation; visual editing; visual composition; visual programming

 doi:10.20944/preprints202609.0682.v1

Article

Beyond Visual Generation: Frontiers, Challenges, and Future Directions in Agentic Visual Creation

Hefei Mei [1,2,†], Hangzhou He [1,3,†], Haiyi Qiu [1,4,†], Longrong Yang [1,†], Lunhao Duan [1,†], Shanshan Zhao [1,*], Pengxin Zhan [1], Qing-Guo Chen [1], Zhao Xu [1] and Weihua Luo [1]

[1] Alibaba Group, China
[2] City University of Hong Kong, China
[3] Peking University, China
[4] Zhejiang University, China
* Correspondence: zhaoshanshan.zss@alibaba-inc.com
† These authors contributed equally to this work.

Abstract

Recent advances in image and video generation have improved visual quality and controllability, yet multi-stage creative tasks still require planning, coordination, and revision. Agentic visual creation plays a crucial role in bringing visual generation models into real-world creative workflows by using language or multimodal agents to coordinate generative models, editing tools, and creative software. This survey provides a structured overview to help researchers quickly understand the field, its representative approaches, and their connections. We organize existing work into four categories based on the agent's main role in production: visual generation, visual editing, visual composition, and visual programming. Within each category, we compare how agents plan, use tools, maintain memory, incorporate feedback, and collaborate. We also review benchmark design, training data, and evaluation methods, with attention to feedback that supports revision. Finally, we discuss current limitations and future directions toward more reliable and interactive visual creation systems. The references associated with this survey are available on GitHub.

Keywords: agentic visual creation; multimodal agents; visual generation; visual editing; visual composition; visual programming

1. Introduction

Visual content generation has advanced rapidly in recent years. Diffusion models, autoregressive visual generators, and unified multimodal models can now produce high-quality images and videos, support reference-guided creation, and follow editing instructions [1–9]. These advances have made it substantially easier to generate visual content from natural-language instructions. Yet interaction remains largely prompt-centric: users describe a desired image or video, inspect one or more generated candidates, and respond to mismatches by revising the prompt or moving to another tool. This pattern is effective for short, self-contained requests, but once creation becomes a multi-stage production process, the burden of planning and coordination falls largely on users.

That burden becomes clear in practice, where producing a plausible sample is only one part of the task. A creator may first need to develop an ambiguous concept into a sequence of shots while maintaining character identity across scenes. During production, they may have to select and coordinate generation and editing models, repair local artifacts without altering unaffected regions, and synchronize visual events with audio. Depending on the task, producing the final result may involve assembling raw footage into a coherent long-form video or composing a layout-sensitive document such as a poster or slide deck. Supporting such workflows requires planning, memory, tool orchestration, verification, and iterative revision. These workflows also benefit from intermediate representations that preserve the evolving state of the work in a form that can be inspected and revised

 doi:10.20944/preprints202609.0682.v1

throughout production. Figure 1 summarizes the rapid growth of representative works across these four research themes.

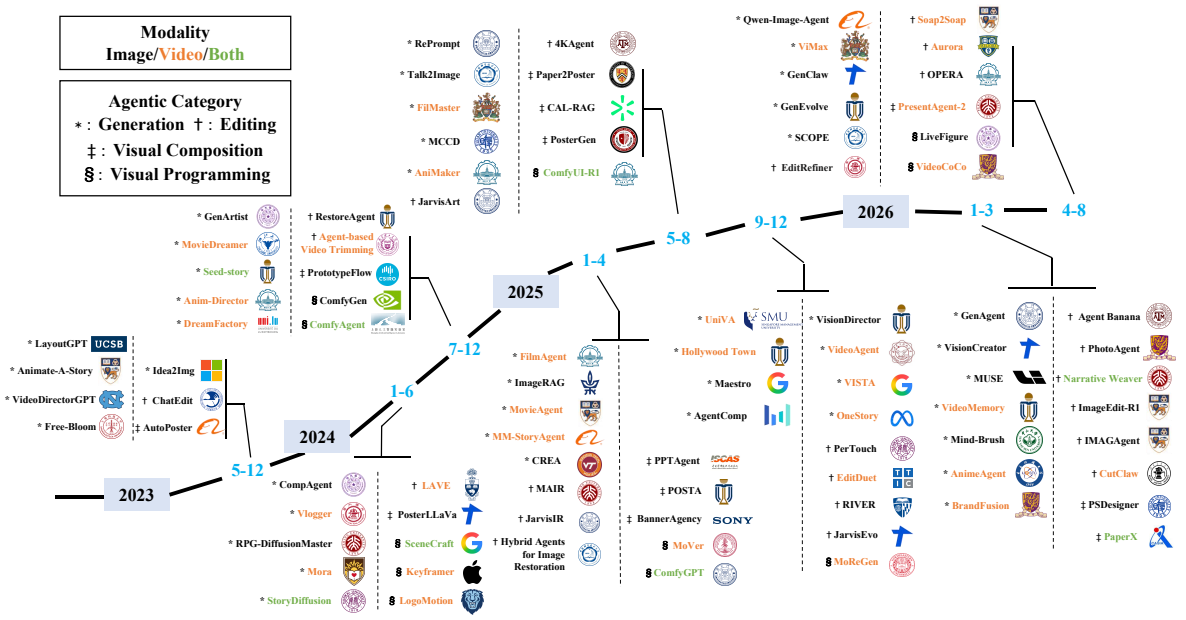

Figure 1. Timeline of Representative Works in Agentic Visual Production. The works are categorized by their release years, from before 2023 to 2026. Works shown in black, orange, and green represent image, video, and both modalities, respectively. The symbols *, †, ‡, § denote the four research themes: Agentic Visual Generation, Agentic Visual Editing, Agentic Visual Composition and Agentic Visual Programming, respectively. The timeline highlights the rapid growth of agentic methods for visual content creation.

Agents built on large language models (LLMs) and multimodal large language models (MLLMs) provide a natural framework for the shift from isolated generation to coordinated production. General-purpose agent systems extend these models beyond response generation by linking reasoning to external action and feedback [10–16]. In visual production, this structure allows an agent to coordinate generative models and creative software within a larger workflow rather than merely produce a richer prompt for another model. It can turn underspecified intent into an actionable plan, track how the work evolves across stages, and use visual feedback to determine what should happen next.

We use the term *agentic visual creation* to describe systems in which one or more agents interpret creative goals and make production decisions through models or creative software. The emphasis is on the decisions assigned to the agent, rather than the presence of an LLM or MLLM alone. Systems vary in their use of feedback, from executing a plan prepared in advance to revising decisions in response to intermediate results. We also review closely related model-based methods that contribute individual capabilities to this process. This scope extends beyond text-to-image and text-to-video generation to visual editing, composition, and programming, including workflows in which agents operate creative software directly.

Despite rapid progress, research on agentic visual creation has developed along largely separate lines. In text-to-image and text-to-video generation, agents typically turn a request into prompt planning, tool routing, and self-refinement [17–22]. Long-form and multi-shot systems place greater emphasis on continuity across the production process, so that prior decisions and reusable assets remain consistent as the work moves from one scene to the next [23–27]. Editing research instead focuses on how an agent interprets an instruction and intervenes selectively in existing media, from correcting a local defect to reorganizing a video over time [28–33]. Another line of work moves away from direct pixel synthesis and asks the agent to produce structured instructions or executable procedures for downstream software [34–39]. Because each community is organized around a different artifact or interface, closely related agentic ideas are often discussed in isolation.

 doi:10.20944/preprints202609.0682.v1

This survey offers a unified, production-oriented view of the field. We organize existing methods according to the dominant responsibility assigned to the agent in the visual production process, covering *visual generation*, *visual editing*, *visual composition*, and *visual programming*. Each category is then organized around the distinction most relevant to its task. For visual generation, that distinction is the scope of agent-side coordination, from the current target to identifiable components linked by shared specifications, assets, or state. In editing, it is whether the intervention is primarily spatial image content or temporally organized video content. Composition separates artifacts contained within one canvas from those that extend across multiple pages, screens, or interactive states, while programming is organized by whether the executable procedure takes the form of a workflow graph or program code. Across these categories, we examine how an agent acts on a plan, keeps track of the evolving work, and uses what it observes to decide what should happen next. We also consider when this process is distributed across specialized agents or mediated by executable intermediate representations. Finally, we review datasets and evaluation protocols that move beyond judging isolated outputs. Particular attention is given to datasets that describe production goals in a structured form and to metrics that capture cinematic continuity and audiovisual coherence across shots. We also consider whether evaluation operates at multiple levels and produces feedback that can guide subsequent revision.

The remainder of this survey is organized as follows. Section 2 introduces the foundations of visual generation and agent systems. Sections 3.1–3.4 review agentic visual creation from four perspectives: visual generation, visual editing, visual composition, and visual programming. We then review data and evaluation protocols for agentic visual creation, followed by open challenges and future directions.

2. Preliminary

2.1. Visual Generation

Diffusion Models.

Diffusion models [1,40] generate visual content by learning to transform noise into data through an iterative denoising process. Compared with direct generation in pixel space, Latent Diffusion Models [2] perform denoising in a compressed latent space, substantially reducing computational cost while preserving high visual fidelity. This latent formulation has become the foundation of many modern text-to-image (T2I) and text-to-video (T2V) systems, and provides the dominant backbone for agentic visual generation considered in this survey.

A key advantage of diffusion models lies in their flexible conditioning mechanisms [41], which make them well suited to agentic visual generation. T2I systems commonly inject representations from pretrained text encoders, such as CLIP [42] or T5 [43], into the denoising network through cross-attention, while spatial conditioning methods such as ControlNet [3] enable fine-grained control using edges, depth maps, poses, and other structural signals. Classifier-free guidance [44] further adjusts the trade-off between prompt fidelity and sample diversity during inference. Meanwhile, diffusion backbones have evolved from convolutional U-Nets to transformer-based architectures [4], providing better scaling abilities [45]. For video generation, temporal attention, 3D convolutions, or related temporal modules are introduced to model cross-frame dependencies [5,6,46–49].

Autoregressive Models.

A common approach to autoregressive visual generation represents images or videos as sequences of discrete visual tokens, often obtained through vector quantization [50]. A transformer then learns to predict each token from the preceding tokens, as in language modeling. Emu3 [51] follows this approach by representing text, images, and videos as discrete token sequences, supporting both multimodal understanding and visual generation through next-token prediction. Another approach is next-scale prediction, introduced by VAR [52]. Rather than generating individual tokens sequentially, VAR generates multi-scale token maps from coarse to fine. At each stage, it predicts all tokens at the

 doi:10.20944/preprints202609.0682.v1

next scale in parallel, conditioned on the previously generated coarser scales. This reduces the number of sequential generation steps compared with token-by-token prediction.

Unified Multimodal Systems.

Unified multimodal models support both visual understanding and generation, with designs based on autoregressive modeling, diffusion, or a combination of the two [53]. Emu3.5 [54] is pretrained through next-token prediction over interleaved text and discrete visual tokens, and introduces discrete diffusion adaptation to accelerate image generation through parallel decoding. Other methods connect multimodal language models to diffusion decoders. Ovis-U1 [55] uses a bidirectional token refiner to prepare multimodal features for a diffusion-based visual decoder, supporting image understanding, generation, and editing. BLIP3-o [56] first generates CLIP image features through flow matching conditioned on language-model outputs, then converts these features into images using a diffusion decoder. BAGEL [9] instead connects understanding and generation experts through shared self-attention, combining autoregressive text prediction with rectified-flow generation of visual latents. Lance [57] similarly uses separate experts over a shared multimodal context, with autoregressive text prediction and flow-based visual generation. Through multi-task training, it supports understanding, generation, and editing for both images and videos.

2.2. Agent

An *agent* is an autonomous system that perceives its environment, reasons about goals, plans actions, and executes them to complete tasks, often through external tools and feedback from the environment. In LLM-based systems, the language model typically serves as the reasoning core, while tools, memory, and environmental feedback extend its ability to act beyond text generation. For agentic visual creation, the design space can be summarized along five dimensions.

Planning and Reasoning.

Agents decompose high-level goals into executable sub-tasks through structured reasoning. ReAct [10] provides a representative pattern in which an LLM alternates reasoning, tool invocation, and observation in a closed loop, enabling the agent to select tools, interpret feedback, and adapt its plan. Other representative techniques include CoT prompting for step-by-step reasoning [58], Tree of Thoughts [59] for multi-path exploration, and explicit task decomposition before execution [60]. In visual creation, planning often translates underspecified user intent into detailed specifications or conditioning signals for generation models.

Tool Use.

Agents extend their capabilities by invoking external tools, including APIs, pretrained models, code interpreters, and retrieval systems. Tool-use behavior can be acquired and improved through training, allowing language models to learn when to invoke external tools, how to incorporate tool outputs, and how to optimize multi-step interactions under feedback [11,61–63]. In visual generation, tool use includes selecting T2I or T2V backends, applying image editing operations, and invoking verification or evaluation modules.

Memory.

Agents preserve context through memory mechanisms [64]. Short-term memory resides in the LLM context window and records the current trajectory, and long-term memory persists across sessions through external stores such as vector databases or structured knowledge bases, which are important for coherent long-horizon behavior [65–67]. In visual creation, memory stores reusable assets, character identities, scene layouts, and production states to remain consistent across generated components.

 doi:10.20944/preprints202609.0682.v1

Feedback and Refinement.

Agents can evaluate outputs and improve them iteratively. Frameworks such as Reflexion [12] and Self-Refine [13] formulate this process as a loop of generation, evaluation, and revision, where the agent observes results, identifies failures, and updates subsequent actions. Specifically, Self-Refine [13] uses the same LLM to generate, critique, and revise an output. Reflexion [12] converts task feedback into verbal reflections stored in episodic memory to guide subsequent attempts. Both improve behavior without updating model parameters. In visual domains, VLMs [68] serve as perceptual interfaces for inspecting generated content. This mechanism supports prompt revision, local editing, sample reranking, and coherence checking in coordinated production.

Multi-Agent Collaboration.

Complex tasks can be assigned to multiple specialized agents, each responsible for a distinct role or sub-goal [14,16]. These agents coordinate through shared states, structured dialogue, or a central orchestrator [15,16,69]. In visual creation, multi-agent systems can separate responsibilities with specialized roles, enabling pipelines that satisfy multiple constraints simultaneously.

We use planning, tool use, memory, feedback, and multi-agent collaboration as recurring analytical dimensions rather than as a fixed agent architecture. In target-centered generation, these dimensions describe decisions around the current target artifact, especially planning, tool routing, and refinement. In coordinated generation, they additionally describe how shared specifications, assets, and state connect identifiable production components. Feedback-driven systems can use these connections for cross-component checking and targeted revision.

3. Agentic Visual Creation

Agentic visual creation assigns production decisions to agents that plan and act through models or software. We organize this literature according to the dominant production operation performed by the agent, rather than the output modality or underlying model architecture. We distinguish four paradigms. *Visual generation* creates new visual content, whereas *visual editing* changes existing media while preserving what should remain unchanged. *Visual composition* arranges content into structured artifacts whose layout is part of the intended result, while *visual programming* expresses creative intent as an executable procedure for downstream software. A system may combine several of these operations, so the boundaries between the four paradigms are not absolute. We therefore classify each system according to the operation that defines its main responsibility.

3.1. Agentic Visual Generation

Agentic visual generation uses agents to coordinate generative models and supporting tools to create new visual content. We distinguish two settings according to how the agent organizes and uses production state. In *target-centered visual generation*, planning, tool use, and refinement focus on the current generation target. In *coordinated visual generation*, the agent manages identifiable components, such as shots or layers, and uses shared specifications, assets, or state to connect their production. The distinction concerns agent-side coordination rather than output complexity or duration. A complex image or a long video can still be target-centered when the agent focuses on generation conditions and candidates rather than managing linked production units. Output coherence or local editing alone does not imply coordinated generation. Some coordinated systems additionally track dependencies to support local revision. Figures 2 and 3 illustrate the two settings, respectively.

 doi:10.20944/preprints202609.0682.v1

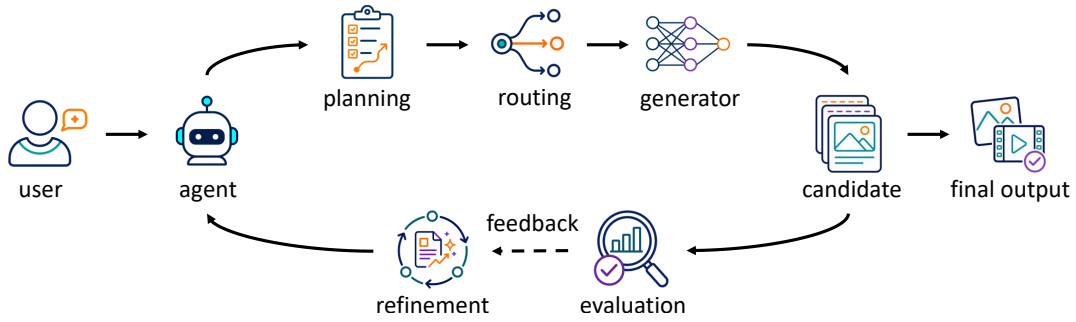

Figure 2. Agentic pipeline for target-centered visual generation. The agent organizes its working state around a current target artifact, translates underspecified user intent into concrete generation inputs, routes the request to suitable models or tools, and iteratively evaluates and refines candidates throughout the generation process.

3.1.1. Target-Centered Visual Generation

The working state of a target-centered agent may include reference information, structured generation controls, and feedback from earlier candidates. The agent uses this information to prepare generation inputs, select suitable models or tools, and refine the current result. We organize these methods around three recurring mechanisms: planning and context construction, tool and model routing, and prompt revision and output refinement, as summarized in Table 1.

Planning and Context Construction.

Agentic planning turns an underspecified creative request into a working representation that a downstream generator can execute. LayoutGPT [17], LMD [18], and LayoutLLM-T2I [70] convert descriptions of objects and relations into explicit spatial layouts, while DivCon [71] makes numerical and positional constraints available to layout-conditioned generation. Richer plans describe the target through region-level prompts, semantic panels, or object-wise generation sequences [72–74]. SCOPE [75] maintains an evolving specification of entities, constraints, and unknowns, using verification failures to route further retrieval, reasoning, or visual repair. MCCD [76] uses role-specialized MLLM agents to extract scene elements and bounding boxes for compositional diffusion, while 3D Space as a Scratchpad [77] externalizes placement, orientation, and viewpoint decisions into an editable 3D scene before image synthesis. For video, VideoDirectorGPT [19] and VideoStudio [78] organize prompts into scene scripts, while Vlogger [79] plans at the level of individual actors. GenClaw [80] moves toward visual programming by using executable SVG, HTML, or Three.js sketches as an intermediate canvas.

The quality of a plan also depends on whether the initial request provides enough information to construct it. ProactiveT2I [81] recovers missing constraints through user questions, whereas co-DrawAgents [82] uses internal agent dialogue to incrementally construct and check a compositional layout. Retrieval-based methods instead supply external visual or semantic evidence. ImageRAG [83] retrieves references for concepts unfamiliar to the generator, FineRAG [84] combines query decomposition with self-reflective retrieval, and Unify-Agent [85] incorporates multimodal world knowledge. WeAgent-MMGenEdit [86] stores user-provided and retrieved evidence in a per-trajectory workspace with stable identifiers and provenance. Dedicated vision and code tools verify candidate images and organize the selected evidence into a rendered carrier that binds facts and identities to spatial positions before generation. Recent systems make context acquisition itself a planning decision. Qwen-Image-Agent [87] assembles sufficient context by coordinating reasoning with search, memory, and feedback. RS-Gen [88] diagnoses logical and knowledge gaps through a questioning-and-solving loop, whereas SearchGen [89] models the boundary between knowledge already available to a generator and knowledge that should be retrieved. Mind-Brush [90] routes detected cognitive gaps through multimodal search or reasoning before consolidating the evidence into a generation prompt. CMA [91] instead retrieves task-relevant episodes from structured visual memory to construct context for long-horizon multimodal interaction.

 doi:10.20944/preprints202609.0682.v1

Table 1. An overview of representative agentic systems and closely related model-based methods for target-centered visual generation. Methods are compared in terms of output, input reference, agent-side representation, execution target, feedback signal, and optimization action. Input references: T = text, D = dialogue, R = reference image/video, K = retrieved knowledge, and A = audio/context. Agent-side representations: Lyt = layout, Reg = region prompts, Scr = script, Ctrl = control signals, TG = tool graph, Ret = retrieved evidence, Crit = critic trace, and Traj = experience trajectory. A dash in Agent-Side Rep. denotes contextual state internal to the generator. Optimization actions: PR = prompt rewriting, RR = rerouting/regeneration, LE = local editing, SR = sample ranking, LR = learned reflection, and MU = memory/evidence update. Parenthetical tags indicate the stage of a feedback signal: train, pre-render, runtime-output.

Method	Venue	Output	Input Ref.	Agent-Side Rep.	Execution Target	Feedback Signal	Optimization Action
Planning / Context Construction							
LayoutGPT [17]	NeurIPS 2023	Layout image / scene	T	Lyt	T2I / scene renderer	—	—
LMD [18]	TMLR 2024	Compositional image	T	Lyt	T2I diffusion	—	—
LayoutLLM-T2I [70]	ACM MM 2023	Layout-guided image	T	Lyt	T2I diffusion	—	—
ProactiveT2I [81]	ICML 2025	Interactive image	T, D	Crit / Lyt	T2I generator	user questions	PR
RPG-DiffusionMaster [72]	ICML 2024	Region-composed image	T	Reg / Lyt	T2I diffusion	MLLM	PR
Ranni [73]	CVPR 2024	Instruction-following image	T	Reg / semantic panels	T2I diffusion	—	PR
MuLan [74]	arXiv 2024	Multi-object image	T, D	Reg / Ctrl	T2I diffusion	user / VLM	PR / LE
VideoDirectorGPT [19]	COLM 2024	Short video	T	Scr	T2V generator	—	PR
DirectorLLM [92]	arXiv 2024	Human-centric video	T	Scr / Ctrl	T2V generator	—	PR
Free-Bloom [93]	NeurIPS 2023	Multi-scene video	T	Scr	LLM director + LDM animator	—	PR
VideoStudio [78]	ECCV 2024	Multi-scene video	T	Scr	LLM director + animator	—	PR
Vlogger [79]	CVPR 2024	Vlog video	T	Scr / Ctrl	LLM director + video model	—	PR
DirecT2V [94]	arXiv 2024	Short video	T	Scr	frame-level director	—	PR
LVD [95]	ICLR 2024	Dynamic-scene video	T	Ctrl	T2V diffusion	—	control update
3D-aware Video Director [96]	NeurIPS 2024	3D-aware video	T	Scr / Ctrl	3D concept compositor	—	control update
VideoTetris [97]	NeurIPS 2024	Compositional video	T	Reg / Ctrl	T2V diffusion	—	control update
MEVG [98]	ECCV 2024	Multi-event video	T	Scr / Ctrl	T2V diffusion	—	PR
DivCon [71]	ECAI 2025	Constrained image	T	Lyt / Reg	T2I diffusion	constraint check	LE
MotionAgent [99]	ICCV 2025	Motion video	T, R	Ctrl	diffusion controller	motion misalignment (runtime-output)	control update
CogPortrait [100]	arXiv 2026	Portrait animation	T, R, A	Scr / Ctrl	DiT animator	constraint check	control update
FlowZero [101]	arXiv 2023	Dynamic-scene video	T	Ctrl	T2V generator	LLM layout verification (pre-render)	control update
coDrawAgents [82]	CVPR Findings 2026	Compositional image	T	Lyt / Crit	T2I diffusion	checker	PR / LE
SCOPE [75]	arXiv 2026	Complex image	T	semantic spec.	skill-conditioned T2I	entity / constraint verification	PR / LE / RR
GenClaw [80]	arXiv 2026	Image generation	T, K	Code / Lyt	code renderer + T2I	VLM / user review (runtime-output)	control update
ImageRAG [83]	ICLR 2026	Knowledge-rich image	T, K	Ret	T2I diffusion	VLM gap diagnosis (runtime-output)	MU / RR
FineRAG [84]	COLING 2025	Knowledge-rich image	T, K	Ret	T2I diffusion	self-reflection	MU
Unify-Agent [85]	arXiv 2026	World-grounded image	T, K	Ret / TG	multimodal T2I agent	visual-reference selection (pre-render)	PR / MU
WeAgent-MMGenEdit [86]	arXiv 2026	Knowledge-intensive image	T, (R), K	Ret / TG	multimodal harness + image backend	visual evidence verification (pre-render)	MU
KGEdit [102]	arXiv 2026	Video generation / editing	T	KG / Ctrl	T2V diffusion	semantic constraints	control update
Qwen-Image-Agent [87]	arXiv 2026	Knowledge-grounded image	T, (D), (R), K	Ret / Traj	T2I agent	context / visual feedback	PR / MU
RS-Gen [88]	arXiv 2026	Reasoning-rich image	T, D, (R), K	Ret / Crit	T2I / I2I model	gap diagnosis	PR / MU
SearchGen [89]	arXiv 2026	Open-world image	T, K	Ret / Crit	T2I generator	knowledge boundary	MU
MetaPoint [103]	arXiv 2026	Spatially controlled image	T	Lyt / Ctrl	visual generator	spatial error localization (runtime-output)	control update
MCCD [76]	CVPR 2025	Compositional image	T	Lyt / Reg	compositional diffusion	MLLM evaluator	PR / control update
3D Space as a Scratchpad [77]	CVPR 2026	Compositional image	T	3D scene / Ctrl	T2I + image-to-3D	render / VLM	control update
Mind-Brush [90]	arXiv 2026	Knowledge-grounded image	T, (R), K	Ret	search/reasoning + T2I	retrieved evidence	PR / MU
CMA [91]	arXiv 2026	Dialog image	T, D, R	Ret / TG	visual toolset	episodic retrieval	PR / MU
Ontology-guided Generation [104]	Array 2026	Rare-event image	T, (R)	ontology / Crit	image APIs	LMM panel + metrics	PR / RR / LE
SceneDecorator [105]	NeurIPS 2025	Scene-oriented story	T	Scr / Ctrl	VLM planner + T2I diffusion	—	PR
Narrative Weaver [106]	CVPR 2026	Storyboard / video	T, R	Scr / Lyt	AR MLLM + T2I diffusion	—	PR / control update
DreamRunner [107]	AAAI 2026	Story-to-video	T, R	Scr / Lyt / Ret	LLM planner + T2V diffusion	—	PR
Animate-A-Story [108]	ECCVW 2024	Storytelling video	T, R	Ret / Ctrl	video retrieval + T2V diffusion	—	control update
FairyGen [109]	SIGGRAPH Asia 2025	Cartoon video	R	Scr / Ctrl	MLLM planner + I2V diffusion	—	PR / control update
VGoT [110]	arXiv 2025	Multi-shot video	T	Scr / Ctrl	LLM storyline planner + T2V diffusion	storyline validation	PR / control update
STAGE [111]	CVPR 2026	Multi-shot narrative	T	Scr / Ctrl	director agent + STEP2 + I2V diffusion	—	PR / control update
MV-Crafter [112]	ACM TIIS 2025	Music video	T, A	Scr / Ctrl	LLM + T2I/I2V + synchronizer	user edits	PR / LE / control update
Closely Related Model-Based Methods							
StoryDiffusion [113]	NeurIPS 2024	Image/video story	T, (R)	—	T2I diffusion + transition model	—	—
StoryGPT-V [114]	CVPR 2025	Image story	T, R	—	LLM-conditioned T2I diffusion	—	—
SEED-Story [115]	ICCV 2025	Long story generation	T, (R)	—	MLLM + T2I diffusion	—	—
Story-Iter [116]	ICLR 2026	Long story visualization	T	—	iterative T2I diffusion	previous-iteration frames	RR
OneStory [117]	CVPR 2026	Multi-shot video	T, (R)	—	memory-conditioned I2V DiT	—	—
Motion by Queries [118]	arXiv 2025	Multi-shot video	T, R	—	T2V feature injection	—	—
ShotAdapter [119]	CVPR 2025	Multi-shot video	T	Scr / Ctrl	T2V diffusion	—	shot control
MovieDreamer [120]	ICLR 2025	Long visual sequence	T, (R)	—	AR keyframe model + I2V diffusion	—	—
Plot'n Polish [121]	AAAI 2026	Story visualization/editing	T, (R)	—	T2I diffusion + multi-frame editor	user edit prompts	LE
Tool and Model Routing							
DiffAgent [20]	CVPR 2024	Image generation	T	TG	T2I API pool	—	RR
DiffusionAgent [122]	arXiv 2024	Image generation	T	TG	expert model tree	—	RR
OctoT2I [123]	CVPR 2026	Cost-aware image	T	TG	T2I generator pool	cost/perf. score	RR / LR
Talk2Image [124]	AAAI 2026	Dialog image	T, D	TG / Reg	T2I + editing tools	multi-view score / user feedback	RR / LE
T2I-Copilot [125]	ICCV 2025	Dialog image	T, D	TG / Crit	T2I backends	VLM	RR
GenArtist [126]	NeurIPS 2024	Image generation/editing	T, R	TG / Reg	visual toolset	VLM	RR / LE
Mora [127]	arXiv 2024	Generalist video	T, R	TG	image/video agent toolkit	—	RR
SPAgent [128]	IEEE TIP 2026	Video generation/editing	T, R	TG	expert video models	—	RR
UniVA [129]	arXiv 2025	Universal video agent	T, R	TG / Crit	video generalist toolkit	VLM	RR
NEWTON [130]	arXiv 2026	Physics-grounded video	T	TG / Crit	T2V + physics tools	verifier	RR / PR
VideoWeaver [131]	arXiv 2026	Long video	T, R, A	TG / Traj	evolving video skills	trace + video judge	RR / LR
GenRouter [132]	arXiv 2026	Image generation	T	TG / Traj	agentic workflows + T2I pool	quality / cost / latency	RR / MU
ToolArtist [133]	arXiv 2026	Open-world image	T, K	Ret / Traj	native UMM generation + search tools	intent / image-quality rewards	policy update
Self-Refinement / Prompt Evolution							
CompAgent [21]	arXiv 2024	Compositional image	T	Reg / Crit	T2I + editing tools	VLM / detector	PR / LE
PAE [134]	CVPR 2024	Text-to-image	T	Crit	T2I generator	image-based PPO reward (train)	PR
Idea2Img [135]	ECCV 2024	Image design	T, (R)	Crit	T2I generator	GPT-4V	PR / SR
SLD [136]	CVPR 2024	Compositional image	T	Reg / Crit	LLM-controlled diffusion	detector	LE
Maestro [22]	arXiv 2025	Text-to-image	T	Crit	T2I generator	VLM preference	PR / SR
PromptSculptor [137]	EMNLP 2025	Prompt-optimized image	T	Crit	T2I generator	multi-agent feedback	PR
RePrompt [138]	arXiv 2025	Text-to-image	T	Crit	T2I generator	RL reward	PR / LR
RAPO [139]	CVPR 2025	Text-to-video	T, K	Ret / Crit	T2V generator	discriminator verdict (pre-render)	PR / prompt selection
PRIS [140]	CVPR 2026	Text-to-visual	T	Crit	visual generator	failure statistics	PR / SR
RAISE [141]	CVPR 2026	Text-to-image	T	Crit / Traj	T2I + I2I models	tool-grounded requirement checklist	PR / RR / LE / SR
VISTA [142]	CVPR 2026	Short video	T	Crit	T2V generator	VLM	PR
Agentic Self-Improvement [143]	arXiv 2026	I2V video	T, R	Crit	black-box I2V generator	DSG/CMQ VQA + metrics	PR / parameter search / SR
VideoRepair [144]	ACL 2026 Findings	Short video	T	Reg / Crit	T2V + local refiner	mismatch detector	LE
PhotoFlow [145]	arXiv 2026	3D virtual photo	T	Ctrl / Crit	Blender renderer	reviewer score	SR / control update
Agentic Retoucher [146]	CVPR 2026	Text-to-image	T	Crit / Reg	T2I + retouching tools	defect detector	LE
AgentComp [147]	arXiv 2025	Compositional image	T	Crit / Traj	multimodal generator	agentic reasoning	LR
SILMM [148]	CVPR 2025	Compositional image	T	Crit / Traj	multimodal generator	self-feedback	LR
GenMAC [149]	AAAI 2026	Compositional video	T	Reg / Crit	T2V + layout control	MLLM agents	PR / control update / RR
CREA [150]	NeurIPS 2025	Creative image	T, R	TG / Crit	T2I + editing tools	multi-agent feedback	PR / LE
ReflectDiT [151]	ICCV 2025	Text-to-image	T	Traj / Crit	diffusion transformer	in-context reflection	LR
ReflectionFlow [152]	ICCV 2025	Text-to-image	T	Crit	T2I diffusion	reflection model	LR
DreamSync [153]	NAACL 2025	Text-to-image	T	Crit	T2I generator	image-understanding feedback	LR
InterleaveThinker [154]	arXiv 2026	Interleaved text-image	T	Traj / Crit	interleaved generator	planner-critic feedback	LR
Generation Navigator [155]	arXiv 2026	Text-to-image	T	Crit / Traj	T2I / I2I generator	reviewer score	RR / LE / SR
PromptEnhancer [156]	CVPR 2026	Prompt-optimized image	T	Crit / Traj	T2I generator	AlignEvaluator	PR / LR
VisionDirector [157]	CVPR 2026	Image generation/editing	T, R	Crit / goal state	T2I / I2I generator	VLM goal check	LE / RR / SR
GenAgent [158]	arXiv 2026	Text-to-image	T	Traj / Crit	T2I tool	self-reflection	LR
GenEvolve [159]	arXiv 2026	Image generation	T	TG / Traj	visual toolset	self-reflection	LR
SIDiffAgent [160]	arXiv 2026	Image generation	T	Traj / Crit	diffusion agent	self-improvement	LR
GEMS [67]	arXiv 2026	Image generation	T	Traj / Crit / TG	T2I generator	multi-agent feedback	LR
MemoGen [161]	arXiv 2026	Knowledge-rich image	T, K	Ret / Crit / Traj	T2I generator	visual feedback	PR / MU / LR
APE [162]	arXiv 2026	Image generation / editing	T	Crit / TG	visual generator	task-aware reward	PR
Product Grid-Collage [163]	arXiv 2026	Narrative collage	T, R	Lyt / Crit	image generator	content / quality gates	PR / RR
MAVEN [164]	arXiv 2026	Culturally grounded video	T	Scr / Crit	T2V generator	—	PR
Genflow [165]	CAIS 2026	Brand-aligned video	T, K	Ret / Crit	video generator	multi-agent QC	PR / RR
CHIEF [166]	ICMLW 2026	Creator-driven video	T, D	Crit	video generator	human + persona critique	PR / RR

 doi:10.20944/preprints202609.0682.v1

Once sufficient context has been assembled, the plan must be expressed in controls understood by the selected generator. MetaPoint [103] encodes point and box coordinates as compositional tokens for precise spatial control. KGEdit [102] organizes identity, relation, attribute, and negative constraints in an ambiguity-aware knowledge graph. For rare-event synthesis, Ontology-guided Generation [104] instantiates a formal domain ontology as a structured semantic specification for controllable image generation. DirectorLLM [92] grounds human-centric video plans in pose tokens, while MotionAgent [99] derives trajectories, camera extrinsics, and optical-flow guidance. CogPortrait [100] produces event-level eye-motion plans, and PhotoFlow [145] turns photography requests into executable camera parameters. At a broader temporal scale, DirecT2V [94] and Free-Bloom [93] distribute instructions across frames or successive stages of generation. FlowZero [101], LVD [95], and 3D-aware Video Director [96] construct dynamic scene syntax, spatiotemporal layouts, or 3D-aware controls. VideoTetris [97] and MEVG [98] adapt such planning to compositional and multi-event synthesis, respectively. In these systems, planning serves as an interface between high-level creative intent and the control space of a particular generator, rather than remaining a free-form reasoning process.

The same planning principle extends to longer target-centered sequences without changing the unit around which state is organized. SceneDecorator [105] separates global and local scene prompts, while Narrative Weaver [106] produces storyboards and layouts. DreamRunner [107] combines hierarchical plans with motion retrieval, and Animate-A-Story [108] similarly uses retrieved motion structure as generation control. Starting from a character sketch, FairyGen [109] constructs shot-level storyboards together with motion controls. Other systems make the temporal organization of the sequence more explicit. VGoT [110] expands a prompt into validated shot specifications, STAGE [111] converts a text storyboard into start–end frame pairs, and MV-Crafter [112] constructs music-conditioned scripts and synchronizes generated clips along a timeline.

Closely related model-based methods preserve story context within the generation process rather than through a separately managed production state. StoryDiffusion [113], StoryGPT-V [114], SEED-Story [115], Story-Iter [116], OneStory [117], Motion by Queries [118], ShotAdapter [119], MovieDreamer [120], and Plot'n Polish [121] follow this route. Some carry context through shared attention or multi-modal tokens. Others revisit selected prior frames or iterative visual references. Query injection, auto-regressive visual-token histories, and grid-based correspondences provide further ways to propagate information across the artifact. We discuss these methods alongside target-centered agents because they improve coherence through generator-level context and conditioning. Such mechanisms do not by themselves establish agent-side coordination of separate production units.

Tool and Model Routing.

Tool and model routing matches a visual request to the execution capability best aligned with its requirements. In its simplest form, this is a choice among models, APIs, and parameter settings. DiffAgent [20] selects a text-to-image service and configures its parameters from a diffusion-model ecosystem, while DiffusionAgent [122] searches a hierarchical model tree for an appropriate domain expert. OctoT2I [123] makes the decision cost-aware by learning capability profiles for available generators and balancing expected quality against inference cost. These systems treat generators as alternative backends whose strengths can be matched to the requirements of each prompt.

When a request spans several operations, routing becomes part of the execution plan rather than a one-time model choice. Talk2Image [124] resolves intentions expressed over multiple dialogue turns and schedules the corresponding generation or editing actions. T2I-Copilot [125] separates this process across agents that interpret the request, select a text-to-image backend and its control inputs, and assess whether the result should be regenerated. GenArtist [126] decomposes a request into object, background, and operation nodes, searches a planning tree over candidate tools, and follows an alternative branch when verification exposes an execution failure. CMA [91] combines episodic visual retrieval with an executive controller that selects the required operation at each turn.

 doi:10.20944/preprints202609.0682.v1

The route therefore defines how a request moves from interpretation to execution. It determines what information must be prepared before each call and how the plan should change when a tool fails.

At a broader scale, routing governs complete production workflows. GenRouter [132] jointly selects a workflow and generator from predefined templates, using demand profiles and routing memories to balance expected visual quality against cost and latency. Mora [127] composes specialized image and video modules for generation and editing tasks, while SPAgent [128] couples task decomposition with capability-based model selection. UniVA [129] extends this organization to a general video toolkit whose operations are coordinated with multi-modal assessment. In physically grounded video generation, NEWTON [130] connects keyframe generation, physics computation, prompt refinement, and verification within one executable workflow. VideoWeaver [131] makes the workflow itself adaptable by assembling long-video pipelines from reusable skills and refining those skills using process-aware evaluation.

ToolArtist [133] takes a different approach by learning routing and generation jointly. Its multi-modal policy learns when generation should proceed, when external evidence is needed, and when the current result should be inspected and revised. Reason-Act-Draw GRPO propagates intent and image-quality rewards across the resulting trajectory. At this level, effective routing depends on a current operational model of each component that captures how it should be invoked and whether its output can support the next stage. If that model becomes outdated, or if verification fails to detect an interface mismatch, this account also guides rerouting, parameter revision, or regeneration after execution failures.

Prompt Revision and Output Refinement.

Self-refinement turns visual generation into a feedback loop in which an agent evaluates and ranks candidates, diagnoses failures, and selects among prompt rewriting, local repair, and regeneration. CompAgent [21] verifies object and attribute relations in compositional images and repairs detected errors through local editing. Maestro [22] instead treats prompt design as iterative search, using specialized critics and pairwise comparisons to guide targeted rewrites. PRIS [140] broadens the evidence available to this loop by aggregating element-level failures across multiple samples before redesigning the prompt. RAISE [141] implements a training-free, requirement-adaptive loop that evolves a candidate population through prompt rewriting, noise resampling, and instructional editing. A tool-grounded binary checklist tracks unresolved requirements and provides a stopping rule. These methods frame refinement as a decision about how much of the current attempt needs to change.

Some systems act primarily on the prompt while leaving the generator unchanged. PromptSculptor [137] uses role-specialized agents to enrich intent and revise the prompt in response to critique. The Agentic Prompt Enhancer [162] assigns prompt optimization into routing, rewriting, and composition. MAVEN [164] distributes descriptions of people, actions, and locations across specialized agents to improve cultural fidelity in video generation. Other methods learn or retrieve prompt transformations more directly. RePrompt [138] learns reasoning-guided reprompting, while PAE [134] learns modifier tokens, denoising ranges, and weights from image-based training rewards. RAPO [139] retrieves modifiers and uses a discriminator trained from video evaluations to select a prompt before generation. VISTA [142] draws on visual and audio observations as well as broader context. PromptEnhancer [156] trains a chain-of-thought rewriter with fine-grained rewards from generated images, allowing feedback into a model-agnostic reformulation. Across these methods, the object of refinement is the condition supplied to the generator rather than the visual artifact itself.

Other systems use feedback to alter the current candidate or the controls that produced it. Idea2Img [135] compares drafts and uses multimodal reflection to guide the next attempt, whereas SLD [136] detects compositional errors and corrects them in latent space. VideoRepair [144] localizes text-to-video mismatches for partial regeneration. Agentic Retoucher [146] identifies defects before applying local edits. PhotoFlow [145] searches over virtual-camera configurations using reviewer feedback and region memory. Agentic Self-Improvement [143] extends feedback-driven I2V refinement from language to stochastic controls. It uses negative MLLM answers to prompt-specific Davidsonian

 doi:10.20944/preprints202609.0682.v1

Scene Graph and Common Mistake Questions, then applies Bayesian optimization over random seeds and classifier-free guidance scales under adherence and quality scores.

Choosing an appropriate intervention is itself part of refinement. Generation Navigator [155] learns the higher-level decision of whether to stop, refine the current image, or regenerate from a revised prompt. VisionDirector [157] decomposes long instructions into goal-level state and alternates generation with verified micro-edits and rollback. Ontology-guided Generation [104] converts evaluator recommendations into targeted refinement or regeneration. When feedback comes from several evaluators or participants, the loop must also coordinate their judgments. Self-Reasoning Grid-Collage framework [163] uses quality gates and failure attribution for targeted retries. GenMAC [149] and CREA [150] divide generation, critique, and redesign among collaborating agents. Genflow [165] evaluates video against retrieved brand constraints, whereas CHIEF [166] combines creator-directed revisions with critiques from audience-oriented personas.

Some methods retain feedback after the immediate correction and use it to improve later behavior. AgentComp [147] turns agent-generated edits and hard negatives into preference-optimization data. ReflectDiT [151] and ReflectionFlow [152] internalize reflection through in-context adaptation and reflection tuning, respectively. SILMM [148] learns through self-questioning, while DreamSync [153] filters training examples using feedback. GenAgent [158] learns from full multi-turn refinement trajectories, whereas InterleaveThinker [154] trains a planner and critic to interact during interleaved generation. Experience-oriented systems preserve reusable guidance in a form that can be retrieved later. GenEvolve [159] distills tool-use experience, and SIDiffAgent [160] learns from successful and failed attempts. GEMS [67] couples trajectory memory with an extensible skill library. MemoGen [161] preserves a structured record of how earlier tasks were interpreted and corrected, together with the evidence and feedback behind those decisions. At this point, self-refinement becomes a source of cumulative improvement rather than a correction mechanism confined to one output.

These mechanisms support production around the current target artifact. Planning prepares generation inputs, routing selects models or tools, and refinement uses feedback to improve candidates. Individual methods may implement only part of this process. Coordinated visual generation extends agent-side management to identifiable components linked through shared specifications, assets, or state, as summarized in Figure 3.

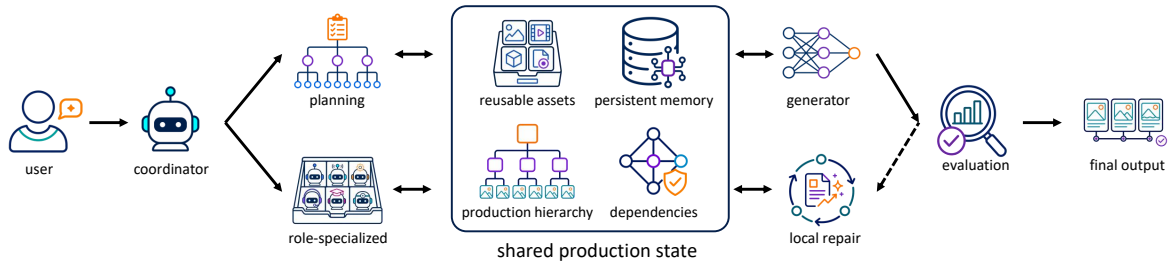

Figure 3. Overview of agentic pipelines in coordinated visual generation. The agent maintains shared production state and explicit dependencies across components that can be generated or revised separately, supporting four mechanisms: hierarchical production planning; persistent state, consistency management, and asset reuse; role-based production coordination; and consistency checking and local regeneration.

3.1.2. Coordinated Visual Generation

Coordinated visual generation treats a visual artifact as a set of separately addressable components, such as shots in a video or layers in an image, that must remain consistent as a whole. Coordination is mediated through shared specifications, persistent memories, reusable assets, or explicit relationships that carry decisions across components. A character reference may condition later shots, while a storyboard or event graph may specify how individual units contribute to the larger artifact. Some systems go further by updating persistent state, tracing dependencies, or revising only the affected components. Existing systems realize this coordination through four recurring mechanisms, as summarized in Table 2:

 doi:10.20944/preprints202609.0682.v1

Table 2. An overview of representative agentic systems for coordinated visual generation. Methods are compared using high-level coordination tags rather than method-specific role names. Unit tags denote the dominant component granularity: Panel, Shot, Segment, Asset, and Layer/Obj. State tags summarize maintained production state: Plan, Memory, Spatial, Timeline, and Constraint. Coordination modes include Plan-Gen = planner–generator pipeline, Dir-Crew = director with specialized production agents, Mem-Crit = memory/critic loop, Mem-Gen = memory-conditioned generation, HITL = human-in-the-loop, and Domain-Team = domain-specialized agents. Consistency axes are Entity, Sequence, Form, and Grounding. Control modes include StateCond = state-conditioned generation, Reuse = reference/asset reuse, PlanAsm = planning or assembly, ReviewRepair = audit/review/local repair, and ConstraintRev = constraint-guided revision.

Method	Venue	Output	Unit	State	Coordination	Consistency Axis	Control Mode
Story / Storyboard Generation							
TaleCrafter [23]	SIGGRAPH Asia 2023	Image story	Panel+Asset	Plan+Memory	HITL+Plan-Gen	Entity+Sequence	Reuse+StateCond
AutoStory [167]	IJCV 2025	Image story	Panel+Asset	Plan+Memory	HITL+Plan-Gen	Entity+Sequence	StateCond+Reuse
MUSE-story [168]	arXiv 2026	Story visualization	Panel+Shot	Plan+Memory	Dir-Crew+Mem-Crit	Sequence+Form	ReviewRepair
StoryAgent [169]	arXiv 2024	Storytelling video	Panel+Shot	Plan+Timeline	Dir-Crew	Sequence	PlanAsm
AutoStudio [170]	CVPRW 2026	Image story	Panel+Asset	Memory+Spatial	Mem-Crit	Entity+Form	ReviewRepair
TheaterGen [171]	arXiv 2024	Image story	Panel+Asset	Plan+Memory	Plan-Gen	Entity+Form	StateCond+Reuse
Audit and Repair [172]	arXiv 2025	Story visualization	Panel+Asset	Memory+Constraint	Mem-Crit	Entity+Sequence	ReviewRepair
S2ED [173]	ICME 2026	Image story	Panel+Asset	Plan+Memory	Plan-Gen	Entity+Sequence	StateCond+ConstraintRev
BookAgent [174]	ACL Findings 2026	Illustrated storybook	Panel+Asset	Plan+Memory+Constraint	Dir-Crew+Mem-Crit	Entity+Sequence+Grounding	ReviewRepair+ConstraintRev
VisionCreator [175]	arXiv 2026	Multi-output image/video	Panel+Shot	Plan+Constraint	Plan-Gen	Entity+Sequence	PlanAsm+Reuse
Cinematic and Long-Form Video							
FilmAgent [25]	arXiv 2025	Cinematic video / 3D film	Shot+Asset	Plan+Memory	Dir-Crew+Mem-Crit	Entity+Sequence+Form	ReviewRepair
FilmWorld [176]	arXiv 2026	Novel-to-film video	Shot+Asset	Plan+Memory+Timeline	Dir-Crew+Mem-Crit	Entity+Sequence+Form+Grounding	StateCond+Reuse+ReviewRepair
MovieAgent [26]	arXiv 2025	Movie generation	Shot	Plan	Dir-Crew	Sequence+Form	PlanAsm
FilMaster [177]	arXiv 2025	Cinematic video	Shot	Plan+Constraint	Plan-Gen	Sequence+Form	ConstraintRev
DreamFactory [27]	arXiv 2024	Multi-scene video	Shot	Plan	Dir-Crew+Mem-Crit	Sequence+Form	ReviewRepair
Kubrick [178]	CVPRW 2025	Synthetic video	Shot+Layer/Obj	Plan+Spatial	Plan-Gen+Mem-Crit	Sequence+Form	PlanAsm+ReviewRepair
Hollywood Town [179]	arXiv 2025	Long video	Shot+Segment+Asset	Plan+Memory+Timeline	Dir-Crew	Sequence+Grounding	PlanAsm+ReviewRepair
One Sentence One Drama [180]	arXiv 2026	Short-form drama	Shot+Segment+Asset	Plan+Memory+Spatial	Dir-Crew+Mem-Crit	Entity+Sequence+Grounding	PlanAsm+ReviewRepair
ScripterAgent [181]	arXiv 2026	Dialogue video	Shot+Segment	Plan+Timeline	Plan-Gen	Sequence+Grounding	StateCond
CoAgent [182]	arXiv 2025	Coherent video	Shot+Asset	Memory	Mem-Crit	Entity+Sequence	ReviewRepair
Co-Director [183]	arXiv 2026	Generative video story	Shot+Segment	Plan+Constraint	Plan-Gen+Mem-Crit	Sequence+Form+Grounding	ConstraintRev
ViMax [184]	arXiv 2026	Agentic video generation	Shot+Asset	Memory	Mem-Crit	Entity+Form	StateCond+ReviewRepair
A2RD [185]	arXiv 2026	Long video	Shot+Asset	Memory+Timeline	Mem-Crit	Entity+Sequence	ReviewRepair
VideoMemory [24]	arXiv 2026	Long narrative video	Shot+Asset	Memory	Mem-Gen	Entity+Sequence	Reuse+StateCond
GroundShot [186]	arXiv 2026	Multi-shot video	Shot+Asset	Plan+Memory	Plan-Gen+Mem-Crit	Entity+Sequence	Reuse+StateCond
CineAGI [187]	ICME 2026	Cinematic video	Shot+Asset	Plan+Memory+Timeline	Dir-Crew	Entity+Sequence+Form	PlanAsm+StateCond
Animation and Audiovisual Production							
Anim-Director [188]	SIGGRAPH Asia 2024	Animation	Shot+Asset	Plan+Timeline	Dir-Crew+Mem-Crit	Sequence+Form	PlanAsm+ReviewRepair
AniME [189]	SIGGRAPH Asia 2025 Posters	Long animation	Shot+Asset	Plan+Memory	Plan-Gen+Mem-Crit	Sequence+Form	ReviewRepair
AniMaker [190]	SIGGRAPH Asia 2025	Animated storytelling	Shot+Segment	Plan+Timeline	Plan-Gen+Mem-Crit	Sequence+Form	ConstraintRev
AnimAgents [191]	arXiv 2025	Animation pre-production	Asset+Panel	Plan+Memory	HITL+Dir-Crew	Sequence+Form	PlanAsm
AnimeAgent [192]	arXiv 2026	Anime video	Shot+Asset	Plan+Memory	Dir-Crew+Mem-Crit	Entity+Sequence+Form	ReviewRepair
MAViS [193]	EACL 2026	Long-sequence video	Shot+Segment+Asset	Plan+Memory+Timeline	Dir-Crew	Entity+Sequence+Grounding	PlanAsm+ReviewRepair
MM-StoryAgent [194]	arXiv 2025	Narrated storybook video	Panel+Segment	Plan+Timeline	Dir-Crew+Mem-Crit	Sequence+Grounding	PlanAsm+ReviewRepair
AutoMV [195]	arXiv 2025	Music video	Shot+Segment	Timeline	Domain-Team	Sequence+Grounding	PlanAsm
LVAS-Agent [196]	arXiv 2025	Long-video audio synthesis	Shot+Segment	Timeline	Domain-Team+Mem-Crit	Sequence+Grounding	ConstraintRev
Domain-Specific Production							
Preacher [197]	ICCV 2025	Paper-to-video	Segment+Shot	Plan+Constraint	Domain-Team+Mem-Crit	Grounding+Form	ReviewRepair
SciTalk [198]	arXiv 2025	Scientific short-form video	Segment+Shot	Plan+Constraint	Domain-Team+Mem-Crit	Grounding+Form	ConstraintRev+ReviewRepair
VideoAgent [199]	ICMR 2026	Scientific video	Segment+Shot	Plan+Constraint	Domain-Team	Grounding+Sequence	StateCond
LASEV [200]	KDD 2026	Educational video	Segment+Shot	Plan+Constraint	Domain-Team+Mem-Crit	Grounding	ConstraintRev
Code2Video [201]	arXiv 2025	Educational video	Segment+Layer/Obj	Plan+Constraint	Domain-Team+Mem-Crit	Grounding+Form	PlanAsm+ConstraintRev
Script2Screen [202]	IUI 2026	Scripted audiovisual	Shot+Segment	Plan+Timeline	HITL+Plan-Gen	Sequence+Grounding	PlanAsm
Sima [203]	arXiv 2026	Documentary video	Shot+Segment	Plan+Constraint	Domain-Team	Grounding+Sequence	ConstraintRev
BrandFusion [204]	CVPR Findings 2026	Brand video	Shot+Asset	Memory+Constraint	Domain-Team+Mem-Crit	Entity+Grounding	Reuse+ConstraintRev
PosterCopilot [205]	arXiv 2025	Graphic design	Layer/Obj	Spatial+Constraint	HITL+Plan-Gen	Form+Grounding	PlanAsm+ConstraintRev
MUSE-3D [206]	arXiv 2026	3D scene/design	Layer/Obj+Asset	Spatial+Memory	Plan-Gen+Mem-Crit	Form+Grounding	StateCond+ReviewRepair
SimWorlds [207]	arXiv 2026	Dynamic 4D scene	Layer/Obj+Asset	Plan+Spatial+Timeline	Plan-Gen+Mem-Crit	Form+Grounding	PlanAsm+ReviewRepair

Hierarchical Production Planning.

Hierarchical production planning turns a creative goal into identifiable production units and specifies how higher-level decisions guide their generation. FilmAgent [25], for example, develops a story idea into a production plan that connects narrative structure with character, staging, and camera decisions before rendering the result in Unity. MovieAgent [26], DreamFactory [27], ScripterAgent [181], MUSE [168], and Co-Director [183] follow the same principle, organizing production from high-level narrative structure to shot-level execution. One Sentence One Drama [180] adapts this hierarchy to personalized short-form drama, linking episode planning and visual assets to 3D-grounded keyframes and post-production. CineAGI [187] constructs cinematic blueprints that specify characters, scenes, and cross-modal requirements. VisionCreator [175] learns long-horizon trajectories that translate storyboard and multi-output requests into sequences of executable tool calls. In these systems, the hierarchy is not merely a task decomposition. It records how high-level creative decisions constrain the assets and shots produced.

The form of the plan depends on how downstream modules use it. FilMaster [177] keeps planning at the shot level but grounds each decision in cinematic principles. S2ED [173] turns a story into executable descriptions that preserve the context needed for continuity across panels. GEST Authoring Agents [208] make the structure more explicit by representing a scene as an event graph that a 3D game engine can execute. These representations give generation modules a shared specification and provide identifiable units for later state management and consistency checking.

The same planning principle extends beyond cinematic video. TaleCrafter [23] and AutoStory [167] organize narratives into separately managed panels and character-centered events. Animation

 doi:10.20944/preprints202609.0682.v1

systems further separate story structure from character design, settings, storyboards, motion, and clip generation: Anim-Director [188] and AniMaker [190] construct multi-stage animation plans, while AniME [189] and AnimAgents [191] organize longer productions around adaptive or role-based planning. Source-grounded systems use a similar structure when the output must remain faithful to external material. SciTalk [198], Code2Video [201], and VideoAgent [199] convert scientific content into coordinated audiovisual scenes rather than treating narration and visuals as independent outputs. PosterCopilot [205] applies hierarchical planning to editable design layers, while MUSE-3D [206] turns scene requirements into incremental authoring steps. Across these domains, decomposition becomes operational only when each production unit remains identifiable after planning, allowing later generation, state tracking, and verification to act on it directly.

Persistent State, Consistency Management, and Asset Reuse.

Persistent state distinguishes coordinated generation from pipelines that treat every component as an isolated request. TheaterGen [171] uses a prompt book to preserve the information needed to reproduce characters and scene arrangements in later shots. VideoMemory [24] organizes recurring visual entities into records that connect semantic attributes with reference images. MUSE-3D [206] extends this idea to scene authoring through working, scene, and skill memories that track requirement progress, persistent scene structure, protected bindings, and reusable decomposition patterns. These representations turn earlier decisions and generated content into named state that later components can access.

Consistency requires more than storing state. The system must also determine which part of that state should condition each component. GroundShot [186] builds an entity-level visual memory online, verifies candidate references before storing them, and retrieves accepted appearances when an entity returns in a later shot. FilmWorld [176] constructs a plot-driven sequence of entity states and associates them with visual anchors before rendering, then verifies keyframes and video segments against this shared state and locally regenerates failed units. AutoStudio [170] carries subject and layout information across interactive story generation, while CoAgent [182] and ViMax [184] use entity-relevant context to condition later components. These methods make consistency a retrieval and verification problem grounded in explicit state, rather than a property left entirely to the generator.

Persistent state also allows generated material to remain part of the production process as a reusable asset. BrandFusion [204] treats brand identity and insertion assets as constraints that must remain stable across generated videos, while PosterCopilot [205] preserves editable layer structures for subsequent design operations. The broader principle is that reuse depends on stable identity and addressability. Once an asset can be retrieved and updated without losing its links to dependent components, later revisions no longer require the workflow to reconstruct it from scratch. Persistent state therefore provides the basis for both cross-component consistency and targeted revision in coordinated visual production.

Role-Based Production Coordination.

Role-based production coordination assigns different parts of a shared production process to specialized agents. The aim is not simply to divide the work, but to make clear which role controls each decision and how the resulting state is passed on. FilmAgent [25] mirrors a film crew by separating the roles that develop and realize a scene from those that critique and judge the result. DreamFactory [27] and Hollywood Town [179] organize production as a sequence of handoffs from script development to final assembly, while MAViS [193] extends the same principle across a longer multimodal pipeline. StoryAgent [169] uses specialized roles to carry a story from narrative design through visual production and review. One Sentence One Drama [180] adapts this organization to personalized short-form drama, whereas BookAgent [174] separates content development from illustration and assigns a further role to global repair.

Another common design separates the agents that construct an artifact from those that supervise its construction. Kubrick [178] pairs a programmer with a reviewer so that executable scene con-

 doi:10.20944/preprints202609.0682.v1

struction remains distinct from inspection. MUSE-3D [206] assigns requirement interpretation, scene construction, and verification to Architect, Sculptor, and Inspector roles. SimWorlds [207] similarly uses separate agents to plan, implement, and review dynamic 4D scenes, with the implementation carried out through Blender procedures. In these systems, specification, execution, and acceptance are placed under different authority. No single agent must both produce and validate every change to the production state.

Role boundaries may also reflect modality or domain expertise. MM-StoryAgent [194] coordinates specialists across the verbal and audiovisual parts of storybook video production. AutoMV [195] organizes music-video generation around rhythm and clip selection, while LVAS-Agent [196] assigns long-video audio synthesis and alignment to dedicated agents. Sima [203] uses role-based coordination to connect factual material with documentary production. Co-Director [183] instead organizes its specialists around a shared creative strategy so that decisions made in one part of the production remain compatible with the rest. Across these designs, the value of specialization depends less on the number of roles than on the interfaces between them. Each handoff must expose enough of the current production state for the receiving agent to continue the work without reconstructing decisions that have already been made.

Consistency Checking and Local Regeneration.

Consistency checking asks whether a newly generated component remains compatible with decisions already encoded in the shared production state. This differs from evaluating the component as an isolated output. Audit and Repair for Story Visualization [172] compares generated panels with entity metadata, distinguishes intentional story changes from accidental appearance drift, and sends only the affected content for repair. Persistent state thus provides a reference for locating inconsistencies without regenerating the full sequence.

Once a failure has been localized, the system must decide how much of the existing result can be preserved and where revision should begin. CoAgent [182] and ViMax [184] use consistency checks to make this decision at the component level. MUSE [168] and A2RD [185] place review-and-repair loops around story or video units, preserving accepted material while revisiting failed shots or intermediate representations. When components share constraints, a local repair should account for its effects on other units. Revising one component in isolation may leave related decisions out of sync. Explicit dependency tracking can help identify the affected units and determine whether repair should extend beyond the component where the error first appears.

Consistency checks can also reflect production goals beyond cross-shot consistency. Co-Director [183] uses factored MLLM rewards to assess whether the production remains aligned with a shared creative strategy. SciTalk [198], Code2Video [201], and LASEV [200] apply related checks to scientific and educational videos, where a visually plausible result may still be unacceptable if it departs from the source material or fails to execute correctly. BrandFusion [204] and Sima [203] similarly judge outputs against commitments associated with brand identity, factual grounding, or the intended audience. In these settings, evaluation becomes part of the production logic rather than a final quality filter. This reliance on automated evaluation also introduces a point of failure. When closely related multimodal models both generate and judge the content, an inaccurate assessment may miss a genuine conflict or reopen a component that was already acceptable. Poor localization can then turn an unnecessary repair into a new source of inconsistency.

Coordinated visual generation is therefore defined by agent-side management of identifiable components rather than by the number of shots or agents. Hierarchical planning organizes these units, while shared state and role-based coordination connect their production. When runtime feedback is available, consistency checks can guide targeted revision. Explicit dependency tracking further helps determine whether a change should extend to other components.

 doi:10.20944/preprints202609.0682.v1

Table 3. Comparison of systems for agentic visual editing. Systems are grouped by editing scope. *Agent State* records the plan, memory, trace, or intermediate state used for decisions; *Execution Substrate* names the editor, tools, or software used to execute edits. Inputs: I = image, V = video, T = text, D = dialogue, R = visual reference, G = spatial guidance, A = audio, S = structured metadata, and E = initial edited result; parentheses denote optional inputs. Unmarked feedback and actions occur at inference time; (train) and (HITL) mark training and human-in-the-loop operation, respectively; —denotes none.

Method	Venue	Inputs	Agent State	Execution Substrate	Feedback Signal	Revision Action
Instruction-Guided Image Editing						
IMAGAgent [209]	arXiv 2026	I, T, D	subtask plan, history	retrieval, vision, editing tools	VLM critique, score	revise tools / re-edit
EditRefiner [210]	arXiv 2026	I, E, T	saliency maps, diagnoses	local image editor	saliency/MOS (train); quality scores	re-edit / stop
CAMEO [211]	arXiv 2026	I, T, (R)	constraints, prompt, critic trace	editor, reference tools	constraint deviations	revise prompt / re-edit / stop
ImageEdit-R1 [212]	arXiv 2026	I, T	edit tuples, subrequest order	diffusion editor	format/tuple rewards (train)	planner optimization (train)
Agent Banana [213]	arXiv 2026	I, T, D	state graph, layer plan	MCP editing tools	compliance, quality	retry / rollback / replan
From Plans to Pixels [214]	arXiv 2026	I, T	subgoals, tool/region plan	editing toolkit	VLM reward (train); step/identity/quality	re-rank
GMO-E^{2}DIT [215]	arXiv 2026	I, T, (G), (R)	edit agenda, masks, reflection	mask-conditioned editor	plan/reflection rewards (train); outcome label	continue / rollback / stop
MSRAMIE [216]	arXiv 2026	I, T	state tree, reference graph	plug-in editor	instruction, preservation, quality	select / resample / trace back / stop
MIRA [217]	CVPR Findings 2026	I, T	source/current images, edit trace	plug-in editor	semantic/perceptual rewards (train); visual state	edit / stop
PSBench [218]	ICML 2026	I, T	GUI state, action trace	Photoshop GUI	—	—
Photographic Retouching and Image Restoration						
JarvisArt [29]	NeurIPS 2025	I, T, (G)	CoT, ROC file	Lightroom (A2L)	format/operation/quality (train); user edits (HITL)	revise ROC / run (HITL)
JarvisEvo [219]	CVPR 2026	I, T	iMCoT, intermediate images	Lightroom tools	preference/human scores (train); self-score	revise
PhotoAgent [220]	ICML 2026	I, (T)	MCTS tree, memory	generative and CV tools	aesthetic labels (train); aesthetic/instruction	select / rollback
PerTouch [221]	AAAI 2026	I, T, (D)	parameter maps, scene memory	diffusion retoucher	map/image pairs (train); intent/result check	adjust / re-edit
RestoreAgent [222]	NeurIPS 2024	I, T	degradations, tool sequence, history	expert pool	tool traces (train); image/history	rollback / reorder / stop
AgenticIR [223]	ICLR 2025	I	schedule, experience	expert pool	IQA labels (train); reflection	rollback / reschedule
MAIR [224]	IJCV 2026	I, T	three-stage schedule, tool registry	expert pool	degradation checks	advance / retry / select
Domain-Grounded Selection [225]	arXiv 2026	I, T, G	probe, candidates, evaluator trace	generative editor	physics, preservation checks	retry / filter / select
JarvisIR [226]	CVPR 2025	I, T	task–model sequence	weather-restoration experts	CleanBench, IQA ranks (train)	controller optimization (train)
HybridAgent [227]	CVPR 2026	I, T	fast/slow route, tool history	single/mixed tools	route/stop labels (train); stop check	continue / stop
4KAgent [30]	NeurIPS 2025	I	degradation profile, plan	expert pool	HPSv2, no-reference IQA	select / rollback / replan
OPERA [228]	arXiv 2026	I	tool-composition plan	co-trained tools	IQA, plan, tool loss (train)	planner/tool optimization (train)
SEAR [229]	arXiv 2026	I	P-MCTS tree, episodic memory	expert pool	no-reference reward, MLLM ranking	select / update memory
Visual Content Transformation: Within-Clip Visual Content Editing						
Aurora [230]	arXiv 2026	V, T, (R)	condition plan	search, segmentation, video DiT	plan/reference/tool preferences (train)	agent optimization (train)
RIVER [231]	arXiv 2025	V, T	digital twin, edit plan	perception, diffusion editor	reasoning/generation rewards (train); execution	refine reasoning
Visual Content Transformation: Long-Form Remaking						
Soap2Soap [232]	arXiv 2026	V, R	JSON screenplay, visual anchors	keyframe/video pipeline	identity, stability, alignment	regenerate shots/windows
Timeline Composition: Non-Linear Editing						
LAVE [31]	IUI 2024	V, D	narrations, timeline plan	NLE tools	preview, adjustment (HITL)	revise timeline (HITL)
EditDuet [32]	SIGGRAPH 2025	V, T	draft timeline, critic trace	NLE environment	request-alignment critique	revise / render
VideoAgent [233]	arXiv 2026	V, T, (A)	storyboard, agent graph	editing agents/tools	graph and intent checks	revise graph
Crayotter / GRPB [234,235]	arXiv 2026	T, (V)	coverage pool, blueprint, trace	retrieval, timeline, audio tools	rules/rankings (train); diagnostics	segment credit (train); re-execute
Timeline Composition: Music-Driven Montage						
GLANCE [236]	arXiv 2026	V, T, A	task graph, sub-timelines	NLE timeline	beat, duration, conflict checks	rerun stages / repair regions
CutClaw [237]	arXiv 2026	V, T, A	footage summaries, script	retrieval, NLE assembly	plot, aesthetic, instruction review	accept / reject / replace clip
DIRECT [238]	arXiv 2026	V, T, A	footage index, plan, guidance	retrieval, beam search	guidance validation	revise query / search
Selection and Compression: Highlight Extraction						
AgentVideoTrimming [239]	arXiv 2024	V	clip captions, story plan	filtering, timeline composition	—	—
DIAMOND [33]	REALM 2025	S, T	play scores, rankings	timestamp retrieval, timeline	preferences, narrative constraints	re-rank plays
Selection and Compression: Video Summarization						
CineAgents [240]	arXiv 2026	V, T	narrative memory, blueprint/script	shot retrieval, video tools	grounding, script validation	revise blueprint / shots
Prompt-Driven [241]	arXiv 2025	V, T	semantic index, storyboard	retrieval, rendering	consistency, boundary checks	refine index / adjust cuts

3.2. Agentic Visual Editing

Agentic visual editing operates on existing images or videos. The agent interprets an editing objective, determines the extent of the required intervention, and carries it out while preserving content that should remain unchanged. The central problem is therefore not simply how to produce the desired result, but how to control where and how the source is altered. We organize this spectrum into two categories. Image editing, retouching, and restoration (Section 3.2.1) operate primarily on individual images and their spatial content, while video editing, remaking, montage, and summarization (Section 3.2.2) extend editing to temporally organized shots, segments, and timelines.

3.2.1. Image Editing, Retouching, and Restoration

Image editing agents modify an existing image while preserving content outside the intended change. Figure 4 summarizes the corresponding runtime loop. Given a source image and, when applicable, an editing request, the agent first diagnoses the relevant degradation or mismatch. It then formulates an editing plan, translates that plan into an ordered sequence of tool calls, and inspects the intermediate result. This evaluation determines whether the edit is complete or whether the workflow should revisit an earlier stage through retry, rollback, or replanning. The lower part of the figure groups methods according to the type of editing objective they address. Instruction-guided editing grounds a natural-language request in localized operations or structured subproblems. Photographic retouching translates aesthetic intent into professional tool settings and editing actions. Image restoration instead identifies the degradation affecting the source and assembles the tools

 doi:10.20944/preprints202609.0682.v1

needed to correct it. A separate training-time branch covers methods that learn planning or evaluation from demonstrations and feedback. It is shown independently because learned components and runtime revision are complementary design choices, rather than features that necessarily appear together in every system.

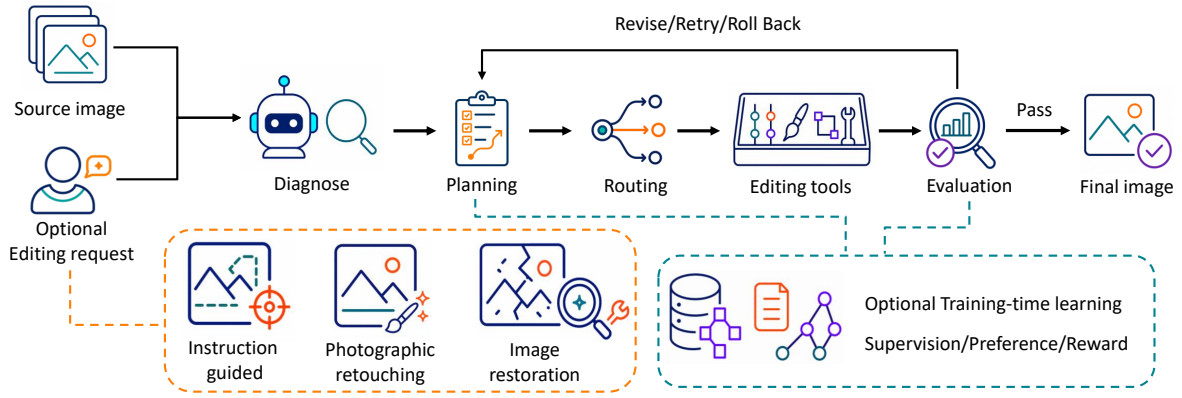

Figure 4. Schematic workflow of agentic image editing. A source image, optionally accompanied by an editing request, passes through diagnosis, planning, routing, tool execution, and evaluation before a final image is accepted. The workflow preserves untargeted source content, while intermediate-result evaluation may trigger retry, rollback, or replanning through the return path. The lower groups distinguish instruction-guided editing, photographic retouching, and image restoration from optional training-time learning based on supervision, preferences, or rewards. Solid arrows denote the main inference path and its evaluation loop, while dashed connections organize task families and training influences. Individual systems may instantiate only part of this workflow.

Instruction-guided image editing.

Natural-language requests often express the desired effect without fully defining its scope or the boundary of what should remain unchanged. Multi-turn interaction allows an agent to resolve this ambiguity as the edit proceeds. ChatEdit [28] provides an early example in facial image editing. It tracks requests across turns but applies each new operation to the original image rather than the preceding result, limiting error accumulation and attribute forgetting. IMAGAgent [209] makes the control process more explicit by extracting constraints from the instruction, constructing a compliant sequence of tool calls, and revising later operations after inspecting intermediate outputs.

Even a well-specified plan can introduce local artifacts or alter content outside the target region. EditRefiner [210] addresses this problem through repeated diagnosis, correction, and evaluation. Its perception stage predicts saliency maps for visible defects, after which the reasoning stage converts the localized evidence into human-aligned diagnoses. An action agent applies a focused correction, and an evaluator stops the process when another iteration no longer improves the result. CAMEO [211] extends the same correction pattern to multi-constraint conditional editing. Its Strategic Director activates the constraints relevant to the current request and introduces reference grounding when needed. A Quality Critic then converts the remaining deviations into instructions for a Refinement Editor until the required criteria are met or the iteration budget is exhausted.

Other methods treat decomposition itself as a learned editing policy. ImageEdit-R1 [212] parses the request and source image into tuples with action, subject, and goal fields. A sequencing agent orders the resulting subrequests, and the complete sequence conditions a diffusion-based editor. Group Relative Policy Optimization (GRPO) updates only the decomposition policy, rewarding both valid output structure and accurate tuple fields. GMO-E^{2}DIT [215] keeps the plan revisable during execution. A VLM converts an underspecified request into a region-grounded edit agenda, which is then compiled into operation-aware masks and optional references. After each operation, the system decides whether to accept the current progress, continue editing, or return to an earlier state. This allows correct partial edits to be retained while harmful changes are discarded. ImageEdit-R1 therefore

 doi:10.20944/preprints202609.0682.v1

commits to a structured plan before execution, whereas GMO-E^{2}DIT updates the execution path in response to intermediate results.

As the editing horizon grows, maintaining a stable working state becomes as important as planning the next operation. Agent Banana [213] separates global planning from localized execution. Context Folding compresses the interaction history into asset-, execution-, and planning-level schemas, Image Layer Decomposition isolates high-resolution regions, and an image-state graph retains prior results. A failed quality test can therefore trigger retry, rollback, or replanning. From Plans to Pixels [214] instead learns its decomposition and orchestration policies. Its planner acquires atomic subgoal decomposition through checklist-guided self-distillation. Its orchestrator learns tool and region selection from a VLM judge that scores instruction adherence, identity preservation, and visual quality, with infeasible subtasks pruned before retraining. At inference time, the orchestrator executes its top-k tool–region candidates for each subtask, and a verifier distilled from the judge advances the highest-scoring edit. Agent Banana relies on explicit state to recover after a failed operation, while From Plans to Pixels learns to choose among candidate executions before committing to the next state.

Long-horizon systems also differ in how they explore alternative editing trajectories. MSRAMIE [216] performs explicit inference-time search through iterative interactions between an MLLM Instructor and a plug-in editing Actor, building a Tree-of-States for alternative trajectories and traceback and a Graph-of-References for cross-state information aggregation, while every state retains access to the original input. MIRA [217] avoids an explicit search topology and instead follows receding-horizon control. A trained VLM observes both the original and current images, issues one atomic instruction to an interchangeable editor, and repeats from the updated result. A jointly trained termination controller decides when the process should stop. The action policy is trained through supervised fine-tuning and GRPO. MSRAMIE therefore preserves alternatives as an explicit search structure, whereas MIRA learns a stepwise control policy that replans after every edit.

GUI execution provides another route for instruction-guided editing by treating professional software as the agent's action space. PSBench [218] evaluates agents that translate editing requests into layer-aware Photoshop action trajectories across canvas, layer, and filter operations. Its success and non-destructive editing metrics are post-hoc benchmark evaluations rather than runtime feedback.

Photographic retouching and image restoration.

Photographic retouching and image restoration both rely on coordinated tool use, but they begin from different questions. Retouching asks how an aesthetic intention should be translated into controllable adjustments. Restoration first identifies what has degraded the image and then decides how to remove it without damaging content that should be preserved. Agentic systems model both expert workflows through explicit planning, tool execution, and visual assessment. JarvisArt [29] trains a multi-modal large language model (MLLM) to translate user intent into a retouching operation configuration (ROC) file. The model is first supervised on chain-of-thought (CoT) annotated samples and then optimized with Group Relative Policy Optimization for Retouching (GRPO-R) using format, operation-accuracy, and perceptual-quality rewards. Because the model does not observe intermediate edited images during reasoning, runtime correction remains user-driven through transparent changes to the ROC file. JarvisEvo [219] closes this loop with interleaved multimodal chain-of-thought (iMCoT), where each tool invocation returns an intermediate image for self-evaluation. Its Synergistic Editor–Evaluator Policy Optimization (SEPO) derives intrinsic editor rewards from self-evaluation while calibrating the evaluator with human-annotated scores. PhotoAgent [220] follows a search-based approach to the same problem. It preserves the editing history as working memory and uses Monte Carlo tree search (MCTS) to explore possible action sequences. A reward model trained on human aesthetic judgments assesses the resulting candidates. When the best candidate does not improve on the preceding state, the system rolls back rather than allowing a weak edit to enter the subsequent workflow.

PerTouch [221] moves from professional tool sequences to semantic and personalized control. Its VLM agent distinguishes weak instructions, which it completes from scene-conditioned preference

 doi:10.20944/preprints202609.0682.v1

memory, from strong instructions that specify a region, attribute, and strength. It converts either form into spatial parameter maps for colorfulness, contrast, color temperature, and brightness. These maps condition a ControlNet-based retoucher trained with semantic replacement and parameter perturbation for regional boundary perception. When the requested degree remains ambiguous, the agent compares the edited image with both the source and the instruction, adjusts the regional control values, and repeats the operation. Information retained from earlier sessions helps the system estimate user preferences when editing later images.

Image restoration poses a related but more uncertain decision problem because the degradation itself must first be inferred. A poor early choice can introduce artifacts that later operations cannot easily undo. RL-Restore [242] uses deep Q-learning to compose twelve specialized convolutional restoration tools under a PSNR-improvement reward and jointly fine-tunes the tools along the learned toolchains. RestoreAgent [222] expands this formulation to more degradations, tasks, and expert models. It trains on optimal pipelines found by exhaustive comparison, along with erroneous and post-rollback trajectories that teach the policy to continue, roll back, or stop. AgenticIR [223] makes this adaptive cycle explicit at inference time. A fine-tuned VLM identifies degradations and evaluates intermediate results, while an LLM builds an experience-guided tool schedule. Harmful operations return the system to the preceding state with a revised schedule. A separate self-exploration stage evaluates alternative tool sequences and summarizes their success statistics in documents that guide subsequent inference-time scheduling.

Several methods reduce the restoration search space by introducing stronger priors. MAIR [224] uses a real-world degradation prior that reverses the assumed image formation process, addressing compression, imaging, and scene degradations in that order. After each operation, DepictQA checks whether the target degradation has been sufficiently reduced. A failed candidate leads to another model, while pairwise comparison retains the strongest result when none meets the target threshold. Domain-Grounded Candidate Selection [225] specializes this strategy to shadow removal. A shared shadow-formation prior grounds both the generative editor and the evaluator, allowing the system to retry failed candidates while balancing shadow removal against preservation of the surrounding scene. JarvisIR [226] instead learns a VLM controller for coupled adverse-weather restoration in autonomous-driving imagery. At inference, the controller handles task planning, model selection, execution, and response generation.

Other systems reduce planning cost by deciding when more extensive deliberation is necessary. HybridAgent [227] reduces the cost and error propagation of iterative restoration by routing explicit requests through a lightweight FastAgent and ambiguous requests through a fine-tuned MLLM SlowAgent. A separately trained FeedbackAgent uses the restored image and tool history to decide whether restoration should terminate, while a mixed-distortion tool removes coupled degradations jointly. 4KAgent [30] constructs a restoration pipeline for each input. Its Perception Agent combines image-quality tools with VLM reasoning to diagnose the degradations affecting the image. OPERA [228] enlarges the planning space by generating a complete tool-invocation plan in one forward pass. Group Relative Policy Optimization (GRPO) trains this policy on final-image restoration quality with auxiliary rewards for degradation prediction, output format, and reasoning–action consistency. The tools are jointly fine-tuned under agent-generated plans to cooperate under sequential composition.

Recent work further treats accumulated restoration experience as an explicit state that can evolve without updating model parameters. SEAR [229] operationalizes this idea through an Intuitive Executor and a Deliberate Planner. The executor retrieves trajectories from an episodic memory indexed by degradation-aware state fingerprints and accepts them only after they pass a reward gate. The planner combines an LLM-generated macro agenda with Pruning-Aware Monte Carlo Tree Search (P-MCTS), using a hybrid no-reference reward and an MLLM pairwise tournament to reduce metric exploitation. Successful trajectories are then written back to memory, allowing the system to refine its external experience from executed trajectories without updating agent parameters.

 doi:10.20944/preprints202609.0682.v1

Table 3 distinguishes three roles that feedback can play in these systems. Training-only methods learn decomposition, tool assignment, or complete compositions without inspecting deployed outputs. Inference-only methods use fixed evaluators, intermediate images, or external experience to revise the current trajectory. Hybrid methods train a policy, retoucher, verifier, or evaluator that also participates in runtime revision. Only the latter two form deployed feedback loops. Evaluator reliability is therefore critical when its judgment controls regeneration, rollback, rescheduling, or termination, whereas errors in a training reward affect the policy before deployment. Over longer horizons, the central difficulty is maintaining a stable account of the intended correction and the content that must remain protected. A locally successful restoration step may otherwise invalidate decisions that earlier stages had already accepted.

3.2.2. Video Editing, Remaking, Montage, and Summarization

Video editing agents modify and reorganize existing footage from within-frame content to shots, segments, and timelines. Figure 5 organizes this range by the primary transformation applied to source material. Visual content transformation changes rendered content while retaining the source's temporal scaffold. Timeline composition changes clip selection, duration, and order without necessarily altering pixels. Selection and compression shorten long footage while preserving salient events or narrative structure. These operations define six task categories: within-clip visual content editing and long-form remaking transform source content; non-linear editing (NLE) and music-driven montage compose timelines; highlight extraction selects salient moments; and video summarization compresses long footage into a coherent short-form account. Across these tasks, the agent grounds an editing objective in the relevant content and time span, keeps track of the evolving edit, and inspects intermediate results before deciding what should happen next. The upper path in Figure 5 traces how a source video, optionally paired with an editing request, moves from analysis through planning and execution to evaluation. The lower group captures temporal state, human guidance, and training-time learning. The return path applies only to systems that revise content, timeline structure, or selection after evaluation.

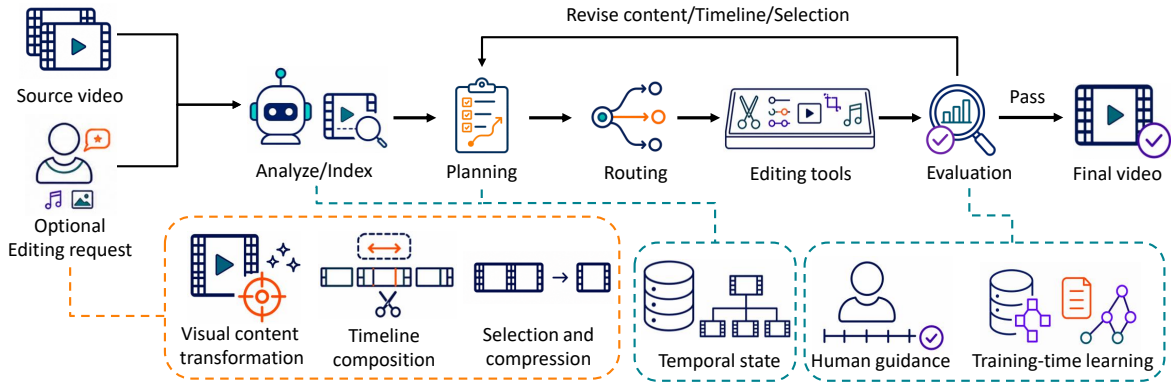

Figure 5. Schematic workflow of agentic video editing. A source video, optionally accompanied by an editing request, passes through analysis and indexing, planning, routing, tool execution, and evaluation before a final video is accepted. Evaluation may trigger revisions to content, timeline structure, or selection. The lower groups distinguish visual content transformation, timeline composition, and selection and compression, while also highlighting temporal state, human guidance, and training-time learning. Solid arrows denote the main inference path and its evaluation loop, while dashed connections organize task families and supporting factors. Individual systems may instantiate only part of this workflow.

Within-clip visual content editing.

Within-clip visual content editing changes what appears in an existing clip without reorganizing its temporal structure. The agent must localize the requested intervention across frames and prevent it from affecting content that should remain unchanged. Aurora [230] treats underspecified requests as a condition-construction problem. Its VLM agent produces a structured plan containing a rewritten

 doi:10.20944/preprints202609.0682.v1

instruction, a task label, an optional image-search query, and an optional mask phrase. Web image search and grounded segmentation supply missing appearance references and spatial masks. RIVER [231] addresses implicit requests by constructing a digital twin of object semantics, spatial locations, depth, masks, and temporal trajectories. An LLM performs multi-hop reasoning over this representation to instruct a diffusion-based editor, while reinforcement learning (RL) rewards reasoning accuracy and generation quality. Both systems transform content within a source video rather than rearranging its timeline.

Long-form video remaking.

Long-form video remaking applies content transformation across an entire episode or film rather than to a localized clip. The challenge is to restyle or recast hundreds of shots while preserving how the source narrative unfolds and how recurring characters move and appear. Soap2Soap [232] conditions this process on target appearance references for the main characters. A scene-aware screenplay represented in JavaScript Object Notation (JSON) and dynamically allocated visual anchors maintain series-level consistency. Batched keyframe generation reduces identity drift before video synthesis, while a verification agent audits identity, stability, and contextual alignment and selectively regenerates failed shots or time windows.

Non-linear editing.

Non-linear editing reorganizes existing footage into a coherent timeline. A locally plausible cut may still weaken the sequence as a whole, so agents must reason about both individual clips and their narrative relations. LAVE [31] generates language descriptions of user footage and lets an LLM plan and execute edits while keeping the timeline under direct user control. EditDuet [32] automates more of this process. Its Editor searches the footage and revises the timeline, while a Critic reviews successive drafts and triggers rendering once the result is satisfactory. VideoAgent [233] broadens the operation set for instructions accompanied by raw clips, audio tracks, and auxiliary documents. It filters more than thirty specialized agents by intent and composes them into a directed graph. Textual-gradient optimization then checks acyclicity, connectivity, intent coverage, and tool compatibility, providing inference-time verification of the workflow representation rather than the rendered video. Crayotter [234] treats intermediate artifacts as explicit production state shared across planning, research, and execution roles. Artifact-level diagnostics localize failures so that only the affected segment needs to be re-executed. Building on the same environment, the GRPB-trained Crayotter model [235] converts same-task rankings of final videos into zero-sum advantages and assigns bounded credit to earlier semantic editing segments using a lagged Bradley–Terry allocator. This links subjective judgments of the final result to the editing decisions that produced it without requiring score calibration across tasks.

Music-driven montage.

Music-driven montage uses music as an external temporal structure for arranging footage. The agent must align the developing visual sequence with musical form and affect, rather than merely match individual cuts to beats. GLANCE [236] coordinates this process through two nested loops. The outer loop constructs a music-aware editing graph, while the inner loop repeatedly executes and verifies local edits under a preventive context controller. Once the sub-timelines have been assembled, a diagnostic agent identifies conflicts across segments, and only the affected parts are re-edited. CutClaw [237] scales this process to hours-long footage by decomposing both the video and soundtrack hierarchically. Its Playwriter aligns narrative development with musical sections, after which Editor and Reviewer agents retrieve candidate segments and validate them before insertion. DIRECT [238] connects global structure with local shot selection through a hierarchy of Screenwriter, Director, and Editor agents. It converts the musical plan into segment-level retrieval queries and pacing constraints, then searches for a valid shot sequence. If the validator rejects all candidates, the Director revises the query and repeats retrieval.

 doi:10.20944/preprints202609.0682.v1

Highlight extraction.

Highlight extraction selects salient moments from long footage without altering their visual content. The main challenge is to consider not only the value of each interval, but also how the selected moments relate when placed together. Agent-based Video Trimming (AVT) [239] argues that conventional highlight detection and moment retrieval often evaluate intervals in isolation. Its agent therefore decides jointly which segments to retain and how to order them. DIAMOND [33] applies this idea to baseball using play-by-play logs rather than visual or audio signals. It converts the logs into structured game events with contextual statistics, uses an LLM to assess their significance, and reranks the candidates according to configurable priorities before retrieving the corresponding clips. Since these preferences affect inference-time ranking rather than a learned user model, DIAMOND provides interpretable preference conditioning rather than learned personalization.

Video summarization.

Video summarization compresses long footage into a coherent short-form account. Selection must preserve long-range narrative relations while removing material that does not contribute to the resulting story. CineAgents [240] reconstructs the source as a hierarchical narrative memory. A Director proposes a story blueprint, and an Orchestrator grounds each stage in evidence retrieved from that memory. Unsupported proposals are returned for revision before a Manager validates the script and an Editor assembles the selected shots. Feedback therefore changes the blueprint and shot sequence before rendering rather than evaluating only the completed video. Prompt-Driven Video Editing [241] scales summarization to hours-long, story-driven footage by building a persistent semantic index from temporal segmentation and compressed multi-scale memory. It refines its synopsis and scene records to resolve missing identities, contradictions, and unsupported details. Specialized agents then turn a free-form request into a timestamped recap plan and rendered video. A micro-cut agent further adjusts clip boundaries using ElevenLabs transcription.

Agent roles vary by transformation level: content agents ground targets across frames, timeline agents connect shot semantics to pacing and order, and selection agents balance local salience with narrative coherence. The systems also differ in their use of feedback. Aurora relies on training rewards without runtime revision, whereas RIVER also uses code-execution results to refine digital-twin reasoning before a single diffusion edit. Soap2Soap, EditDuet, Crayotter, GLANCE, CutClaw, and Prompt-Driven modify outputs or trajectories through runtime verification. DIAMOND re-ranks event candidates during inference, while VideoAgent and CineAgents revise pre-render agent graphs or narrative blueprints. LAVE retains human input during timeline refinement. AVT provides post-hoc assessment without an architectural feedback loop. In long-form settings, compressed representations may still omit subtle events that later decisions require.

3.3. Agentic Visual Composition

Agentic visual composition transforms heterogeneous content elements and design intent into explicit, layout-sensitive artifacts. Unlike agentic visual generation and editing, its main concern is not to create or alter perceptual content, but to decide how information should be selected and arranged so that the result remains legible and editable. We classify this area by the structural scope of the composition state being controlled: graphic and document composition (Section 3.3.1) organizes mostly static content on a single canvas, whereas presentation and interface composition (Section 3.3.2) coordinates content across multiple pages, screen regions, interactive states, or related media while preserving global consistency. Figure 6 summarizes the shared high-level pipelines of visual composition and visual programming.

 doi:10.20944/preprints202609.0682.v1

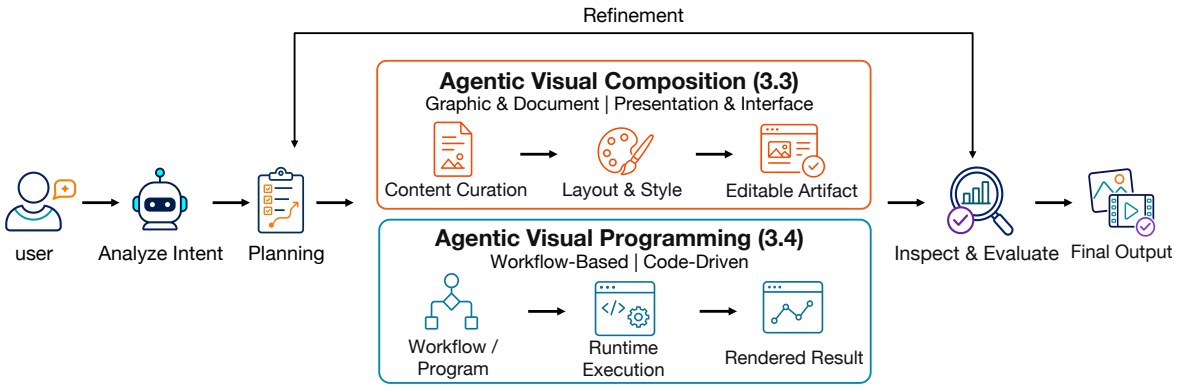

Figure 6. Overview of agentic pipelines for visual composition and visual programming. The two paradigms are presented in a common high-level sequence of intent analysis, planning, production, inspection, and refinement. In visual composition, agents curate content and determine layout and style to produce editable artifacts. In visual programming, agents construct workflows or programs whose runtime execution produces rendered results. Evaluation may trigger branch-specific refinement of the composition state or the executable procedure before the final visual output is accepted.

3.3.1. Graphic and Document Composition

Graphic and document composition covers static artifacts organized primarily on a single canvas, such as posters, banners, and text-heavy documents. The agent works through an editable representation that records the content structure and its spatial arrangement. This representation gives the agent a concrete action space for inspecting and revising content, layout, style, layers, or components.

Early pipeline-based systems establish much of this structural foundation. AutoPoster [243], PosterLLaVa [244], and POSTA [245] predict structured content, layout, or typography before rendering. By making document elements explicit and addressable, they support controlled composition, but offer only limited support for inspecting a completed design and deciding how to repair it.

More agentic systems assign composition decisions to roles and, increasingly, close the rendering loop. PosterGen [246] distributes paper parsing, content curation, layout, style, and PPTX rendering across specialist agents, although its VLM rubric evaluates completed posters rather than controlling another generation round. Paper2Poster [34] closes this loop. A Parser builds an asset library, a Planner constructs a binary-tree layout, and a Painter-Commenter loop executes panel-rendering code and corrects overflow or misalignment. PosterMELD [247] similarly uses capacity-aware template slots, deterministic gates, and VLM review to route failures into bounded repair before exporting editable PPTX and PNG artifacts. Agentic control therefore comes not from role naming alone, but from assigning a detected defect to the document state or composition stage responsible for correcting it.

Tool-grounded document states extend this control beyond academic posters. Any2Poster [248] turns heterogeneous sources into editable HTML/CSS panels and can translate panel-level diagnoses into localized edits. PSDesigner [249] alternates asset collection, graphic planning, and tool execution over the current PSD state. BannerAgency [250] organizes its production roles around a foreground blueprint that reviewer feedback can update. CAL-RAG [251] combines retrieved exemplars with a layout recommender, a threshold-based vision-language grader, and a feedback agent to revise structured layouts after rejection. Although their tools differ, these systems primarily exercise control through explicit document states rather than undifferentiated pixel-level repainting.

Diagram and figure composition applies this control at the object level. EvoDiagram [252] coordinates agents over an editable canvas and accumulates reusable design knowledge, whereas SciFig [253] assigns extraction, XML layout, component construction, and feedback to specialized agents. In graphic and document systems, rendered images serve as observations, while repair actions remain grounded in responsible content items, regions, layers, or components.

 doi:10.20944/preprints202609.0682.v1

3.3.2. Presentation and Interface Composition

Presentation and interface composition extends visual composition beyond a single canvas to a collection of related views. The agent must preserve a shared structure and visual language while managing the state of each view locally, so that revising one part does not weaken the coherence of the whole artifact.

Slide systems realize this global-to-local control through different executable representations. PPTAgent [35] converts an outline and reference schemas into editing actions and self-corrects failed executions; PreGenie [254], DeepPresenter [255], and SeaSlides [256] instead coordinate content, code, and rendered-page inspection in Slidev, HTML, or Typst environments. Their implementation details differ, but the shared agentic mechanism is to expose local build or visual failures while retaining deck-level structure.

Persistent state makes later interaction more targeted. MemSlides [257] separates user-profile, working, and tool memory, allowing a new instruction to be scoped to the affected slide region without discarding deck-level preferences. PrototypeFlow [258] exposes the current design at several levels but leaves selective regeneration under human control. GameUIAgent [259] automates more of this process by instantiating a structured design specification in Figma and accepting a VLM-guided revision only when it does not degrade the current design.

Other systems coordinate content across media or deliverables. PresentAgent-2 [260] plans research, retrieved media, slides, narration, and speech as a presentation video. PaperX [261] shares a Scholar DAG across PPT, poster, and promotion modules, while OmniPresent [262] uses a canonical knowledge stream and cross-artifact verification before format-specific layout and styling. Here, agentic control lies in deciding where information belongs and propagating shared structure or corrections without losing global coherence. Table 4 summarizes these states and feedback channels.

Table 4. Comparison of systems for agentic visual composition. Systems are grouped by the structural scope of their composition state into graphic and document composition (single-canvas artifacts) and presentation and interface composition (multi-page, multi-screen, or interactive artifacts). *Intermediate Artifact* denotes the explicit composition state maintained before rendering. Agentic-control tags denote planning (Plan), multi-agent coordination (Multi), tool use (Tool), inference-time feedback (Loop), and memory (Mem).

Method	Venue	Output	Intermediate Artifact	Agentic Control	Feedback Signal
Graphic and Document Composition					
AutoPoster [243]	ACM MM 2023	advertising poster	poster layout, tagline, style attributes	—	—
PosterLLaVa [244]	arXiv 2024	editable SVG poster	JSON layout specification	—	—
POSTA [245]	CVPR 2025	artistic poster	layout and typography plan	—	—
PosterGen [246]	CVPR Findings 2026	PPTX poster	storyboard, layout plan, style metadata	Multi	VLM rubric (evaluation only)
Paper2Poster (PosterAgent) [34]	NeurIPS 2025	editable PPTX poster	asset library, layout tree, rendering code	Multi, Loop	VLM commenter (panels)
PosterMELD [247]	arXiv 2026	editable PPTX, PNG	capacity-aware template and slot state	Multi, Loop	gated VLM repair
Any2Poster [248]	arXiv 2026	poster (HTML/CSS)	parsed source, content and layout plan	Plan, Tool, Loop	panel-level VLM (optional)
PSDesigner [249]	CVPR 2026	editable PSD	PSD state and tool calls	Multi, Tool, Loop	rendered design / layer state
BannerAgency [250]	EMNLP 2025	Figma / SVG banner	foreground blueprint, editable components	Multi, Mem, Loop	reviewer critique
CAL-RAG [251]	arXiv 2025	poster layout	retrieved exemplars, JSON layout	Multi, Loop	grader rejection
EvoDiagram [252]	arXiv 2026	editable diagram	object-level canvas schema	Multi, Tool, Loop, Mem	rendered-canvas VLM diagnosis
SciFig [253]	arXiv 2026	editable XML figure	layout plan and XML figure	Multi, Loop	human / VLM
Presentation and Interface Composition					
PPTAgent [35]	EMNLP 2025	slide deck (PPTX)	outline, slide schema, code actions	Plan, Tool, Loop	action-execution self-correction
PreGenie [254]	EMNLP 2025 Findings	Slidev deck	Slidev Markdown code	Multi, Loop	code and rendered-slide review
DeepPresenter [255]	ACL Findings 2026	HTML deck	manuscript, design plan, HTML slides	Multi, Loop	environment-grounded reflection
MemSlides [257]	arXiv 2026	slide deck	hierarchical memory and deck state	Mem, Loop	multi-turn slide-local revision
SeaSlides [256]	arXiv 2026	HTML / Typst deck	semantic HTML / Typst content	Tool, Loop	build, constraint, visual
PresentAgent-2 [260]	arXiv 2026	presentation video	slide plan, media plan, narration structure	Plan, Tool	—
PaperX [261]	arXiv 2026	PPT, poster, promo suite	Scholar DAG	Plan, Multi	—
OmniPresent [262]	arXiv 2026	slides, poster, video, page	renderable HTML suite representation	Multi, Loop	cross-format verification
PrototypeFlow [258]	TOCHI 2026	UI prototype	UI prototype and intermediate guidance	Loop	designer (human)
GameUIAgent [259]	arXiv 2026	Figma UI design	Design Spec JSON	Tool, Loop	VLM reflection

3.4. Agentic Visual Programming

Agentic visual programming refers to systems in which an agent maps visual intent, source assets, or runtime state to an executable production procedure whose execution yields the visual artifact. Unlike direct prompting, this representation gives the agent direct control over the artifact's structure and behavior while keeping the underlying decisions inspectable and revisable. A planning or coding agent therefore constructs an intermediate program, invokes a runtime, and uses validation, execution traces, or rendered evidence to select the next repair, so that program-level edits converge on the intended visual result. We distinguish workflow-based systems (Section 3.4.1), which manipulate node

 doi:10.20944/preprints202609.0682.v1

graphs in a bounded tool ecosystem, from code-driven systems (Section 3.4.2), which modify scripts or domain-specific specifications under program semantics. Table 5 summarizes this comparison.

Table 5. Comparison of systems for agentic visual programming. Systems are grouped by executable substrate. *Executable Substrate* combines the representation and runtime using standardized labels. Agentic-control tags denote planning (Plan), multi-agent coordination (Multi), tool use (Tool), inference-time feedback (Loop), memory (Mem), knowledge guidance (Know), and search.

Method	Venue	Output	Executable Substrate	Agentic Control	Feedback Signal
Workflow-Based Visual Programming					
ComfyAgent [36]	CVPR 2025	image / video	workflow code (ComfyUI interpreter)	Plan, Multi, Tool	execution success (post-hoc)
ComfyGen [263]	arXiv 2024	image	workflow graph (ComfyUI)	—	offline preference / evaluation
ComfyUI-Copilot [264]	ACL 2025 Demo	image / video	workflow graph (ComfyUI)	Multi, Tool	recommendation and user feedback
ComfyGPT [265]	arXiv 2025	image / video	workflow graph (ComfyUI)	Multi, Tool, Loop	execution and instruction checks
ComfyUI-R1 [266]	ACL Findings 2026	image / video	workflow graph (ComfyUI)	Plan	rule–metric reward (training)
Knowledge-Centric Agents [267]	ECCV 2026	image	workflow graph (ComfyUI)	Know	structural and execution evaluation
ComfyMind [268]	NeurIPS 2025	image / video	workflow graph (ComfyUI)	Plan, Search, Loop	localized execution feedback
ComfySearch [269]	arXiv 2026	image / video	workflow graph (ComfyUI)	Search, Tool, Loop	graph validation and execution
COMFYCLAW [270]	arXiv 2026	image	workflow graph (ComfyUI)	Mem, Tool, Loop	rollback, VLM repair, skill evolution
Code-Driven Visual Programming					
PairCoder++ [271]	arXiv 2026	charts, figures, CAD, 3D	artifact code (compiler / renderer / simulator)	Multi, Tool, Loop	diagnostics, execution, rendering
GVR-Coder [272]	ACM MM 2026	SVG diagram	vector code (SVG renderer)	Loop	attributed visual feedback
Feynman [273]	arXiv 2026	domain diagram	diagram program (Penrose)	Plan, Loop	visual refinement
LiveFigure [274]	ICML 2026	editable vector figure	drawing script (PowerPoint)	Plan, Tool, Loop	targeted visual diagnostics
SceneCraft [37]	ICML 2024	3D scene	scene script (Blender)	Plan, Tool, Loop, Mem	rendered VLM feedback
SceneCode [275]	arXiv 2026	articulated 3D scene	scene script (Blender / physics simulator)	Plan, Multi, Tool, Loop	execution-guided repair
HDSL [276]	arXiv 2026	3D indoor scene	scene DSL (asset / layout runtime)	Plan, Multi, Tool, Loop	bounded verification; local repair
MoReGen [277]	CVPR 2026	physics-grounded video	simulation code (simulator / renderer)	Multi, Tool, Loop	evaluator / trajectory feedback
VideoCoCo [278]	arXiv 2026	physics-consistent video	scene script (Blender / video editor)	Tool	deterministic rendered draft
GEST Authoring Agents [208]	arXiv 2026	narrative video	event graph (GEST engine)	Plan, Multi, Tool	state-constrained validity
MoVer [279]	ACM TOG 2025	SVG motion graphics	motion DSL (SVG verifier)	Tool, Loop	failed-predicate report
MapStory [280]	UIST 2025	map animation	animation blocks (Mapbox / Fabric.js)	Plan, Multi, Tool	user-guided revision
Keyframer [38]	VL/HCC 2025	SVG animation	animation code (browser)	—	human prompt / code / property edits
LogoMotion [281]	CHI 2025	logo animation	animation code (browser)	Loop	visual checking and program repair
ManimAgent [282]	arXiv 2026	educational animation	animation code (Manim)	Multi, Mem, Loop	VLM keyframe scores; episodic memory

3.4.1. Workflow-Based Visual Programming

Workflow-based visual programming represents a production procedure as an executable graph within a bounded environment such as ComfyUI. The agent selects nodes and models, configures typed connections, and executes partial or complete graphs; validation, runtime errors, or rendered output then guide node replacement, dependency repair, parameter revision, or backtracking.

Existing systems range from graph prediction to explicit task allocation. ComfyGen [263] predicts prompt-adaptive flows without inference-time repair and therefore provides a non-repair baseline. ComfyAgent [36], ComfyUI-Copilot [264], and ComfyGPT [265] instead divide planning, retrieval, graph construction, refinement, and execution among named agents. Despite architectural differences, their shared contribution is to make the workflow plan inspectable and executable under node availability and interface constraints.

Another line makes workflow reasoning more structured before execution. ComfyUI-R1 [266] learns node selection and workflow planning from reasoning traces and rule-metric rewards, but is a trained policy rather than a multi-agent repair loop. Knowledge-Centric Agents [267] distill existing workflows into strategies, skeletons, and pseudo-code, enabling the generation agent to move between high-level intent and executable graphs. Both expose more structure than direct graph prediction, although neither contribution alone implies online correction.

Execution-guided systems close this loop by using partial execution to determine what should change next. ComfyMind [268] executes intermediate functional modules so that failure can be assigned locally, while ComfySearch [269] validates partial graphs during component-space exploration. COMFYCLAW [270] combines typed graph editing with runtime rejection and region-level visual verification. It can return to an earlier state after a failed change and preserve successful procedures as reusable skills. In these systems, execution is not merely a final test. It identifies the subgraph responsible for a failure and provides evidence for the next repair.

3.4.2. Code-Driven Visual Programming

Code-driven visual programming expresses a visual task as source code or a domain-specific executable specification. Compared with pixel-space generation, this representation makes the artifact's structure and behavior explicit, giving the agent finer control and allowing later edits to target the source. Execution success, however, is not sufficient: a program may compile and render without

 doi:10.20944/preprints202609.0682.v1

producing the intended visual result. Code-driven agents therefore execute the program, treat runtime traces or rendered outputs as evidence, and repair the source representation rather than the pixels alone.

For diagrams and figures, source-level structure preserves visual elements as editable, addressable objects. Feynman [273] states diagram content as declarative Penrose programs and lets optimization-based rendering solve the layout, so the same semantics can be re-rendered without breaking spatial relations. LiveFigure [274] emits PowerPoint drawing scripts and thereby produces a vectorized figure whose shapes and text remain adjustable in a familiar authoring tool. This addressability also makes feedback actionable. GVR-Coder [272] traces defects in a rendered SVG to the elements responsible for them. PairCoder++ [271] separates code writing from the review of compiler, execution, and rendering evidence between Driver and Navigator agents, exchanging their roles after persistent failure. In each case, the agent converts a visual diagnosis into an edit of the specific object, constraint, or code region that caused it, so defects can be repaired through source-level changes rather than undifferentiated pixel-level regeneration.

For 3D scenes and physically grounded video, the goal is to maintain geometry, state, and motion consistently across viewpoints and frames. This favors executable specifications that compute spatial relations, physical parameters, and temporal evolution rather than depicting each frame independently. SceneCraft [37] translates spatial relations among up to a hundred assets into numerical constraints in a Blender script, revises the scene from rendered views, and distills recurring solutions into a reusable library. HDSL [276] keeps rooms, regions, objects, and support surfaces in a hierarchical specification with local coordinates, verifies subtrees, and rewrites only the subtree an edit implicates, so resolving one collision does not disturb the rest of the room. SceneCode [275] likewise composes part-wise programs and repairs them from execution evidence, exporting articulated assets that remain valid under physics simulation. For video, MoReGen [277] parses a prompt into physical parameters and coordinates code writing, rendering, and evaluation, so that motion follows from simulated dynamics rather than from learned appearance priors, whereas VideoCoCo [278] executes its program into a deterministic spatiotemporal draft that conditions a generative video editor, carrying the simulated dynamics into a photorealistic result. GEST Authoring Agents [208] constrain event-graph construction through a stateful tool layer that rejects invalid operations and returns explanatory errors for replanning. This creates an authoring-stage agent-environment loop. Rendering remains offline, and the paper does not establish a rendered-video review-and-repair cycle.

For animation, the result is judged over time, and code keeps timing, easing, and motion as named parameters rather than baked pixels so that an incorrect movement can be retimed or otherwise revised at the source level. MoVer [279] synthesizes an animation program alongside a verification program over spatiotemporal predicates, and returns failed predicates to the synthesizer for repair, while ManimAgent [282] scores rendered keyframes and retains successful rationales and validated failure patterns in episodic memory for later tasks. Other systems retain human control over revision. MapStory [280] renders map-animation blocks on an editable timeline with explicit timing and style arguments, while Keyframer [38] presents generated CSS animation next to its rendered preview so that the user drives the next iteration through prompts or direct property edits. LogoMotion [281] closes the loop autonomously by analyzing the source logo, expressing a design concept as JavaScript animation code, and repairing visually detected defects at the program level. Across these systems, the executable representation determines where a correction can be made, while execution evidence indicates what should change.

4. Data and Evaluation

Agentic visual creation involves both producing visual content and coordinating the steps needed to produce it. Evaluation therefore concerns not only the quality of individual images or clips, but also whether the completed work satisfies its instructions, reference conditions, and cross-component constraints.

 doi:10.20944/preprints202609.0682.v1

On the data side, benchmarks supplement text prompts with structured specifications, including storyboards, shot- or event-level descriptions, recurring entity schedules, multi-modal references, audio requirements, and user preferences. These inputs support assessment of outputs under complex production requirements. Training data may further encode cross-shot relationships, record the decisions and tool interactions that precede generation, or derive supervision from reward signals and comparisons between generated outcomes.

On the evaluation side, fine-grained criteria help identify errors that overall quality scores can miss. For images, evaluation distinguishes failures in atomic requirements, reference preservation, and factual correctness, and can separate the contributions of the prompter and renderer. For video, evaluation checks whether identities, states, and relationships remain consistent across shots while allowing changes explained by the action or camera viewpoint. Recent evaluation pipelines combine within-shot and cross-shot assessment and select criteria relevant to the content. They pair VLM judgments with specialist measurements and provide feedback for refinement. Their automatic scores can also be checked against human judgments. Assessing the agent's decisions additionally requires evidence from the production process, rather than final outputs alone.

4.1. Data

Most resources reviewed in this section are benchmarks for image and video generation and editing, including agentic systems. We organize them according to the role their data play in the creation process. Benchmark data define what should be generated through prompts, references, and hierarchical production specifications. Training data supervise either the visual outputs or the agent actions that produce them. Because a single resource may serve both purposes, this distinction is based on data use rather than dataset identity. We first examine benchmark instance design and then training supervision. Evaluation metrics and protocols are discussed in the following subsection.

4.1.1. Benchmark Instance Design

Image Prompts and Requirements.

Several image benchmarks make prompt requirements explicit through verification questions, evaluation rubrics, or reference descriptions. ConceptMix [283] combines one object with k sampled visual concepts and generates one verification question per concept, making difficulty controllable. GenEval 2 [284] decomposes prompts into 3 to 10 visual atoms for atom-level diagnosis and strict prompt-level assessment. Qwen-Image-Bench [285] organizes prompts from professional creation scenarios with hierarchical rubrics. R2I-Bench [286] covers seven reasoning categories, and each instance includes a reasoning-oriented prompt, a reference caption, and an explanation.

Source and Reference Images.

Some image benchmarks include source or reference images alongside textual instructions, specifying visual content to preserve or incorporate into the output. FactIP [85] pairs an instruction with two seed images and metadata for a factual or culturally grounded concept. ICE-Bench [287] distinguishes four creation and editing settings by the presence of source and reference images, with masks used for local edits. MMIG-Bench [288] provides multi-view references for humans, animals, objects, and artistic styles to test identity or style preservation. MultiBanana [289] supports up to eight references and includes domain and scale mismatches, rare concepts, and multilingual text. These benchmarks test whether a model can use the relevant attributes of each reference while following the instruction.

Video Prompt Categories.

Video benchmarks organize prompts by application scenario or by the generation capabilities being tested. AVGen-Bench [290] groups 235 text-to-audio-video prompts into three application domains and 11 categories. T2V-CompBench [291] contains 1,400 prompts across seven compositional categories covering attribute, motion, and action binding, spatial relations, object interactions, and numeracy.

 doi:10.20944/preprints202609.0682.v1

Multi-Shot and Story Descriptions.

Benchmarks for multi-shot and story generation often describe individual shots or events, together with shared characters, reference assets, and continuity requirements. FilmBench [292] derives 1,169 text-to-video and reference-to-video prompts from clips of award-winning films selected by professional directors. Of these, 1,056 contain multiple shots. The prompts use scene, role, and prop tags with shot-level cinematic directions, while reference-conditioned cases provide a scene, prop, or character image. UniVBench [293] pairs 200 multi-shot videos with shot-level captions, editing instructions in several formats, and reference images for six tasks spanning understanding, generation, editing, and reconstruction.

MSAVBench [294] represents synchronized audio-video requests as global-to-shot scripts of up to 15 shots. Each script specifies visual, audio, and cinematic conditions, while conditioned cases may add character, scene, or audio references. MSVBench [295] organizes stories through global character and environment assets, scene segments, and shot conditions. Its shot annotations identify on-screen characters, a reference frame, visual states, actions, and camera movement. EntityBench [296] adds per-shot schedules for characters, objects, and locations to test entity recurrence over sequences of up to 50 shots.

Story-level resources specify how these units serve a longer narrative. ViStoryBench [297] decomposes each story into storyboard shots containing the setting, plot correspondence, onstage characters, visual state, and camera perspective, with shared character descriptions and reference images. LongAV-Compass [298] describes 284 minute-scale audio-video cases through a global description and timed events that record actions, completion criteria, visual elements, audio, and persistent constraints. DirectorBench [299] pairs production metadata for story, shots, camera control, consistency, and audio with user profiles that encode priorities and hard constraints. These specifications define the intended output but do not record the agent decisions or tool interactions used to produce it.

4.1.2. Training Supervision

Cross-Shot Supervision for Visual Generators.

MuSS [300] contains over 30K captioned multi-shot clips and more than 1,000 hours of video drawn from over 3,000 movies. It pairs consecutive shots with coherent, shot-aligned captions. For subject-to-video training, each target clip is paired with an image of the same subject from a separate shot, encouraging identity preservation across changes in pose, viewpoint, and context without using a frame from the target clip.

Interaction Trajectories for Image Agents.

Image agents also need supervision for the process before synthesis. Unify-Agent [85] provides 143K trajectories containing a user instruction, textual and visual research traces, and an evidence-grounded recaption. ToolArtist [133] provides 7,132 supervised trajectories that retain reasoning, text and image search actions, observations, a final visual caption, and generated image tokens. Both datasets expose how an agent gathers evidence and converts it into a generation condition.

Gen-Searcher [301] divides 16K samples into 10K for supervised fine-tuning and 6K for reinforcement learning. The first split teaches multi-turn search, reference selection, and grounded prompt construction. The second optimizes tool use with textual and visual feedback while keeping the image generator fixed. GenEvolve [159] contains 8,800 supervised examples that preserve the full tool loop. It also exports 3,175 filtered image cases without teacher trajectories or final programs for self-evolution and held-out evaluation. Multiple rollouts for the same request are compared, and their differences become structured visual experience for token-level supervision.

WeDataset-MMGenEdit [86] links trajectory collection to stage-specific verification. It contains approximately 23K SFT trajectories and 14.7K RL tasks, with separate checklists for the agentic chain, the generation input, and the final image. Examples that pass all three quality gates are retained for

 doi:10.20944/preprints202609.0682.v1

SFT. Structurally valid examples that fail at least one gate are assigned to RL and tagged by failure type.

Together, these resources cover trajectory imitation, reward-guided search, and supervision derived from generated outcomes. Benchmark data define increasingly structured creation tasks, while training data determine whether learning targets the visual result or the process used to produce it.

4.2. Evaluation

Evaluation of complex visual generation must determine which properties to measure, how to separate different sources of error, and how to turn those judgments into reliable feedback. We first summarize the main evaluation dimensions, then examine automated and agent-based pipelines for images, multi-shot videos, and agentic settings.

4.2.1. Evaluation Dimensions

Conventional quality metrics remain useful, but complex image and video tasks also require evaluation of instruction adherence, reference preservation, and relationships within the generated content. Multi-shot narratives add entity consistency, cinematic control, temporal continuity, and audiovisual coherence. We organize these criteria into four representative dimensions covering consistency and controllability, visual composition and cinematic language, knowledge and reasoning, and audio-visual coordination.

Consistency and controllability.

For image generation and editing, consistency measures whether the output preserves relevant content from source and reference images. Controllability concerns whether the output follows the requested content, edits, and structural conditions. ICE-Bench [287] separates image quality and prompt following from source consistency, reference consistency, and controllability. MMIG-Bench [288] organizes assessment into visual artifacts and identity preservation, aspect-level semantic matching, and aesthetics and human preference. MultiBanana [289] examines failure modes such as subject omission and compositional distortion when several references must be combined.

Multi-shot evaluation must compare generated shots with external references and with one another. MSVBench [295] separates face, character, background, clothes-and-color, and relative-size consistency. EntityBench [296] uses per-shot fidelity to gate cross-shot comparisons, then combines DINOv2 similarity with LLM-based pairwise judgments. ViStoryBench [297] distinguishes cross-similarity to reference images from self-similarity among outputs in the same story for both character identity and style. Similarity to a reference can also reward direct copying. MuSS [300] addresses this ambiguity with ACP-Var, which measures pose variation relative to the reference, and CP-Rate, which detects appearance overfitting through DINOv2 feature similarities. MSAVBench [294] extends consistency assessment to visual content and audio properties across shots. LongAV-Compass [298] adds long-form continuity and diagnoses event collapse in minute-scale generation.

Visual composition and cinematic language.

Evaluation for practical image creation extends beyond generic visual quality. Qwen-Image-Bench [285] supplements Quality, Aesthetics, and Text-Image Alignment with Real-world Fidelity and Creative Generation. Video evaluation must also account for how shots are staged and connected. EvalVerse [302] maps its criteria to pre-production, production, and post-production, covering script and scene coherence, shot composition and camera control, and editing choices such as transitions and color grading. MSAVBench [294] evaluates adherence to specified camera parameters, while MuSS [300] uses TransNetV2 to detect shot boundaries and measure transition timestamp deviation.

Knowledge, reasoning, and compositional accuracy.

This dimension asks whether an image satisfies explicit constraints and correctly renders conclusions that depend on implicit intent or external knowledge. ConceptMix [283] controls compositional

 doi:10.20944/preprints202609.0682.v1

difficulty through the number of sampled concepts and checks each requirement separately. R2I-Bench [286] jointly evaluates reasoning accuracy, text-image alignment, and image quality, while PhyBench [303] focuses on physical commonsense. WISE [304] and WorldGenBench [305] test world knowledge and implicit inference. Mind-Bench [90] examines whether reasoning or retrieval can resolve implicit visual constraints. For search-assisted generation, KnowGen [301] separates prompt faithfulness, visual correctness, text accuracy, and aesthetics. KVBench [306] evaluates knowledge-intensive scientific images through atomic constraints on structure and symbolic accuracy. SciIR-Bench [307] separates scientific reasoning into entity structure, scientific process, and scientific law.

Audio-visual coordination.

Joint audio-video generation requires the soundtrack to match visible events in content and time. AVGen-Bench [290] evaluates audio and video quality together with cross-modal alignment, including general and lip synchronization. LongAV-Compass [298] separates audio-video synchronization, event-level audio quality, and coherence across the full soundtrack. MSAVBench [294] further evaluates audiovisual relations at global, cross-shot, and intra-shot levels, including sound attribution and speaker timbre consistency.

4.2.2. Automated and Agent-Based Evaluation Pipelines

Complex outputs are difficult to assess with one evaluator or one fixed score. Automated pipelines therefore decompose the task, select applicable criteria, combine different evaluators, and preserve enough detail to support diagnosis and revision.

Hierarchical evaluation.

For images, evaluation often proceeds from task requirements to atomic checks, preventing strong overall quality from masking an omitted object, attribute, or relation. IA-Bench [87] uses fine-grained binary checklists and reports both average checklist accuracy and a strict pass rate that requires every item to be satisfied. GenEval 2 [284] aggregates per-atom VQA results through Soft-TIFA, while Qwen-Image-Bench [285] uses Q-Judger to score active fine-grained facets before aggregating them through its capability hierarchy. The resulting scores retain the connection between an aggregate result and the requirements that produced it.

Video pipelines must also follow temporal structure. MSVBench [295] organizes its scripts through global priors, scene segments, and shot conditions, then combines specialist models with LMM reasoning to assess story alignment and cross-shot consistency. MSAVBench [294] explicitly separates global, cross-shot, intra-shot, and reference-level metrics. It first obtains shot boundaries with TransNetV2, then lets a VLM inspect the segments and invoke merge or split operations when the initial segmentation is unsuitable. This correction reduces the risk that one boundary error will distort all subsequent shot-level scores.

Dynamic dimension selection.

A fixed metric set can penalize an output for properties that its prompt never requested. Uni-VBench [293] addresses this problem by decomposing the input and producing a task-specific checklist of relevant dimensions. Qwen-Image-Bench [285] excludes inapplicable facets from aggregation, while FAGER [308] constructs and verifies a factual rubric for each prompt-reference pair from proposed and visually observable facts. AtelierJudge [309] routes evaluation skills into separate subjective and objective branches instead of applying one scoring procedure to every criterion.

The same principle applies to long-form video, where relevant criteria depend on the events and modalities present in the content. DirectorBench [299] builds a content profile and activates only applicable checkpoints. Its user profiles then modify the weighting of narrative, visual, audio, and cross-modal criteria and specify hard constraints. Criterion selection determines what should be judged, while profile weighting determines how applicable judgments contribute to the final assessment.

 doi:10.20944/preprints202609.0682.v1

Scoring methodology.

Scoring must account for the source of an error and the type of evidence available. AtelierEval [309] evaluates prompts and images separately across T2I backends to distinguish prompting proficiency from rendering performance. WeBench-MMGenEdit [86] uses isolated judges for the pre-generation agentic chain, the multimodal input delivered to the image model, and the final image. This separation locates errors in the agentic process, input construction, or visual realization while keeping the evidence available to each judge distinct. In the related policy-training setting, Gen-Searcher [301] combines image-based and text-based rewards because the final image also depends on the capability and stochasticity of a fixed downstream generator. The text reward assesses the gathered information more directly, while the image reward retains the generation outcome.

Video evaluators often assign semantic judgments to VLMs and measurable perceptual properties to specialist models. EvalVerse [302] extracts evidence with dedicated operators before applying expert-guided reasoning in a fine-tuned VLM. Its two-stage training first learns pairwise preferences and then calibrates pointwise scores and rationales. MuSS [300] similarly combines LMM visual reasoning with TransNetV2, RAFT, and DINOv2 measurements for transition timing, motion, and identity. EntityBench [296] grounds scheduled entities with GroundingDINO and CLIP, then combines DINOv2 similarity with LLM pairwise judgments. Its cross-shot evaluation uses per-shot fidelity as an admission condition, so repeated but poorly rendered entities do not receive high consistency scores. MSAVBench [294] assigns well-defined measurements to specialist models, uses instance-specific rubrics for subjective criteria, and allows its evaluation agent to invoke perception tools for complex spatial judgments.

Automatic scores also require direct validation against human judgments. GenEval 2 [284] compares Soft-TIFA's prompt-level estimates with human prompt-level annotations using AUROC. MSAVBench [294] derives expert system rankings from anonymized pairwise comparisons and measures their agreement with automatic rankings using Spearman's ρ. Scores used to compare systems or guide revision should be checked against human judgments in this way.

Actionable evaluation feedback.

Qwen-Image-Agent [87] converts failed checklist items into feedback context and uses it to refine the next generation prompt. FAGER [308] returns question-level judgments and natural-language feedback for regeneration or editing. Evaluation records can also supervise evaluators. MSVBench [295] converts evaluation traces into instruction data and trains a Qwen3-VL-4B evaluator with GRPO. UniVBench [293] produces structured weakness checklists for targeted optimization, while Director-Bench [299] reports low-scoring checkpoints and prioritized bottlenecks. Each output connects an evaluation result to a specific correction or training target beyond the aggregate score.

5. Challenges and Future Directions

Agentic visual creation remains at an early stage. Existing systems show that agents can coordinate multi-stage production by translating creative goals into plans, acting through visual models and software, and revising the work in response to intermediate results. Yet most are still organized as task-specific workflows. Their planning logic, tool graphs, role assignments, state-update rules, and repair procedures are largely hand-designed around powerful foundation models. While effective for building task-specific systems, such designs often generalize poorly to new creation tasks, tools, and production environments. Moving beyond these prototypes raises fundamental challenges in learning reusable creation policies, maintaining reliable long-horizon state, evaluating both artifacts and agent behaviors, and supporting efficient human–agent interaction.

 doi:10.20944/preprints202609.0682.v1

5.1. Challenges

Data for Training Agentic Creation Policies.

Many existing methods rely on LLM prompting or manually engineered workflows to make production decisions. While flexible, this paradigm does not provide a clear recipe for learning generalizable policies for agentic visual creation. Rather than only decomposing a request into a predefined plan, a capable creation policy must decide, conditioned on the current production state, how to decompose the task, which assets or references to retrieve, which models and tools to invoke, how to construct their inputs, whether an intermediate result should be accepted or revised, and when to replan or terminate the process.

Training such policies requires richer supervision than conventional paired image-text or video-text datasets. A useful trajectory must connect the initial user intent and evolving production state to the agent's decisions about tools and parameters. It should then record the generated artifacts, evaluation or human feedback, revision actions, and final outcome. Efficiency signals, such as generation cost and the number of repair attempts, provide additional supervision. Failures and alternative decisions are as important as successful trajectories because recovery often depends on recognizing an incorrect plan, unsuitable tool choice, or unsuccessful generation.

Constructing these data is challenging for several reasons. First, collecting these data is expensive because high-quality production traces often require expertise from designers, editors, animators, or filmmakers. Second, the action space is also difficult to standardize. Visual creation may involve natural-language plans and storyboards, layout trees and asset libraries, ComfyUI graphs and Blender programs, GUI actions, or memory states. Third, long trajectories introduce a separate credit-assignment problem because final quality may depend on decisions made many steps earlier. A locally reasonable action can still create downstream inconsistencies. Finally, synthetic trajectories generated by existing agents may reproduce their own planning biases, evaluation errors, or tool-use failures. Future data pipelines therefore need to combine instrumented production logs, expert demonstrations, automatically verified executions, human preferences, controlled failure trajectories, and outcome-aware filtering to provide reliable supervision for agentic decision making.

Reliable State and Memory for Long-Horizon Creation.

Long-horizon visual production requires agents to preserve and update information that remains relevant across multiple generation and editing steps. This state may include character identity and appearance, scene layout, object ownership and status, visual style, camera configuration, temporal order, accepted assets, user preferences, and dependencies among production units. Existing systems increasingly use explicit memories, asset repositories, scene representations, or production graphs. Maintaining these representations becomes more difficult as the number of components and interactions grows.

The challenge is not merely to store more context, but to determine what should be retained, retrieved, compressed, updated, or discarded. State can become stale when an earlier asset is revised, contradictory when multiple observations provide incompatible information, or overly restrictive when historical decisions are propagated to situations in which they no longer apply. Errors in state management can consequently cause identity drift, inconsistent object states, invalid dependency propagation, or unnecessary regeneration. More reliable systems need structured and versioned production state with provenance-aware updates. Selective multimodal retrieval, conflict detection, and dependency-aware propagation are also needed to keep that state usable. An important open question is how to combine explicit external state with the increasingly long multimodal context supported by foundation models, rather than treating either representation as universally sufficient.

Evaluation Beyond Final-Artifact Quality.

Current evaluation protocols still focus heavily on the quality of individual generated outputs, such as image quality, alignment between prompt and image, video fidelity, or short-clip temporal

 doi:10.20944/preprints202609.0682.v1

consistency. These metrics remain necessary, but they do not fully characterize an agentic creation system. Two agents may produce artifacts of similar quality while differing substantially in plan validity, tool selection, state maintenance, recovery from failures, generation cost, or responsiveness to user intervention. Evaluation should therefore cover both *what* the system produces and *how* it arrives at the result.

Recent benchmarks broaden artifact-level evaluation by changing the unit and temporal scale of diagnosis. DirectorBench conditions long-form video assessment on structured metadata and user profiles, activates only applicable checkpoints, and returns checkpoint-level diagnoses [299]. MSVBench decomposes multi-shot visual consistency into face, character, background, clothing and color, and relative-size criteria. MSAVBench extends assessment to global, cross-shot, intra-shot, and reference levels for multi-shot audio-video generation. LongAV-Compass measures event fulfillment, long-form continuity, transition stability, and audio-video synchronization over minute-scale outputs [294,295,298]. These benchmarks provide increasingly fine-grained artifact diagnosis, but they capture only part of the behavior of an agentic production system.

A more complete evaluation should additionally examine the execution trajectory. Agent-level criteria should examine whether a plan is executable, whether the selected tools suit the current state, and whether memory updates are correct. It should also test whether the agent localizes a failure to the responsible component and repairs it without unnecessarily modifying accepted content. Practical efficiency depends on repair cost and latency, the number of generation attempts, user intervention, and termination behavior. These factors distinguish useful production agents from systems that depend on expensive trial and error.

Evaluation itself also becomes part of the agentic loop, creating a second reliability problem. When feedback from a VLM or multimodal judge determines whether the agent regenerates, rolls back, or terminates, evaluator errors can propagate into subsequent production decisions. This issue becomes particularly important when similar foundation models are used both to create and to judge the same content. Mature evaluation protocols should therefore measure not only final-artifact quality and trajectory efficiency, but also evaluator calibration, agreement with human judgments, failure localization, and the reliability of feedback as a basis for downstream actions. Ultimately, benchmarks for agentic visual creation should combine artifact-level, trajectory-level, and interaction-level diagnostics.

5.2. Future Directions

Learning Generalizable and Self-Improving Creation Policies.

Learning production policies from multimodal state offers a path beyond manually designed workflows. Rather than following a fixed sequence of model calls, such a policy would decide when to plan or replan, retrieve external information, select a generator or editor, inspect an intermediate result, perform a localized repair, or terminate generation. This turns visual creation from a sequence of manually connected model calls into a sequential decision problem whose behavior can adapt to task requirements and execution outcomes.

Supervised trajectory learning can provide an initial policy, while preference optimization or reinforcement learning can further optimize decisions using artifact quality, instruction following, consistency, execution success, latency, and generation cost as complementary signals. Hierarchical policies may separate production-level reasoning from local execution. A high-level policy determines production goals and dependencies, while lower-level policies select tools and execute local operations. Beyond parameter optimization, agents may also accumulate successful workflows, reusable skills, and failure-recovery strategies in external memory, allowing capabilities to improve as new tasks are solved. A central research question is how to achieve such continual improvement without amplifying erroneous experience or overfitting to a particular set of tools and generators.

 doi:10.20944/preprints202609.0682.v1

Tighter Generation–Understanding Coupling for Closed-loop Creation.

Current agentic pipelines often connect separate components: an LLM plans, a generator synthesizes images or videos, a VLM evaluates the result, and an external controller decides whether to revise. This modular design is flexible and allows specialized tools to be replaced independently. However, each handoff requires information to be translated among textual plans, visual outputs, evaluator feedback, and tool-specific representations. These translations can discard useful context and increase interaction cost. A promising direction is therefore to develop tighter coupling between generation and multimodal understanding within the agentic loop.

Unified multimodal models that support both understanding and generation [9,53,57] provide one possible realization of this idea. A shared model may reason about the user's goal, generate or manipulate intermediate visual states, inspect its own outputs, and update subsequent decisions without repeatedly converting information across separately trained representations. For long-form creation, shared multimodal representations may also provide a compact interface between current visual observations and persistent memories of characters, scenes, and events.

However, closed-loop creation does not necessarily require all capabilities to collapse into a single end-to-end model. External tools, specialized generators, executable representations, and explicit memories remain valuable because they provide controllability, editability, and verifiable execution. The more general opportunity is therefore to reduce the boundary between generation and understanding while preserving access to specialized tools and structured production state. This requires models that can process long visual contexts, ground feedback to editable components, expose controllable intermediate decisions, and support localized regeneration rather than only producing a new artifact from scratch.

Real-Time Interactive Visual Creation.

Real-time interaction would allow users and agents to develop an artifact continuously, rather than alternate between isolated prompts and complete regenerations. Professional creation is inherently incremental: creators inspect partial results, adjust local details, compare alternatives, revise earlier decisions, and change their goals as the artifact evolves. Current generation latency and coarse interaction units make such mixed-initiative workflows difficult.

Supporting such interaction requires updating plans, memories, and visual outputs incrementally as the user types, sketches, drags objects, changes references, or edits a timeline. Instead of regenerating an entire image, video, presentation, or scene after each instruction, the system should identify the affected region, shot, asset, or workflow node and update only the relevant production state. Intermediate previews should make these changes observable before expensive downstream operations are executed.

In this setting, the agent becomes a mixed-initiative creative collaborator rather than an autonomous task executor. It must detect ambiguity without repeatedly interrupting the workflow and present alternatives when several creative choices remain valid. It must also preserve editable history, support rollback and branching, and propagate user changes to dependent components without altering unrelated decisions. This mode of interaction depends on low-latency generation and editing, streaming multimodal understanding, incremental execution, and efficient state synchronization. The interface must expose useful decisions and uncertainty without overwhelming the creator.

6. Conclusions

This survey reviews existing works related to agentic visual creation, organizing the literature by the agent's main responsibility: visual generation, visual editing, visual composition, and visual programming. Within generation, target-centered methods focus on the current artifact and its production history, whereas coordinated methods connect identifiable production components through shared specifications, assets, or state. Editing emphasizes source preservation and temporal constraints, while composition and programming work through editable layouts and executable representations. Across

 doi:10.20944/preprints202609.0682.v1

these categories, planning, tool use, memory, and feedback support the production process around generative models. Benchmarks increasingly use structured multimodal specifications, hierarchical output assessment, and feedback for revision, but evaluation of complete agent trajectories remains limited. Current systems still rely on hand-designed workflows and face shortages of training trajectories, difficulties in maintaining state over long tasks, evaluator errors, and limited support for fine-grained interaction. Further progress requires training data that captures both successful and failed attempts, evaluation of agent behavior alongside artifact quality, and models and interfaces that support efficient, controllable, and incremental revision.

References

1. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **2020**, *33*, 6840–6851.
2. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
3. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 3836–3847.
4. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4195–4205.
5. Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. Video generation models as world simulators, 2024.
6. Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation, 2023, [arXiv:cs.CV/2310.19512].
7. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International conference on machine learning. Pmlr, 2021, pp. 8821–8831.
8. Yu, J.; Xu, Y.; Koh, J.Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B.K.; et al. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research* **2022**. Featured Certification.
9. Deng, C.; Zhu, D.; Li, K.; Gou, C.; Li, F.; Wang, Z.; Zhong, S.; Yu, W.; Nie, X.; Song, Z.; et al. Emerging Properties in Unified Multimodal Pretraining, 2025, [arXiv:cs.CV/2505.14683].
10. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the The Eleventh International Conference on Learning Representations, 2023.
11. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems* **2023**, *36*, 68539–68551.
12. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems* **2023**, *36*, 8634–8652.
13. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems* **2023**, *36*, 46534–46594.
14. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems* **2023**, *36*, 51991–52008.
15. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In Proceedings of the First conference on language modeling, 2024.
16. Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Yau, S.; Lin, Z.; Zhou, L.; et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In Proceedings of the International Conference on Learning Representations, 2024, pp. 23247–23275.

 doi:10.20944/preprints202609.0682.v1

17. Feng, W.; Zhu, W.; Fu, T.j.; Jampani, V.; Akula, A.; He, X.; Basu, S.; Wang, X.E.; Wang, W.Y. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems* **2023**, *36*, 18225–18250.
18. Lian, L.; Li, B.; Yala, A.; Darrell, T. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *arXiv preprint arXiv:2305.13655* **2023**.
19. Lin, H.; Zala, A.; Cho, J.; Bansal, M. VideoDirectorGPT: Consistent Multi-scene Video Generation via LLM-Guided Planning, 2024, [arXiv:cs.CV/2309.15091].
20. Zhao, L.; Yang, Y.; Zhang, K.; Shao, W.; Zhang, Y.; Qiao, Y.; Luo, P.; Ji, R. Diffagent: Fast and accurate text-to-image api selection with large language model. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6390–6399.
21. Wang, Z.; Xie, E.; Li, A.; Wang, Z.; Liu, X.; Li, Z. Divide and Conquer: Language Models can Plan and Self-Correct for Compositional Text-to-Image Generation, 2024, [arXiv:cs.CV/2401.15688].
22. Wan, X.; Zhou, H.; Sun, R.; Nakhost, H.; Jiang, K.; Sinha, R.; Arık, S.Ö. Maestro: Self-improving text-to-image generation via agent orchestration, 2025.
23. Gong, Y.; Pang, Y.; Cun, X.; Xia, M.; He, Y.; Chen, H.; Wang, L.; Zhang, Y.; Wang, X.; Shan, Y.; et al. TaleCrafter: Interactive Story Visualization with Multiple Characters, 2023, [arXiv:cs.CV/2305.18247].
24. Zhou, J.; Du, Y.; Xu, X.; Wang, L.; Zhuang, Z.; Zhang, Y.; Li, S.; Hu, X.; Su, B.; cong Chen, Y. VideoMemory: Toward Consistent Video Generation via Memory Integration, 2026, [arXiv:cs.CV/2601.03655].
25. Xu, Z.; Wang, L.; Wang, J.; Li, Z.; Shi, S.; Yang, X.; Wang, Y.; Hu, B.; Yu, J.; Zhang, M. FilmAgent: A Multi-Agent Framework for End-to-End Film Automation in Virtual 3D Spaces, 2025, [arXiv:cs.CL/2501.12909].
26. Wu, W.; Zhu, Z.; Shou, M.Z. Automated Movie Generation via Multi-Agent CoT Planning, 2025, [arXiv:cs.CV/2503.07314].
27. Xie, Z.; Tang, D.; Tan, D.; Klein, J.; Bissyand, T.F.; Ezzini, S. DreamFactory: Pioneering Multi-Scene Long Video Generation with a Multi-Agent Framework, 2024, [arXiv:cs.AI/2408.11788].
28. Cui, X.; Li, Z.; Li, P.; Hu, Y.; Shi, H.; Cao, C.; He, Z. ChatEdit: Towards Multi-turn Interactive Facial Image Editing via Dialogue. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023; Bouamor, H.; Pino, J.; Bali, K., Eds. Association for Computational Linguistics, 2023, pp. 14567–14583. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.899.
29. Lin, Y.; Lin, Z.; Lin, K.; Bai, J.; Pan, P.; Li, C.; Chen, H.; Wang, Z.; Ding, X.; Li, W.; et al. Jarvisart: Liberating human artistic creativity via an intelligent photo retouching agent. *Advances in Neural Information Processing Systems* **2025**, *38*, 52088–52130.
30. Zuo, Y.; Zheng, Q.; Wu, M.; Jiang, X.; Li, R.; Wang, J.; Zhang, Y.; Mai, G.; Wang, L.V.; Zou, J.Y.; et al. 4KAgent: Agentic Any Image to 4K Super-Resolution. In Proceedings of the Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025; Belgrave, D.; Zhang, C.; Montoya, L.N.; Lin, H.; Pascanu, R.; Koniusz, P.; Ghassemi, M.; Chen, N.; Ruíz, I.V.M.; Loaiza-Bonilla, A., Eds., 2025.
31. Wang, B.; Li, Y.; Lv, Z.; Xia, H.; Xu, Y.; Sodhi, R. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. In Proceedings of the Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI 2024, Greenville, SC, USA, March 18-21, 2024. ACM, 2024, pp. 699–714. https://doi.org/10.1145/3640543.3645143.
32. Sandoval-Castañeda, M.; Russell, B.C.; Sivic, J.; Shakhnarovich, G.; Heilbron, F.C. EditDuet: A Multi-Agent System for Video Non-Linear Editing. In Proceedings of the Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH Conference Papers 2025, Vancouver, BC, Canada, August 10-14, 2025; Alford, G.; Zhang, H.R.; Schulz, A., Eds. ACM, 2025, pp. 2:1–2:11. https://doi.org/10.1145/3721238.3730761.
33. Kang, J.; Kwon, S.; Lee, J.; Kim, B.H. DIAMOND: An LLM-Driven Agent for Context-Aware Baseball Highlight Summarization. In Proceedings of the Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025); Kamalloo, E.; Gontier, N.; Lu, X.H.; Dziri, N.; Murty, S.; Lacoste, A., Eds., Vienna, Austria, 2025; pp. 386–400. https://doi.org/10.18653/v1/2025.realm-1.28.
34. Pang, W.; Lin, K.Q.; Jian, X.; He, X.; Torr, P. Paper2poster: Towards multimodal poster automation from scientific papers. *Advances in Neural Information Processing Systems* **2025**, *38*.

 doi:10.20944/preprints202609.0682.v1

35. Zheng, H.; Guan, X.; Kong, H.; Zhang, W.; Zheng, J.; Zhou, W.; Lin, H.; Lu, Y.; Han, X.; Sun, L. Pptagent: Generating and evaluating presentations beyond text-to-slides. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 14402–14418.
36. Xue, X.; Lu, Z.; Huang, D.; Wang, Z.; Ouyang, W.; Bai, L. Comfybench: Benchmarking llm-based agents in comfyui for autonomously designing collaborative ai systems. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2025, pp. 24614–24624.
37. Hu, Z.; Iscen, A.; Jain, A.; Kipf, T.; Yue, Y.; Ross, D.A.; Schmid, C.; Fathi, A. Scenecraft: An llm agent for synthesizing 3d scenes as blender code. In Proceedings of the Forty-first International Conference on Machine Learning, 2024.
38. Tseng, T.; Cheng, R.; Nichols, J. Keyframer: Empowering Animation Design using Large Language Models. In Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), 2025.
39. Gao, D.; Ji, L.; Bai, Z.; Ouyang, M.; Li, P.; Mao, D.; Wu, Q.; Zhang, W.; Wang, P.; Guo, X.; et al. Assistgui: Task-oriented desktop graphical user interface automation. *arXiv preprint arXiv:2312.13108* **2023**.
40. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In Proceedings of the International Conference on Learning Representations, 2021.
41. Ma, Y.; Feng, K.; Hu, Z.; Wang, X.; Wang, Y.; Zheng, M.; He, X.; Zhu, C.; Liu, H.; He, Y.; et al. Controllable Video Generation: A Survey. *arXiv preprint arXiv:2507.16869* **2025**.
42. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.
43. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **2020**, *21*, 1–67.
44. Ho, J.; Salimans, T. Classifier-Free Diffusion Guidance. In Proceedings of the NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.
45. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the Forty-first international conference on machine learning, 2024.
46. Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; Zhang, S. ModelScope Text-to-Video Technical Report, 2023, [arXiv:cs.CV/2308.06571].
47. Hu, J.; Liu, J.; Yang, L.; Zhang, X.; Li, K.; Zeng, S.; Li, Y.; Huang, H.; Zhang, C.; Lu, Y. Geometry-as-context: Modulating explicit 3d in scene-consistent video generation to geometry context. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 4258–4268.
48. Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; Chen, Q. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 4117–4125.
49. Ma, Y.; He, Y.; Wang, H.; Wang, A.; Shen, L.; Qi, C.; Ying, J.; Cai, C.; Li, Z.; Shum, H.Y.; et al. Follow-Your-Click: Open-domain Regional Image Animation via Motion Prompts. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 6018–6026.
50. Van Den Oord, A.; Vinyals, O.; et al. Neural discrete representation learning. *Advances in neural information processing systems* **2017**, *30*.
51. Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869* **2024**.
52. Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; Wang, L. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems* **2024**, *37*, 84839–84865.
53. Zhao, S.; Zhang, X.; Guo, J.; Hu, J.; Duan, L.; Fu, M.; Chng, Y.X.; Wang, G.H.; Chen, Q.G.; Xu, Z.; et al. Unified Multimodal Understanding and Generation Models: Advances, Challenges, and Opportunities, 2026, [arXiv:cs.CV/2505.02567].
54. Cui, Y.; Chen, H.; Deng, H.; Huang, X.; Li, X.; Liu, J.; Liu, Y.; Luo, Z.; Wang, J.; Wang, W.; et al. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583* **2025**.
55. Wang, G.H.; Zhao, S.; Zhang, X.; Cao, L.; Zhan, P.; Duan, L.; Lu, S.; Fu, M.; Chen, X.; Zhao, J.; et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044* **2025**.

56. Chen, J.; Xu, Z.; Pan, X.; Hu, Y.; Qin, C.; Goldstein, T.; Huang, L.; Zhou, T.; Xie, S.; Savarese, S.; et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568* **2025**.
57. Fu, F.; Huang, M.; Wu, S.; Jiang, Y.; Huo, Y.; Li, H.; Song, Y.; Ding, F.; Guo, J.; He, Q.; et al. Lance: Unified multimodal modeling by multi-task synergy. *arXiv preprint arXiv:2605.18678* **2026**.
58. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **2022**, *35*, 24824–24837.
59. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* **2023**, *36*, 11809–11822.
60. Wang, L.; Xu, W.; Lan, Y.; Hu, Z.; Lan, Y.; Lee, R.K.W.; Lim, E.P. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In Proceedings of the Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), 2023, pp. 2609–2634.
61. Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. WebGPT: Browser-assisted question-answering with human feedback, 2022, [arXiv:cs.CL/2112.09332].
62. Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.O.; Wang, D.; Zamani, H.; Han, J. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. In Proceedings of the Second Conference on Language Modeling, 2025.
63. Feng, J.; Huang, S.; Qu, X.; Zhang, G.; Qin, Y.; Zhong, B.; Jiang, C.; Chi, J.; Zhong, W. ReTool: Reinforcement Learning for Strategic Tool Use in LLMs, 2025, [arXiv:cs.CL/2504.11536].
64. Zhang, Z.; Dai, Q.; Bo, X.; Ma, C.; Li, R.; Chen, X.; Zhu, J.; Dong, Z.; Wen, J.R. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems* **2025**, *43*, 1–47.
65. Park, J.S.; O'Brien, J.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. In Proceedings of the Proceedings of the 36th annual acm symposium on user interface software and technology, 2023, pp. 1–22.
66. Hu, M.; Chen, T.; Chen, Q.; Mu, Y.; Shao, W.; Luo, P. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 32779–32798.
67. He, Z.; Huang, S.; Qu, X.; Li, Y.; Zhu, T.; Cheng, Y.; Yang, Y. GEMS: Agent-Native Multimodal Generation with Memory and Skills, 2026, [arXiv:cs.CV/2603.28088].
68. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *National Science Review* **2024**, *11*, nwae403.
69. Liu, R.; Liu, Z.; Tang, J.; Ma, Y.; Pi, R.; Zhang, J.; Chen, Q. Longvideoagent: Multi-agent reasoning with long videos. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2026, pp. 40404–40416.
70. Qu, L.; Wu, S.; Fei, H.; Nie, L.; Chua, T.S. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In Proceedings of the Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 643–654.
71. Jia, Y.; Tan, W., DivCon: Divide and Conquer for Complex Numerical and Spatial Reasoning in Text-to-Image Generation. In *ECAI 2025*; IOS Press, 2025. https://doi.org/10.3233/faia251298.
72. Yang, L.; Yu, Z.; Meng, C.; Xu, M.; Ermon, S.; Cui, B. Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. In Proceedings of the International Conference on Machine Learning, 2024.
73. Feng, Y.; Gong, B.; Chen, D.; Shen, Y.; Liu, Y.; Zhou, J. Ranni: Taming text-to-image diffusion for accurate instruction following. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4744–4753.
74. Li, S.; Wang, R.; Hsieh, C.J.; Cheng, M.; Zhou, T. Mulan: Multimodal-llm agent for progressive and interactive multi-object diffusion, 2024.
75. Ren, T.; Yan, Z.; Zhao, Y.; Fang, Z.; Zeng, Y.; Zhang, G.; Xu, H.; Ma, X.; Huang, S.; Xu, K.; et al. SCOPE: Structured Decomposition and Conditional Skill Orchestration for Complex Image Generation, 2026, [arXiv:cs.CV/2605.08043].
76. Li, M.; Hou, X.; Liu, Z.; Yang, D.; Qian, Z.; Chen, J.; Wei, J.; Jiang, Y.; Xu, Q.; Zhang, L. MCCD: Multi-Agent Collaboration-based Compositional Diffusion for Complex Text-to-Image Generation. In Proceedings of the

 doi:10.20944/preprints202609.0682.v1

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 13263–13272.

77. Saha, O.; Krs, V.; Mech, R.; Maji, S.; Gadelha, M.; Blackburn-Matzen, K. 3D Space as a Scratchpad for Editable Text-to-Image Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026, pp. 29233–29243.
78. Long, F.; Qiu, Z.; Yao, T.; Mei, T. Videostudio: Generating consistent-content and multi-scene videos. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 468–485.
79. Zhuang, S.; Li, K.; Chen, X.; Wang, Y.; Liu, Z.; Qiao, Y.; Wang, Y. Vlogger: Make your dream a vlog. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8806–8817.
80. Ye, J.; He, J.; Huang, Z.; Jiang, D.; Yang, X.; Chen, R.; Li, W. GenClaw: Code-Driven Agentic Image Generation, 2026, [arXiv:cs.CV/2605.30248].
81. Hahn, M.; Zeng, W.; Kannen, N.; Galt, R.; Badola, K.; Kim, B.; Wang, Z. Proactive Agents for Multi-Turn Text-to-Image Generation Under Uncertainty. In Proceedings of the International Conference on Machine Learning. PMLR, 2025, pp. 21591–21628.
82. Li, C.; Wu, Q.; Pan, J.H.; Hui, K.H.; Hu, J.; Jiang, Y.; Sheng, B.; Liu, X.; Gong, W.; Liu, Z. coDrawAgents: A Multi-Agent Dialogue Framework for Compositional Image Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings, June 2026, pp. 9802–9812.
83. Shalev-Arkushin, R.; Gal, R.; Bermano, A.; Fried, O. Imagerag: Dynamic image retrieval for reference-guided image generation. In Proceedings of the International Conference on Learning Representations, 2026, pp. 11400–11427.
84. Yuan, H.; Zhao, Z.; Wang, S.; Xiao, S.; Ni, M.; Liu, Z.; Dou, Z. Finerag: Fine-grained retrieval-augmented text-to-image generation. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 11196–11205.
85. Chen, S.; Shou, Q.; Chen, H.; Zhou, Y.; Feng, K.; Hu, W.; Zhang, Y.F.; Lin, Y.; Huang, W.; Song, M.; et al. Unify-Agent: A Unified Multimodal Agent for World-Grounded Image Synthesis, 2026, [arXiv:cs.CV/2603.29620].
86. Zhang, H.; Liu, Z.; Niu, L.; Liu, J.; Li, H.; Cao, Z.; Chen, W.; Zhao, C.; Meng, F. WeAgent-MMGenEdit: A Full-Stack Recipe for Multimodal Agentic Image Generation and Editing, 2026, [arXiv:cs.CV/2609.05171].
87. Zhang, Z.; Li, J.; Zhang, J.; Gao, K.; Yan, K.; Jiang, L.; Tang, N.; Yin, S.; Wu, T.; Chen, X.; et al. Qwen-Image-Agent: Bridging the Context Gap in Real-World Image Generation, 2026, [arXiv:cs.CV/2606.26907].
88. Bian, F.; Zheng, Z.; Deng, W.; Zhou, D.; Luan, J. RS-Gen: A Multi-Stage Agentic Framework for Reasoning and Search-Augmented Image Generation, 2026, [arXiv:cs.CV/2606.23221].
89. Wang, H.; Feng, W.; Yu, J.; Liu, C.; Nie, P.; Lin, F.; Liu, J.; Huang, R.; Lin, J.; Chen, W.; et al. Search Beyond What Can Be Taught: Evolving the Knowledge Boundary in Agentic Visual Generation, 2026, [arXiv:cs.CV/2607.05382].
90. He, J.; Ye, J.; Huang, Z.; Jiang, D.; Zhang, C.; Zhu, L.; Zhang, R.; Zhang, X.; Li, W. Mind-Brush: Integrating Agentic Cognitive Search and Reasoning into Image Generation, 2026, [arXiv:cs.CV/2602.01756]. https://doi.org/10.48550/arXiv.2602.01756.
91. Wang, F.; Fu, C.; Huang, Z.; Li, C.; Lyu, J.; Li, G. Cognitive-structured Multimodal Agent for Multimodal Understanding, Generation, and Editing, 2026, [arXiv:cs.CV/2607.08497]. https://doi.org/10.48550/arXiv.2607.08497.
92. Song, K.; Hou, T.; He, Z.; Ma, H.; Wang, J.; Sinha, A.; Tsai, S.; Luo, Y.; Dai, X.; Chen, L.; et al. Llama Learns to Direct: DirectorLLM for Human-Centric Video Generation, 2024.
93. Huang, H.; Feng, Y.; Shi, C.; Xu, L.; Yu, J.; Yang, S. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems* **2023**, *36*, 26135–26158.
94. Hong, S.; Seo, J.; Shin, H.; Hong, S.; Kim, S. DirecT2V: Large Language Models are Frame-Level Directors for Zero-Shot Text-to-Video Generation, 2024, [arXiv:cs.CV/2305.14330].
95. Lian, L.; Shi, B.; Yala, A.; Li, B.; et al. Llm-grounded video diffusion models. In Proceedings of the International Conference on Learning Representations, 2024, Vol. 2024, pp. 50207–50227.
96. Zhu, H.; He, T.; Tang, A.; Guo, J.; Chen, Z.; Bian, J. Compositional 3d-aware video generation with llm director. *Advances in neural information processing systems* **2024**, *37*, 131618–131644.
97. Tian, Y.; Yang, L.; Yang, H.; Gao, Y.; Deng, Y.; Chen, J.; Wang, X.; Yu, Z.; Tao, X.; Wan, P.; et al. Videotetris: Towards compositional text-to-video generation. *Advances in Neural Information Processing Systems* **2024**, *37*, 29489–29513.

 doi:10.20944/preprints202609.0682.v1

98. Oh, G.; Jeong, J.; Kim, S.; Byeon, W.; Kim, J.; Kim, S.; Kim, S. Mevg: Multi-event video generation with text-to-video models. In Proceedings of the European Conference on computer vision. Springer, 2024, pp. 401–418.
99. Liao, X.; Zeng, X.; Wang, L.; Yu, G.; Lin, G.; Zhang, C. MotionAgent: Fine-grained Controllable Video Generation via Motion Field Agent. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 11305–11316.
100. Feng, H.; Ma, Y.; Di, D.; Fan, L.; Su, T. CogPortrait: Fine-Grained Eye-Region Control in Portrait Animation via Hierarchical Agent Planning, 2026, [arXiv:cs.CV/2605.28056].
101. Lu, Y.; Zhu, L.; Fan, H.; Yang, Y. FlowZero: Zero-Shot Text-to-Video Synthesis with LLM-Driven Dynamic Scene Syntax, 2023, [arXiv:cs.CV/2311.15813].
102. Cai, M.; Zhang, M.; Yang, C.; Li, Y.; Yoshie, O.; Ieiri, Y. KGEdit: Ambiguity-Aware Knowledge Graphs for Training-Free Precise Video Generation and Editing, 2026, [arXiv:cs.CV/2605.29509].
103. Zhou, D.; Huang, X.; Wang, X.; Xie, J.; Zhang, Y.; Li, L.; Li, K.; Yang, Z.; Yang, Y. MetaPoint: Unlocking Precise Spatial Control in Agentic Visual Generation, 2026, [arXiv:cs.CV/2606.05031].
104. Khamis, A. Agentic Ontology-guided Image Generation and Evaluation for Rare-event Data Augmentation in Safety-critical Perception. *Array* **2026**, *30*, 100932. https://doi.org/10.1016/j.array.2026.100932.
105. Song, Q.; Zhou, D.; Lin, J.; Shen, F.; Wang, J.; Hu, X.; Chen, C.; Heng, P. SceneDecorator: Towards Scene-Oriented Story Generation with Scene Planning and Scene Consistency. In Proceedings of the Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025; Belgrave, D.; Zhang, C.; Montoya, L.N.; Lin, H.; Pascanu, R.; Koniusz, P.; Ghassemi, M.; Chen, N.; Ruíz, I.V.M.; Loaiza-Bonilla, A., Eds., 2025.
106. Yao, Z.; Li, Y.; Gao, X.; Chen, Q.; Jiang, P.; Lu, Y. Narrative Weaver: Towards Controllable Long-Range Visual Consistency with Multi-Modal Conditioning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 7707–7718.
107. Wang, Z.; Li, J.; Lin, H.; Yoon, J.; Bansal, M. Dreamrunner: Fine-grained compositional story-to-video generation with retrieval-augmented motion adaptation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 10503–10511.
108. He, Y.; Xia, M.; Chen, H.; Cun, X.; Gong, Y.; Xing, J.; Zhang, Y.; Wang, X.; Weng, C.; Shan, Y.; et al. Animate-A-Story: Storytelling with Retrieval-Augmented Video Generation, 2023, [arXiv:cs.CV/2307.06940].
109. Zheng, J.; Cun, X. Fairygen: Storied cartoon video from a single child-drawn character. In Proceedings of the Proceedings of the SIGGRAPH Asia 2025 Conference Papers, 2025, pp. 1–11.
110. Zheng, M.; Xu, Y.; Huang, H.; Ma, X.; Liu, Y.; Shu, W.; Pang, Y.; Tang, F.; Chen, Q.; Yang, H.; et al. VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention, 2025, [arXiv:cs.CV/2412.02259].
111. Zhang, P.; Jia, Z.; Liu, K.; Weng, S.; Li, S.; Shi, B. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 659–669.
112. Chen, C.; Dang, S.; Liu, Y.; Zhao, N.; Shi, Y.; Cao, N. MV-Crafter: An Intelligent System for Music-guided Video Generation. *ACM Transactions on Interactive Intelligent Systems* **2025**, *15*, 1–27.
113. Zhou, Y.; Zhou, D.; Cheng, M.M.; Feng, J.; Hou, Q. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems* **2024**, *37*, 110315–110340.
114. Shen, X.; Elhoseiny, M. Storygpt-v: Large language models as consistent story visualizers. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 13273–13283.
115. Yang, S.; Ge, Y.; Li, Y.; Chen, Y.; Ge, Y.; Shan, Y.; Chen, Y.C. Seed-story: Multimodal long story generation with large language model. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 1850–1860.
116. Mao, J.; Huang, X.; Xie, Y.; Chang, Y.; Hui, M.; Xu, B.; Zheng, Z.; Wang, Z.; Xie, C.; Zhou, Y. Story-Iter: A Training-free Iterative Paradigm for Long Story Visualization. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
117. An, Z.; Jia, M.; Qiu, H.; Zhou, Z.; Huang, X.; Liu, Z.; Ren, W.; Kahatapitiya, K.; Liu, D.; He, S.; et al. Onestory: Coherent multi-shot video generation with adaptive memory. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 16173–16184.
118. Atzmon, Y.; Gal, R.; Tewel, Y.; Kasten, Y.; Chechik, G. Motion by Queries: Identity-Motion Trade-offs in Text-to-Video Generation, 2025, [arXiv:cs.CV/2412.07750].

119. Kara, O.; Singh, K.K.; Liu, F.; Ceylan, D.; Rehg, J.M.; Hinz, T. Shotadapter: Text-to-multi-shot video generation with diffusion models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 28405–28415.
120. Zhao, C.; Liu, M.; Wang, W.; Chen, W.; Wang, F.; Chen, H.; Zhang, B.; Shen, C. Moviedreamer: Hierarchical generation for coherent long visual sequence. In Proceedings of the International Conference on Learning Representations, 2025, pp. 50060–50090.
121. Akdemir, K.; Shi, J.; Kafle, K.; Price, B.L.; Yanardag, P. Plot'n polish: Zero-shot story visualization and disentangled editing with text-to-image diffusion models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 1694–1702.
122. Qin, J.; Wu, J.; Chen, W.; Lyu, Y. DiffusionAgent: navigating expert models for agentic image generation. *arXiv e-prints* **2024**, pp. arXiv–2401.
123. Jiang, X.; Chen, B.; Li, G.; Duan, Y.; Wang, R.; Zhang, J. OctoT2I: A Self-Evolving Agentic Text-to-Image Router. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 31628–31638.
124. Ma, S.; Guo, Y.; Su, J.; Huang, Q.; Zhou, Z.; Wang, Y. Talk2Image: A multi-agent system for multi-turn image generation and editing. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 32437–32445.
125. Chen, C.Y.; Shi, M.; Zhang, G.; Shi, H. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 19396–19405.
126. Wang, Z.; Li, A.; Li, Z.; Liu, X. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems* **2024**, *37*, 128374–128395.
127. Yuan, Z.; Liu, Y.; Cao, Y.; Sun, W.; Jia, H.; Chen, R.; Li, Z.; Lin, B.; Yuan, L.; He, L.; et al. Mora: Enabling Generalist Video Generation via A Multi-Agent Framework, 2024, [arXiv:cs.CV/2403.13248].
128. Tu, R.C.; Sun, W.; Jin, Z.; Liao, J.; Huang, J.; Tao, D. Spagent: Adaptive task decomposition and model selection for general video generation and editing. *IEEE Transactions on Image Processing* **2026**.
129. Liang, Z.; Zhang, D.; Zhou, H.; Huang, R.; Li, B.; Zhang, Y.; Wu, S.; Wang, X.; Luo, J.; Liao, L.; et al. UniVA: Universal Video Agent towards Open-Source Next-Generation Video Generalist, 2025, [arXiv:cs.CV/2511.08521].
130. Feng, Y.; Wang, J.; Xu, C.; Qian, Y.; Wang, H.; Hou, W.; Liu, Y.; Sun, B.; Liu, Y.; Wang, S. NEWTON: Agentic Planning for Physically Grounded Video Generation, 2026, [arXiv:cs.CV/2605.18396].
131. Wei, J.; Tan, J.; Zhu, H.; Zhang, X.; Zhang, Y.; Chen, Z.; Zhang, D.; Xu, W.; Liu, Z. VideoWeaver: Evaluating and Evolving Skills for Agentic Long Video Generation, 2026, [arXiv:cs.CV/2606.08091].
132. Chen, H.H.; Hou, Z.; Shu, W.J.; Ruan, W.; Xu, Y.; Guo, L.; Chen, Y.C. GenRouter: Unified Workflow Routing for Agentic Image Generation, 2026, [arXiv:cs.CV/2608.16721].
133. Zhao, J.; Yu, X.; Sun, Z.; Teng, F.; Qin, C.; Hu, X.; Xu, J.; Yan, S. ToolArtist: Tool-Using Unified Multimodal Models for Agentic Image Generation. *arXiv preprint arXiv:2608.04436* **2026**.
134. Mo, W.; Zhang, T.; Bai, Y.; Su, B.; Wen, J.R.; Yang, Q. Dynamic prompt optimizing for text-to-image generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26627–26636.
135. Yang, Z.; Wang, J.; Li, L.; Lin, K.; Lin, C.C.; Liu, Z.; Wang, L. Idea2img: Iterative self-refinement with gpt-4v for automatic image design and generation. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 167–184.
136. Wu, T.H.; Lian, L.; Gonzalez, J.E.; Li, B.; Darrell, T. Self-correcting llm-controlled diffusion models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6327–6336.
137. Xiang, D.; Xu, W.; Chu, K.; Ding, T.; Shen, Z.; Zeng, Y.; Su, J.; Zhang, W. Promptsculptor: Multi-agent based text-to-image prompt optimization. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2025, pp. 774–786.
138. Wu, M.; Wang, L.; Zhao, P.; Yang, F.; Zhang, J.; Liu, J.; Zhan, Y.; Han, W.; Sun, H.; Ji, J.; et al. Re-Prompt: Reasoning-Augmented Reprompting for Text-to-Image Generation via Reinforcement Learning, 2025, [arXiv:cs.CV/2505.17540].
139. Gao, B.; Gao, X.; Wu, X.; Zhou, Y.; Qiao, Y.; Niu, L.; Chen, X.; Wang, Y. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 3173–3183.

 doi:10.20944/preprints202609.0682.v1

140. Kim, S.; Mo, S.; Rizve, M.N.; Xu, Y.; Liu, D.; Shin, J.; Hinz, T. Rethinking Prompt Design for Inference-time Scaling in Text-to-Visual Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 22090–22099.
141. Jiang, L.; Chen, R.; Gao, C.; Niu, D. RAISE: Requirement-Adaptive Evolutionary Refinement for Training-Free Text-to-Image Alignment. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2026, pp. 22038–22048.
142. Long, D.X.; Wan, X.; Nakhost, H.; Lee, C.Y.; Pfister, T.; Arik, S.Ö. Vista: A test-time self-improving video generation agent. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 6021–6032.
143. Tyagi, A.; Boinpally, H.; Chen, J.; Gebert, D.; Hickson, S. Beyond Trial-and-Error: Agentic Optimization for Image-to-Video Adherence, 2026, [arXiv:cs.CV/2608.12290].
144. Lee, D.; Yoon, J.; Cho, J.; Bansal, M. Self-Correcting Text-to-Video Generation with Misalignment Detection and Localized Refinement. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 36464–36489. https://doi.org/10.18653/v1/2026.findings-acl.1817.
145. Guo, J.; Wei, H.; Zhang, Y.; Liu, Y.; Gong, Y.; Zhang, H.; Yang, X.; Zhong, Z. PhotoFlow: Agentic 3D Virtual Photography Missions, 2026, [arXiv:cs.CV/2605.23771].
146. Shen, S.; Liang, J.; Cai, C.; Geng, C.; Duan, H.; Zhang, X.; Hu, Q.; Zhai, G. Agentic Retoucher for Text-To-Image Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 29114–29125.
147. Zarei, A.; Pan, J.; Gwilliam, M.; Feizi, S.; Yang, Z. AgentComp: From Agentic Reasoning to Compositional Mastery in Text-to-Image Models, 2025, [arXiv:cs.CV/2512.09081].
148. Qu, L.; Li, H.; Wang, W.; Liu, X.; Li, J.; Nie, L.; Chua, T.S. Silmm: Self-improving large multimodal models for compositional text-to-image generation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 18497–18508.
149. Huang, K.; Huang, Y.; Ning, X.; Lin, Z.; Wang, Y.; Liu, X. Genmac: compositional text-to-video generation with multi-agent collaboration. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 5049–5057.
150. Venkatesh, K.; Dunlop, C.; Yanardag, P. CREA: A Collaborative Multi-Agent Framework for Creative Image Editing and Generation. In Proceedings of the Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025; Belgrave, D.; Zhang, C.; Montoya, L.N.; Lin, H.; Pascanu, R.; Koniusz, P.; Ghassemi, M.; Chen, N.; Ruíz, I.V.M.; Loaiza-Bonilla, A., Eds., 2025.
151. Li, S.; Kallidromitis, K.; Gokul, A.; Koneru, A.; Kato, Y.; Kozuka, K.; Grover, A. Reflect-dit: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 15657–15668.
152. Zhuo, L.; Zhao, L.; Paul, S.; Liao, Y.; Zhang, R.; Xin, Y.; Gao, P.; Elhoseiny, M.; Li, H. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 15329–15339.
153. Sun, J.; Fu, D.; Hu, Y.; Wang, S.; Rassin, R.; Juan, D.C.; Alon, D.; Herrmann, C.; Van Steenkiste, S.; Krishna, R.; et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 5920–5945.
154. Zheng, D.; Lee, H.; Zhang, M.; Feng, K.; Guo, Z.; Zhang, R.; Li, H. InterleaveThinker: Reinforcing Agentic Interleaved Generation, 2026, [arXiv:cs.CV/2606.13679].
155. Liu, J.; Feng, R.; Wang, Y.; Zeng, W.; Jin, X. Generation Navigator: A State-Aware Agentic Framework for Image Generation, 2026, [arXiv:cs.CV/2605.17969].
156. Wang, L.; Xu, Z.; Xing, X.; Cheng, Y.; Zhao, Z.; Li, D.; Hang, T.; Li, Z.; Tao, J.; Wang, Q.; et al. PromptEnhancer: Taming Your Rewriter for Text-to-Image Generation via Fine-Grained Reward. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026, pp. 14895–14904.
157. Chu, M.; Yang, S.; Che, H.; Zhang, S.; Zhang, X.; Yu, S.; Gui, H.; Rao, Z.; Tu, D.; Liu, R.; et al. VisionDirector: Vision-Language Guided Closed-Loop Refinement for Generative Image Synthesis. In Proceedings of the

 doi:10.20944/preprints202609.0682.v1

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026, pp. 9203–9212.
158. Jiang, K.; Wang, Y.; Zhou, J.; Li, P.; Liu, Z.; Xie, C.W.; Chen, Z.; Zheng, Y.; Zhang, W. GenAgent: Scaling Text-to-Image Generation via Agentic Multimodal Reasoning, 2026, [arXiv:cs.CV/2601.18543].
159. Chen, S.; Xing, Z.; Ye, T.; Geng, X.; Lin, Y.; Lai, J.; He, X.; Zhai, F.; Gao, J.; Zhu, L. GenEvolve: Self-Evolving Image Generation Agents via Tool-Orchestrated Visual Experience Distillation, 2026, [arXiv:cs.CV/2605.21605].
160. Garg, S.; Singh, A.; Nayak, G.K. SIDiffAgent: Self-Improving Diffusion Agent, 2026, [arXiv:cs.AI/2602.02051].
161. Chen, W.; Yu, K.; Tian, B.; Song, J.; Liang, S.; Jia, H.; Cheng, K.; Li, H.; Yuan, K.; Wang, L.; et al. MemoGen: Can Past Experience Improve Future Text-to-Image Generation?, 2026, [arXiv:cs.CV/2606.03243].
162. Huang, Z.; Wu, J.Z.; Wang, Z.; Cao, T.; Chen, J.; Fidler, S.; Ling, H.; Ren, X. APE: Agentic Prompt Enhancer for Image Generation and Editing, 2026, [arXiv:cs.CV/2606.00204].
163. Luo, M.; Zhang, Y.; Li, Y.; Wang, X.; Wu, F.; Lee, T.Y.; Deussen, O.; Dong, W. Self-Reasoning Agentic Framework for Narrative Product Grid-Collage Generation, 2026, [arXiv:cs.CV/2604.16958].
164. Li, S.; Zhao, Y.; Bhalerao, P.; Ignat, O. When Cultures Move: Measuring and Improving Multicultural Text-to-Video Generation, 2026, [arXiv:cs.CV/2605.16716].
165. Das, D.; Nigam, L.; Bahadur, S.K.J.; Dhar, G. Genflow Ad Studio: A Compound AI Architecture for Brand-Aligned, Self-Correcting Video Generation. In Proceedings of the Proceedings of the ACM Conference on AI and Agentic Systems, 2026, pp. 1193–1198.
166. Savytski, D.; Lei, A.; Liu, H.; Yang, W.; Liang, S.; Liu, A.; Zhao, Z. Bridging Creative Intent and Visual Quality: Creator-Driven Recurrent Video Generation with Agentic Feedback Loops, 2026, [arXiv:cs.CV/2606.18591].
167. Wang, W.; Zhao, C.; Chen, H.; Chen, Z.; Zheng, K.; Shen, C. Autostory: Generating diverse storytelling images with minimal human effort. *International Journal of Computer Vision* **2025**, *133*, 3083–3104.
168. Sun, W.; Wang, Z.; Hu, Z.; Wang, C.; Li, H.; Chen, W. MUSE: A Multi-agent Framework for Unconstrained Story Envisioning via Closed-Loop Cognitive Orchestration, 2026, [arXiv:cs.CV/2602.03028].
169. Hu, P.; Jiang, J.; Chen, J.; Han, M.; Liao, S.; Chang, X.; Liang, X. StoryAgent: Customized Storytelling Video Generation via Multi-Agent Collaboration, 2024, [arXiv:cs.CV/2411.04925].
170. Cheng, J.; Lu, X.; Li, H.; Zai, K.L.; Yin, B.; Cheng, Y.; Yan, Y.; Liang, X. AutoStudio: Crafting Consistent Subjects in Interactive Story Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2026, pp. 4770–4779.
171. Cheng, J.; Yin, B.; Cai, K.; Huang, M.; Li, H.; He, Y.; Lu, X.; Li, Y.; Li, Y.; Cheng, Y.; et al. Theatergen: Character management with llm for consistent multi-turn image generation. *arXiv preprint arXiv:2404.18919* **2024**.
172. Akdemir, K.; Kazimi, T.; Yanardag, P. Audit & Repair: An Agentic Framework for Consistent Story Visualization in Text-to-Image Diffusion Models, 2025, [arXiv:cs.CV/2506.18900].
173. Yin, S.; Liu, J.; Tang, X.; Shakib, Y.; Liu, Q. S2ED: From Story to Executable Descriptions for Consistency-Aware Story Illustration, 2026, [arXiv:cs.AI/2605.22448].
174. Gao, B.; Liu, C.; Miao, Y.; Ma, S.; Lim, S.N. BOOKAGENT: Orchestrating Safety-Aware Visual Narratives via Multi-Agent Cognitive Calibration. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026, 2026, pp. 2275–2292.
175. Lai, J.; Lu, Z.; He, J.; Quan, R.; Zhao, W.; Yang, Q.; Chen, Q.; Lin, Q.; Li, C.; Gao, T.; et al. VisionCreator: A Native Visual-Generation Agentic Model with Understanding, Thinking, Planning and Creation, 2026, [arXiv:cs.CV/2603.02681]. https://doi.org/10.48550/arXiv.2603.02681.
176. Zuo, J.; Zuo, H.; Zhang, S.; Wang, X.; Li, C.; Sang, N.; Gao, C.; Bai, X. FilmWorld: Agentic Novel-to-Film Generation through Dynamic Cinematic World Modeling. *arXiv preprint arXiv:2607.19038* **2026**.
177. Huang, K.; Huang, Y.; Wang, X.; Lin, Z.; Ning, X.; Wan, P.; Zhang, D.; Wang, Y.; Liu, X. FilMaster: Bridging Cinematic Principles and Generative AI for Automated Film Generation, 2025, [arXiv:cs.CV/2506.18899].
178. He, L.; Song, Y.; Huang, H.; Liu, P.; Tang, Y.; Aliaga, D.; Zhou, X. Kubrick: Multimodal agent collaborations for synthetic video generation. *arXiv preprint arXiv:2408.10453* **2024**.
179. Wei, Z.; Li, M.; Zhang, Z.; Yuan, R.; Hui, P.; Qu, H.; Evans, J.; Agrawala, M.; Rao, A. Hollywood Town: Long-Video Generation via Cross-Modal Multi-Agent Orchestration, 2025, [arXiv:cs.MA/2510.22431].
180. Shi, Y.; Yan, W.; Huang, N.; Chen, Y.; Zhang, C.; He, T.; Yeo, S.Y.; Li, M. One Sentence, One Drama: Personalized Short-Form Drama Generation via Multi-Agent Systems, 2026, [arXiv:cs.CV/2605.22144].
181. Mu, C.; He, X.; Yang, Q.; Chen, W.; Yao, J.; Liu, H.; Yi, Z.; Zhao, B.; Chen, X.; Ma, R.; et al. The Script is All You Need: An Agentic Framework for Long-Horizon Dialogue-to-Cinematic Video Generation, 2026, [arXiv:cs.CV/2601.17737].

 doi:10.20944/preprints202609.0682.v1

182. Zeng, Q.; Cai, K.; Chen, R.; Lv, Q.; Wang, K. CoAgent: Collaborative Planning and Consistency Agent for Coherent Video Generation, 2025, [arXiv:cs.CV/2512.22536].
183. Song, Y.; Song, Y.; Losier, N.; Hodson, N.; Jin, Y.; Zhu, R.; Xu, Y.; Vlasic, D.; Claassen, C.; Leon, J.; et al. Co-Director: Agentic Generative Video Storytelling, 2026, [arXiv:cs.AI/2604.24842].
184. Huang, L.; He, S.; Zhou, H.; Nie, L.; Xia, L.; Huang, C. ViMax: Agentic Video Generation, 2026, [arXiv:cs.CV/2606.07649].
185. Long, D.X.; Song, Y.; Kan, M.Y.; Pfister, T.; Le, L.T. A^2RD: Agentic Autoregressive Diffusion for Long Video Consistency, 2026, [arXiv:cs.CV/2605.06924].
186. Lai, Y.; Shao, T.; Zhou, K.; Dou, W.; Zhu, S.; Wang, J. GroundShot: Visually Consistent Multi-Shot Long Video Generation via Entity-Grounded Shot Scheduling, 2026, [arXiv:cs.CV/2606.20799].
187. Xie, T.; Huang, Z.; Wang, M.; Huang, X.; Zhou, J.; Gong, M.; Yi, Z. CineAGI: Character-Consistent Movie Creation through LLM-Orchestrated Multi-Modal Generation and Cross-Scene Integration, 2026, [arXiv:cs.MM/2604.23579].
188. Li, Y.; Shi, H.; Hu, B.; Wang, L.; Zhu, J.; Xu, J.; Zhao, Z.; Zhang, M. Anim-director: A large multimodal model powered agent for controllable animation video generation. In Proceedings of the SIGGRAPH Asia 2024 Conference Papers, 2024, pp. 1–11.
189. Zhang, L.; Xu, B.; Yang, S.; Yin, M.; Liu, J.; Xu, C.; Wang, S.; Wu, Y.; Hong, Y.; Zhang, Z.; et al. AniME: Adaptive Multi-Agent Planning for Long Animation Generation. In *Proceedings of the SIGGRAPH Asia 2025 Posters*; 2025; pp. 1–3.
190. Shi, H.; Li, Y.; Chen, X.; Wang, L.; Hu, B.; Zhang, M. AniMaker: Multi-Agent Animated Storytelling with MCTS-Driven Clip Generation. In Proceedings of the Proceedings of the SIGGRAPH Asia 2025 Conference Papers, 2025, pp. 1–11.
191. Wang, W.F.; Lu, C.T.; Ng, J.P.; Chiu, Y.T.; Lee, T.Y.; Wang, M.; Chen, B.Y.; Chen, X.A. AnimAgents: Coordinating Multi-Stage Animation Pre-Production with Human-Multi-Agent Collaboration, 2025, [arXiv:cs.HC/2511.17906].
192. Yan, H.; Liu, S.; Wang, T.; Zhang, X.; Zhong, Y.; Chen, J.; Zhang, L.; Li, B. AnimeAgent: Is the Multi-Agent via Image-to-Video models a Good Disney Storytelling Artist?, 2026, [arXiv:cs.CV/2602.20664].
193. Wang, Q.; Huang, Z.; Jia, R.; Debevec, P.; Yu, N. MAViS: A multi-agent framework for long-sequence video storytelling. In Proceedings of the Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2026, pp. 2273–2295.
194. Xu, X.; Mei, J.; Li, C.; Wu, Y.; Yan, M.; Lai, S.; Zhang, J.; Wu, M. MM-StoryAgent: Immersive Narrated Storybook Video Generation with a Multi-Agent Paradigm across Text, Image and Audio, 2025, [arXiv:cs.CL/2503.05242].
195. Tang, X.; Lei, X.; Zhu, C.; Chen, S.; Yuan, R.; Li, Y.; Oh, C.; Zhang, G.; Huang, W.; Benetos, E.; et al. AutoMV: An Automatic Multi-Agent System for Music Video Generation, 2025, [arXiv:cs.MM/2512.12196].
196. Zhang, Y.; Xu, X.; Xu, X.; Liu, L.; Chen, Y. Long-Video Audio Synthesis with Multi-Agent Collaboration, 2025, [arXiv:cs.CV/2503.10719].
197. Liu, J.; Yang, L.; Luo, H.; Wang, F.; Li, H.; Wang, M. Preacher: Paper-to-video agentic system. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 17129–17139.
198. Park, J.I.; Taneja, M.; Wang, Q.; Kang, D. Stealing Creator's Workflow: A Creator-Inspired Agentic Framework with Iterative Feedback Loop for Improved Scientific Short-form Generation, 2025, [arXiv:cs.CL/2504.18805].
199. Liang, X.; Li, B.; Chen, Z.; Zheng, H.; Ma, Z.; Wang, D.; Tian, C.; Wang, Q. Videoagent: Personalized synthesis of scientific videos. In Proceedings of the Proceedings of the 2026 International Conference on Multimedia Retrieval, 2026, pp. 1803–1811.
200. Yan, L.; Wu, J.; Xie, D.; Shi, W.; Xia, D.; Huang, J. Beyond End-to-End Video Models: An LLM-Based Multi-Agent System for Educational Video Generation. In Proceedings of the Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, New York, NY, USA, 2026; KDD '26, p. 8369–8378. https://doi.org/10.1145/3770855.3818323.
201. Chen, Y.; Lin, K.Q.; Shou, M.Z. Code2Video: A Code-centric Paradigm for Educational Video Generation, 2025, [arXiv:cs.CV/2510.01174].
202. Wang, Z.; Ma, J.; Grinspun, E.; Grossman, T.; Wang, B. Script2Screen: Supporting Dialogue-Centric Scriptwriting with Interactive Audiovisual Generation. In Proceedings of the Proceedings of the 31st International Conference on Intelligent User Interfaces, New York, NY, USA, 2026; IUI '26, p. 1496–1513. https://doi.org/10.1145/3742413.3789075.

203. Song, Z. Sima 1.0: A Collaborative Multi-Agent Framework for Documentary Video Production, 2026, [arXiv:cs.MA/2604.07721].
204. Zhu, Z.; Wang, R.; Lyu, S.; Zhang, M.; Wu, B. BrandFusion: A Multi-Agent Framework for Seamless Brand Integration in Text-to-Video Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings, June 2026, pp. 8661–8671.
205. Wei, J.; Li, K.; Lao, T.; Wang, H.; Wang, L.; Shan, C.; Si, C. PosterCopilot: Toward Layout Reasoning and Controllable Editing for Professional Graphic Design, 2025, [arXiv:cs.CV/2512.04082].
206. Xu, R.; Zhu, X.; Ying, J.; Dong, D.; Ji, Y.; Tan, X. MUSE: Agentic 3D Scene Authoring via Memory-Grounded Incremental Requirement Satisfaction, 2026, [arXiv:cs.CV/2606.14168].
207. Liu, C.; Wang, X.; Chen, H.; Zhao, Y.; Yang, M.H.; Jeni, L.A. SimWorlds: A Multi-Agent System for Dynamic 3D Scene Creation, 2026, [arXiv:cs.AI/2607.01766].
208. Cudlenco, N.; Masala, M.; Leordeanu, M. Authoring for Living Worlds: Tool-Constrained LLM Agents for Executable Multi-Actor Scenarios, 2026, [arXiv:cs.CV/2604.10383].
209. Shen, F.; Xie, C.; Wang, L.; Zhang, Z.; Jiang, X.; Du, X.; Tang, J. IMAGAgent: Orchestrating Multi-Turn Image Editing via Constraint-Aware Planning and Reflection. *CoRR* **2026**, *abs/2603.29602*, [2603.29602]. https://doi.org/10.48550/ARXIV.2603.29602.
210. Xu, Z.; Duan, H.; Nie, Y.; Du, M.; Wu, S.; Min, X.; Zheng, T.; Zhang, J.; Xu, S.; Chen, J.; et al. EditRefiner: A Human-Aligned Agentic Framework for Image Editing Refinement. *CoRR* **2026**, *abs/2605.07457*, [2605.07457]. https://doi.org/10.48550/ARXIV.2605.07457.
211. Pu, Y.; Zheng, H.; Mo, Z.; Pang, Z.; Zhang, H.; Fan, T.; Wu, S.; Wei, J. CAMEO: A Conditional and Quality-Aware Multi-Agent Image Editing Orchestrator. *arXiv preprint arXiv:2604.03156* **2026**.
212. Zhao, Y.; Ye, Y.; Liu, X.; Shieh, M.Q.; Bui, T. ImageEdit-R1: Boosting Multi-Agent Image Editing via Reinforcement Learning. *CoRR* **2026**, *abs/2603.08059*, [2603.08059]. https://doi.org/10.48550/ARXIV.2603.08059.
213. Ye, R.; Zhang, J.; Liu, Z.; Zhu, Z.; Yang, S.; Li, L.; Fu, T.; Dernoncourt, F.; Zhao, Y.; Zhu, J.; et al. Agent Banana: High-Fidelity Image Editing with Agentic Thinking and Tooling. *CoRR* **2026**, *abs/2602.09084*, [2602.09084]. https://doi.org/10.48550/ARXIV.2602.09084.
214. Rajan, A.S.; Singh, K.K.; Lee, Y.J. From Plans to Pixels: Learning to Plan and Orchestrate for Open-Ended Image Editing. *CoRR* **2026**, *abs/2605.15181*, [2605.15181]. https://doi.org/10.48550/ARXIV.2605.15181.
215. Guo, Z.; Liu, X.; Ma, L.; Wang, C.; He, Y.; Fu, X.; Fu, J.; Shan, X.; Guo, S.; Liu, L.; et al. GMO-E^{2}DIT: Grounded Multi-Operation Editing for E-Commerce Images. *arXiv preprint arXiv:2607.00920* **2026**.
216. Qiu, Z.; Chen, K.; Wang, X.; Xia, Y.; Seneviratne, S.; Halgamuge, S.K. MSRAMIE: Multimodal Structured Reasoning Agent for Multi-instruction Image Editing. *CoRR* **2026**, *abs/2603.16967*, [2603.16967]. https://doi.org/10.48550/ARXIV.2603.16967.
217. Zeng, Z.; Hua, H.; Luo, J. MIRA: Multimodal Iterative Reasoning Agent for Image Editing. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings, June 2026, pp. 9563–9573.
218. Zhang, Y.; Cheng, Z.; Zhao, Z.; Li, Z.; Liu, B.; Liu, Q.; Cai, J.; Chen, X.; Tu, Z.; Chu, D.; et al. PSBench: Editing Image via GUI Agents in Photoshop. In Proceedings of the Forty-third International Conference on Machine Learning, 2026.
219. Lin, Y.; Wang, L.; Lin, K.; Lin, Z.; Gong, K.; Li, W.; Lin, B.; Li, Z.; Zhang, S.; Peng, Y.; et al. Jarvisevo: Towards a self-evolving photo editing agent with synergistic editor-evaluator optimization. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 27291–27302.
220. Yao, M.; You, Z.; Tam, K.M.; Wang, M.; Xue, T. PhotoAgent: Exploratory Visual Aesthetic Planning with Large Vision Models. *arXiv preprint arXiv:2602.22809* **2026**.
221. Chang, Z.; Duan, Z.; Zhang, J.; Guo, C.; Liu, S.; Chun, H.; Park, H.; Liu, Z.; Li, C. PerTouch: VLM-Driven Agent for Personalized and Semantic Image Retouching. In Proceedings of the Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026; Koenig, S.; Jenkins, C.; Taylor, M.E., Eds. AAAI Press, 2026, pp. 2752–2759. https://doi.org/10.1609/AAAI.V40I4.37264.
222. Chen, H.; Li, W.; Gu, J.; Ren, J.; Chen, S.; Ye, T.; Pei, R.; Zhou, K.; Song, F.; Zhu, L. RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems

2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024; Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.M.; Zhang, C., Eds., 2024.

223. Zhu, K.; Gu, J.; You, Z.; Qiao, Y.; Dong, C. An Intelligent Agentic System for Complex Image Restoration Problems. In Proceedings of the The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net, 2025.
224. Jiang, X.; Li, G.; Chen, B.; Zhang, J. Multi-Agent Image Restoration. *Int. J. Comput. Vis.* **2026**, *134*, 205. https://doi.org/10.1007/S11263-026-02792-5.
225. Hu, S.; Xu, J.; Samaras, D.; Le, H. Domain-Grounded Candidate Selection for Agentic Image Editing: A Shadow Removal Case. *arXiv preprint arXiv:2608.06075* **2026**.
226. Lin, Y.; Lin, Z.; Chen, H.; Pan, P.; Li, C.; Chen, S.; Wen, K.; Jin, Y.; Li, W.; Ding, X. JarvisIR: Elevating Autonomous Driving Perception with Intelligent Image Restoration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 22369–22380.
227. Li, B.; Li, X.; Lu, Y.; Chen, Z. Hybrid Agents for Image Restoration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2026, pp. 22636–22647.
228. Zhu, F.; Xie, S.; Zeng, Y.; Liu, M.; Zuo, W. OPERA: An Agent for Image Restoration with End-to-End Joint Planning-Execution Optimization. *CoRR* **2026**, *abs/2605.22104*, [2605.22104]. https://doi.org/10.48550/ARXIV.2605.22104.
229. Cui, S.; Ji, F.; Sun, G.; Guo, Y.; Tang, X.; Li, J.; Xu, F. Self-Evolving Agentic Image Restoration via Deliberate Planning and Intuitive Execution. *CoRR* **2026**, *abs/2606.28971*, [2606.28971]. https://doi.org/10.48550/ARXIV.2606.28971.
230. Yu, Y.; Zeng, Z.; Xiao, Z.; Zhou, Z.; Hua, H.; Xiong, W.; Luo, J. Aurora: Unified Video Editing with a Tool-Using Agent. *CoRR* **2026**, *abs/2605.18748*, [2605.18748]. https://doi.org/10.48550/ARXIV.2605.18748.
231. Shen, Y.; Li, C.; Unberath, M. Text-Driven Reasoning Video Editing via Reinforcement Learning on Digital Twin Representations. *CoRR* **2025**, *abs/2511.14100*, [2511.14100]. https://doi.org/10.48550/ARXIV.2511.14100.
232. Song, Y.; Zhong, H.; Lin, K.Q.; Wang, H.; Shou, M.Z. Soap2Soap: Long Cinematic Video Remaking via Multi-Agent Collaboration. *CoRR* **2026**, *abs/2605.17423*, [2605.17423]. https://doi.org/10.48550/ARXIV.2605.17423.
233. Zhou, H.; Huang, L.; Wang, J.; Zhou, B.; Wu, S.; Xia, L.; Huang, C. VideoAgent: All-in-One Framework for Video Understanding and Editing. *CoRR* **2026**, *abs/2606.23327*, [2606.23327]. https://doi.org/10.48550/ARXIV.2606.23327.
234. Yan, L.; Zhang, Y.; Pan, B.; Zheng, X.; Qian, J.; Wu, A.; Li, W.; Lyu, C. Crayotter: Traceable Multi-Agent Workflows for Long-Form Video Editing. *CoRR* **2026**, *abs/2606.07636*, [2606.07636]. https://doi.org/10.48550/ARXIV.2606.07636.
235. Yan, L.; Lin, J.; Zhang, Y.; Pan, B.; Li, W.; Lyu, C.; Zhou, L.; Gurrin, C. Crayotter: Learning Long-Horizon Video Editing Agents via Group-Relative Preference Backpropagation. *arXiv preprint arXiv:2608.02694* **2026**.
236. Lin, Z.; Wang, H.; Xu, Z.; Dai, S.; Dong, H.; Wang, X.; Tang, Y.Y.; Wang, Y.; Wang, Q.; Huang, L. GLANCE: A Global-Local Coordination Multi-Agent Framework for Music-Grounded Non-Linear Video Editing. *CoRR* **2026**, *abs/2604.05076*, [2604.05076]. https://doi.org/10.48550/ARXIV.2604.05076.
237. Zhao, S.; Hu, Y.; Shan, Y.; Wei, Y.; Cun, X. CutClaw: Agentic Hours-Long Video Editing via Music Synchronization. *CoRR* **2026**, *abs/2603.29664*, [2603.29664]. https://doi.org/10.48550/ARXIV.2603.29664.
238. Li, K.; Li, M.; Chen, J.; Chen, J.; Zheng, Z.; Wang, S.; Chen, X. Direct: Video mashup creation via hierarchical multi-agent planning and intent-guided editing. *arXiv preprint arXiv:2604.04875* **2026**.
239. Yang, L.; Chen, Z.; Li, X.; Jia, P.; Long, L.; Yang, J. Agent-based Video Trimming. *CoRR* **2024**, *abs/2412.09513*, [2412.09513]. https://doi.org/10.48550/ARXIV.2412.09513.
240. Zhang, P.; Zhou, C.; Zhang, Z.; Liu, H.; Zhang, C.; Liu, J.; Zhou, X.; Chen, X.; Weng, S.; Li, S.; et al. A Benchmark and Multi-Agent System for Instruction-driven Cinematic Video Compilation. *CoRR* **2026**, *abs/2604.10456*, [2604.10456]. https://doi.org/10.48550/ARXIV.2604.10456.
241. Ding, Z.; Wang, X.; Chen, J.; Kristensson, P.O.; Shen, J. Prompt-Driven Agentic Video Editing System: Autonomous Comprehension of Long-Form, Story-Driven Media. *CoRR* **2025**, *abs/2509.16811*, [2509.16811]. https://doi.org/10.48550/ARXIV.2509.16811.
242. Yu, K.; Dong, C.; Lin, L.; Loy, C.C. Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 2443–2452. https://doi.org/10.1109/CVPR.2018.00259.

 doi:10.20944/preprints202609.0682.v1

243. Lin, J.; Zhou, M.; Ma, Y.; Gao, Y.; Fei, C.; Chen, Y.; Yu, Z.; Ge, T. Autoposter: A highly automatic and content-aware design system for advertising poster generation. In Proceedings of the Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 1250–1260.
244. Yang, T.; Luo, Y.; Qi, Z.; Wu, Y.; Shan, Y.; Chen, C.W. Posterllava: Constructing a unified multi-modal layout generator with llm. *arXiv preprint arXiv:2406.02884* **2024**.
245. Chen, H.; Xu, X.; Li, W.; Ren, J.; Ye, T.; Liu, S.; Chen, Y.C.; Zhu, L.; Wang, X. Posta: A go-to framework for customized artistic poster generation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 28694–28704.
246. Zhang, Z.; Zhang, X.; Wei, J.; Xu, Y.; You, C. PosterGen: Aesthetic-Aware Multi-Modal Paper-to-Poster Generation Via Multi-Agent LLMs. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings, June 2026, pp. 9813–9823.
247. Hu, H.; Dang, C.; Liu, Y.; Kang, H.; He, C.; Li, W. PosterMELD: Multi-Agent Paper-to-Poster Generation for Controllable Design Diversity with Editable Print-Ready Outputs. *arXiv preprint arXiv:2608.02218* **2026**.
248. Vinaykumar, A.; Li, A.; Huang, S.; Liu, S. Any2Poster: Any-Source Poster Generation Across Modalities and Domains. *arXiv preprint arXiv:2606.02915* **2026**.
249. Shuai, X.; Tang, S.; Huang, Y.; Ding, H.; Tao, D. PSDesigner: Automated Graphic Design with a Human-Like Creative Workflow. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 10165–10175.
250. Wang, H.; Shimose, Y.; Takamatsu, S. Banneragency: Advertising banner design with multimodal llm agents. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 4304–4329.
251. Forouzandehmehr, N.; Maragheh, R.Y.; Kollipara, S.; Zhao, K.; Biswas, T.; Korpeoglu, E.; Achan, K. Cal-rag: Retrieval-augmented multi-agent generation for content-aware layout design. *arXiv preprint arXiv:2506.21934* **2025**.
252. Wang, T.; Ding, L.; Tao, Z.; Zhan, Y.; Ma, Z.; Wu, W.; Lei, Y.; Feng, Y.; Wang, J.; Wu, Y.; et al. EvoDiagram: Agentic Editable Diagram Creation via Design Expertise Evolution. *arXiv preprint arXiv:2604.09568* **2026**.
253. Huang, S.; Zhou, Y.; Gao, Y.; Yin, Z.; Bai, J.; Liu, X.; Chellappa, R.; Lau, C.P.; Peng, C.; Nag, S.; et al. SciFig: Towards Automating Editable Figure Generation for Scientific Papers. *arXiv preprint arXiv:2601.04390* **2026**.
254. Xu, X.; Xu, X.; Chen, S.; Chen, H.; Zhang, F.; Chen, Y.C. Pregenie: An agentic framework for high-quality visual presentation generation. *arXiv preprint arXiv:2505.21660* **2025**.
255. Zheng, H.; Mo, G.; Yan, X.; Yuan, Q.; Zhang, W.; Chen, X.; Lu, Y.; Lin, H.; Han, X.; Sun, L. Deeppresenter: Environment-grounded reflection for agentic presentation generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026, 2026, pp. 31545–31558.
256. Fang, S.; Wu, C.; Zhang, Z. SeaSlides: Semantic Abstraction Layer for Agentic Slide Generation. *arXiv preprint arXiv:2608.03298* **2026**.
257. Jin, Y.; Xu, Y.; Zhu, J.; Yang, Y. MemSlides: A Hierarchical Memory Driven Agent Framework for Personalized Slide Generation with Multi-turn Local Revision. *arXiv preprint arXiv:2606.17162* **2026**.
258. Yuan, M.; Chen, J.; Hu, Y.; Feng, S.; Xie, M.; Mohammadi, G.; Xing, Z.; Quigley, A. Towards Human–AI Synergy in UI Design: Supporting Iterative Generation with LLMs. *ACM Transactions on Computer-Human Interaction* **2026**, *33*, 1–45.
259. Zeng, W.; An, F.; Liu, Z.; Zhao, J. GameUIAgent: An LLM-Powered Framework for Automated Game UI Design with Structured Intermediate Representation. *arXiv preprint arXiv:2603.14724* **2026**.
260. Wu, W.; Xu, Z.; Zhang, Z.; Zhao, Y.; Tang, H. PresentAgent-2: Towards Generalist Multimodal Presentation Agents. *arXiv preprint arXiv:2605.11363* **2026**.
261. Yu, T.; Zhang, M.; Cui, Z.; Wang, H.; Luo, Z.; Chai, S.; Gong, J.; Peng, Y.; Zhou, Y.; Yang, Y.; et al. PaperX: A Unified Framework for Multimodal Academic Presentation Generation with Scholar DAG. *arXiv preprint arXiv:2602.03866* **2026**.
262. Ma, Q.; Xiao, J.; Wang, S.; Tian, Z.; Feng, W.; Wang, S.; Guo, C.; Chang, S.; Liu, Q.; Zhang, Z. OmniPresent: Generating Coherent Presentation Suites from Scientific Papers. *arXiv preprint arXiv:2607.02590* **2026**.
263. Gal, R.; Haviv, A.; Alaluf, Y.; Bermano, A.H.; Cohen-Or, D.; Chechik, G. Comfygen: Prompt-adaptive workflows for text-to-image generation. *arXiv preprint arXiv:2410.01731* **2024**.
264. Xu, Z.; Yang, X.; Wang, Y.; Hu, Q.; Wu, Z.; Wang, L.; Luo, W.; Zhang, K.; Hu, B.; Zhang, M. Comfyui-copilot: An intelligent assistant for automated workflow development. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 2025, pp. 632–643.

265. Huang, O.; Ma, Y.; Zhao, Z.; Wu, M.; Ji, J.; Zhang, R.; Hu, Z.; Sun, X.; Ji, R. Comfygpt: A self-optimizing multi-agent system for comprehensive comfyui workflow generation. *arXiv preprint arXiv:2503.17671* **2025**.
266. Xu, Z.; Wang, Y.; Wang, L.; Luo, W.; Zhang, K.; Hu, B.; Zhang, M.; et al. Comfyui-r1: Exploring reasoning models for workflow generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026, 2026, pp. 2999–3013.
267. Li, Z.; Sun, L.; Ming, R.; Zhang, H.; Paudel, D.P.; Van Gool, L.; Gu, J. Knowledge-Centric Agents for Workflow Generation in ComfyUI. *arXiv preprint arXiv:2607.15845* **2026**.
268. Guo, L.; Xu, X.; Wang, L.; Lin, J.; Zhou, J.; Zhang, Z.; Su, B.; Chen, Y. ComfyMind: Toward General-Purpose Generation via Tree-Based Planning and Reactive Feedback. In Proceedings of the Advances in Neural Information Processing Systems; Belgrave, D.; Zhang, C.; Lin, H.; Pascanu, R.; Koniusz, P.; Ghassemi, M.; Chen, N., Eds. Curran Associates, Inc., 2025, Vol. 38, Main Conference, pp. 45128–45164. https://doi.org/10.52202/085713-1503.
269. Su, J.; Lan, Q.; Wang, Z.; Xia, Y.; Wen, H.; Duan, Y.; Xiao, X.; Shi, T.; Jingsong, Y.; He, L. ComfySearch: Autonomous Exploration and Reasoning for ComfyUI Workflows. *arXiv preprint arXiv:2601.04060* **2026**.
270. Li, Z.; Liu, D.; Liu, F.; Zhou, Y.; Wu, X.; Chen, J.; Xie, J.; Wu, X.; Sun, L. Comfyclaw: Self-evolving skill harnesses for image generation workflows. *arXiv preprint arXiv:2607.01709* **2026**.
271. Chen, J.; Li, X.; Chen, M.; Zhang, B.; Zhang, H.; Xu, Y.; Cui, Y.; Weng, F.; Ma, F.; Tian, Q.; et al. PairCoder++: Pair Programming as a Universal Paradigm for Verified Code-Driven Multimodal and Structured-Artifact Generation. *arXiv preprint arXiv:2607.01883* **2026**.
272. Xu, Y.; Kang, J.; Du, C.; Zhang, Q.; Zhou, W.; Wu, Y.; Li, T.; Song, Q. GVR-Coder: A Visual-Feedback Framework for Structured SVG Generation in Complex Document and Meeting Scenarios. *arXiv preprint arXiv:2607.28073* **2026**.
273. Wen, Z.; Cai, Y.; Lee, K.; Estep, S.; Sunshine, J.; Singh, A.; Chi, Y.; Ni, W. Feynman: Knowledge-Infused Diagramming Agent for Scalable Visual Designs. *arXiv preprint arXiv:2603.12597* **2026**.
274. Shao, C.; Liu, J.; Xu, F.; Li, Y. LiveFigure: Generating Editable Scientific Illustration with VLM Agents. *arXiv preprint arXiv:2605.23527* **2026**.
275. Wang, P.; Wang, Y.; Li, L.; Yang, Z.; Lin, K.Q.; Li, Y.; Cheng, Y. SceneCode: Executable World Programs for Editable Indoor Scenes with Articulated Objects. *arXiv preprint arXiv:2605.19587* **2026**.
276. Li, L.; Shen, C.; Xie, S.; Gu, C.; He, Z.; Meng, Y.; Yang, X.; Jiang, W.; Wang, Z. HDSL: A Hierarchical Domain-Specific Language for Structured 3D Indoor Scene Generation and Localized Editing with LLM Agents. *arXiv preprint arXiv:2606.09738* **2026**.
277. Bai, X.; Liang, H.; Galoaa, B.; Nandi, U.; Moezzi, S.; He, Y.; Ostadabbas, S. MoReGen: Multi-Agent Motion-Reasoning Engine for Code-based Text-to-Video Synthesis. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 7632–7642.
278. Li, H.; Ren, T.; Ma, X.; Qing, C.; Fang, Z.; He, S.; Guo, Z.; Wu, H.; Tian, J.; Zou, Y.; et al. VideoCoCo: Code-as-CoT for Physically-Consistent Video Generation via an Agentic Dual-Engine System. *arXiv preprint arXiv:2607.27380* **2026**.
279. Ma, J.; Agrawala, M. Mover: Motion verification for motion graphics animations. *ACM Transactions on Graphics (TOG)* **2025**, *44*, 1–17.
280. Gunturu, A.; Pearman, B.; Ihara, K.; Faraji, M.; Wang, B.; Kazi, R.H.; Suzuki, R. MapStory: Prototyping Editable Map Animations with LLM Agents. In Proceedings of the Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, 2025, pp. 1–20.
281. Liu, V.; Kazi, R.H.; Wei, L.Y.; Fisher, M.; Langlois, T.; Walker, S.; Chilton, L. Logomotion: Visually-grounded code synthesis for creating and editing animation. In Proceedings of the Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–16.
282. Jiang, W.; Cai, Z.; Shao, Y.; Wang, C.; Han, B.; Song, Z.; Chen, K.; An, S.; Yang, X.; Yang, Z. ManimAgent: Self-Evolving Multimodal Agents for Visual Education. *arXiv preprint arXiv:2606.30296* **2026**.
283. Wu, X.; Yu, D.; Huang, Y.; Russakovsky, O.; Arora, S. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems* **2024**, *37*, 86004–86047.
284. Kamath, A.; Chang, K.W.; Krishna, R.; Zettlemoyer, L.; Hu, Y.; Ghazvininejad, M. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853* **2025**.
285. Li, N.; Hu, G.; Qiao, W.; Ba, Y.; Hong, Q.; Shen, S.; Wang, J.; Zhou, F.; Kang, J.; Shang, X.; et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091* **2026**.

286. Chen, K.; Lin, Z.; Xu, Z.; Shen, Y.; Yao, Y.; Rimchala, J.; Zhang, J.; Huang, L. R2i-bench: Benchmarking reasoning-driven text-to-image generation. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 12606–12641.
287. Pan, Y.; He, X.; Mao, C.; Han, Z.; Jiang, Z.; Zhang, J.; Liu, Y. Ice-bench: A unified and comprehensive benchmark for image creating and editing. In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2025, pp. 16586–16596.
288. Hua, H.; Zeng, Z.; Song, Y.; Tang, Y.; He, L.; Aliaga, D.; Xiong, W.; Luo, J. Mmigbench: Towards comprehensive and explainable evaluation of multi-modal image generation models. *arXiv preprint arXiv:2505.19415* **2025**.
289. Oshima, Y.; Miyake, D.; Matsutani, K.; Iwasawa, Y.; Suzuki, M.; Matsuo, Y.; Furuta, H. Multibanana: A challenging benchmark for multi-reference text-to-image generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 448–460.
290. Zhou, Z.; Lai, Z.; Wang, R.; Yang, Y.; Xing, Z.; Yang, Y.; Dai, Q.; Qiu, L.; Luo, C. AVGen-Bench: A Task-Driven Benchmark for Multi-Granular Evaluation of Text-to-Audio-Video Generation. *CoRR* **2026**, *abs/2604.08540*, [2604.08540]. https://doi.org/10.48550/ARXIV.2604.08540.
291. Sun, K.; Huang, K.; Liu, X.; Wu, Y.; Xu, Z.; Li, Z.; Liu, X. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE, 2025, pp. 8406–8416. https://doi.org/10.1109/CVPR52734.2025.00787.
292. Wang, S.; Li, N.; Hu, G.; Qi, H.; Ding, F.; Qiao, W.; Wang, J.; Lv, X.; Han, P.; Li, Z.; et al. FilmBench: A film-grade benchmark for cinematic video generation. *arXiv preprint arXiv:2607.24241* **2026**.
293. Wei, J.; Zhang, X.; Li, Y.; Wang, Y.; Zhang, Y.; Chen, Z.; Tang, Z.; Xu, W.; Liu, Z. UniVBench: Towards Unified Evaluation for Video Foundation Models. *CoRR* **2026**, *abs/2602.21835*, [2602.21835]. https://doi.org/10.48550/ARXIV.2602.21835.
294. Wei, Y.; Han, Y.; Chen, Z.; Li, Y.; Jiang, K.; Liu, Z.; Li, Q.; Qing, Z.; Wang, X.; Xing, Z.; et al. MSAVBench: Towards Comprehensive and Reliable Evaluation of Multi-Shot Audio-Video Generation. *CoRR* **2026**, *abs/2605.20183*, [2605.20183]. https://doi.org/10.48550/ARXIV.2605.20183.
295. Shi, H.; Li, Y.; Deng, N.; Xu, Z.; Chen, X.; Wang, L.; Hu, B.; Zhang, M. MSVBench: Towards Human-Level Evaluation of Multi-Shot Video Generation. *CoRR* **2026**, *abs/2602.23969*, [2602.23969]. https://doi.org/10.48550/ARXIV.2602.23969.
296. He, R.; Wei, M.; Yang, Z.; Ordonez, V. EntityBench: Towards Entity-Consistent Long-Range Multi-Shot Video Generation. *CoRR* **2026**, *abs/2605.15199*, [2605.15199]. https://doi.org/10.48550/ARXIV.2605.15199.
297. Zhuang, C.; Huang, A.; Cheng, W.; Wu, J.; Hu, Y.; Liao, J.; Huang, Z.; Wang, H.; Liao, X.; Cai, W.; et al. ViStoryBench: Comprehensive Benchmark Suite for Story Visualization. *CoRR* **2025**, *abs/2505.24862*, [2505.24862]. https://doi.org/10.48550/ARXIV.2505.24862.
298. Liu, T.; Shi, Y.; Zhu, X.; Tang, J.; Yang, L.; Wang, Q.; Zhang, Z.; Tang, Y.; Wang, F.; Dong, Y.; et al. LongAV-Compass: Towards Unified Evaluation of Minute-Scale Audio-Visual Generation Across T2AV, I2AV, and V2AV. *CoRR* **2026**, *abs/2605.26244*, [2605.26244]. https://doi.org/10.48550/ARXIV.2605.26244.
299. Chen, J.; Chen, Q.; Zhang, J.; Wu, Y.; Li, Y.; Zhang, X.; Zhou, W.; Ma, C. DirectorBench: Diagnosing Long-Form Video Generation with Personalized Multi-Agent Evaluation. *CoRR* **2026**, *abs/2605.30090*, [2605.30090]. https://doi.org/10.48550/ARXIV.2605.30090.
300. Zhang, H.; Wu, D.; Liu, B.; Zhong, L.; Wei, Y.; Ye, X.; Liu, N.; Liang, Y. MuSS: A Large-Scale Dataset and Cinematic Narrative Benchmark for Multi-Shot Subject-to-Video Generation. *CoRR* **2026**, *abs/2604.23789*, [2604.23789]. https://doi.org/10.48550/ARXIV.2604.23789.
301. Feng, K.; Zhang, M.; Chen, S.; Lin, Y.; Fan, K.; Jiang, Y.; Li, H.; Zheng, D.; Wang, C.; Yue, X. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767* **2026**.
302. Yang, S.; Zhong, H.; Zhang, R.; Zhao, X.; Li, S.; Zheng, K.; Yang, X.; Wang, Z.; Tang, Z.; Li, Y.; et al. EvalVerse: Pipeline-Aware and Expert-Calibrated Benchmarking for Professional Cinematic Video Generation. *CoRR* **2026**, *abs/2605.23271*, [2605.23271]. https://doi.org/10.48550/ARXIV.2605.23271.
303. Meng, F.; Shao, W.; Luo, L.; Wang, Y.; Chen, Y.; Lu, Q.; Yang, Y.; Yang, T.; Zhang, K.; Qiao, Y.; et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802* **2024**.
304. Niu, Y.; Ning, M.; Zheng, M.; Jin, W.; Lin, B.; Jin, P.; Liao, J.; Feng, C.; Meng, F.; Ning, K.; et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265* **2025**.

 doi:10.20944/preprints202609.0682.v1

305. Zhang, D.; Jiang, C.; Xu, R.; Chen, B.; Jin, Z.; Lu, Y.; Zhang, J.; Yong, L.; Luo, J.; Luo, S. Worldgenbench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation. *arXiv preprint arXiv:2505.01490* **2025**.
306. Zhao, R.; Jin, S.; Wu, S.; Liao, K.; Gong, Z.; Guo, Z.; Xiao, Y.; Li, W. Knowledge Visualization: A Benchmark and Method for Knowledge-Intensive Text-to-Image Generation. *arXiv preprint arXiv:2604.22302* **2026**.
307. Ma, Z.; Shi, Z.; An, Y.; Li, P.; Wei, J.; Li, R.; Xiao, J.; Li, J.; Zhou, B. SciIR: A Large-scale Training Dataset and Benchmark for Scientific Image Reasoning Generation. *arXiv preprint arXiv:2606.30124* **2026**.
308. Lim, Y.; Ham, C.; Chen, P.Y.; Ghadiyaram, D. FAGER: Factually Grounded Evaluation and Refinement of Text-to-Image Models. *arXiv preprint arXiv:2605.19111* **2026**.
309. Luo, H.; Huang, Z.; Chung, S.; Wang, Y.; Jin, Y.; Li, J.; Li, J.; Li, X.; Salam, H. AtelierEval: Agentic evaluation of humans & LLMs as text-to-image prompters. *arXiv preprint arXiv:2605.22645* **2026**.