Supplemental Material

1 Literature Collection and Scope

We collected representative literature on diffusion-based visual generation from 2020 to 2026, with the main search completed in May 2026. The search covered major venues and indexing sources, including IEEE Xplore, ACM Digital Library, CVF Open Access, OpenReview, arXiv, and Google Scholar. The scope included text-to-image generation, controllable generation, image editing, personalization, video generation, multi-view and 3D generation, preference alignment, safety alignment, and physically or causally grounded generation.

The search terms covered both diffusion models and consistency-related failure modes. Representative keywords included "diffusion model consistency", "text-to-image prompt alignment", "compositional text-to-image generation", "controllable diffusion generation", "image editing preservation", "personalized diffusion identity preservation", "multi-view diffusion consistency", "video diffusion temporal consistency", "story generation consistency", "diffusion preference alignment", "safe diffusion", "concept erasure in diffusion models", "physical commonsense generation", and "world-consistent video generation". We also expanded the candidate set through citation links and benchmark references from representative papers.

A work was retained if it met at least one of the following criteria: 1) it explicitly identifies or addresses a consistency failure in diffusion-based visual generation; 2) it proposes a mechanism for enforcing agreement with prompts, controls, references, identities, views, temporal states, safety constraints, human preferences, or world-level plausibility; 3) it introduces a benchmark, metric, dataset, or evaluator for diagnosing such agreement; or 4) it provides a closely related survey or conceptual framework needed to position the field. We excluded works focused only on generic image quality, likelihood modeling, acceleration, or non-visual generation when they did not directly affect the consistency relations studied in the survey.

The final bibliography contains 200 references. We do not claim exhaustive coverage of every diffusion-generation paper. Instead, the bibliography is intended to cover representative methods and resources that define mechanism families, evaluation practices, or recurring trade-offs. When a work spans multiple consistency types, we assign it according to its dominant agreement requirement and record secondary relations when they affect mechanism design, evaluation, or trade-off analysis.

2 Benchmark Coverage Rating Protocol

Table 5 in the main paper summarizes the diagnostic coverage of representative benchmarks, datasets, and evaluators. The ratings are qualitative rule-based labels. They indicate diagnostic relevance to a consistency claim, not overall dataset quality, benchmark popularity, dataset scale, or model performance.

For each resource and each consistency type, we assign the rating using three checks: 1) whether the resource's intended diagnostic target matches the consistency claim; 2) whether its observation unit matches the claim, such as a single image, edited pair, identity-conditioned set, multi-view bundle, or video/story sequence; and 3) whether it provides dedicated tasks, annotations, metrics, learned evaluators, or human judgments that directly support the claim.

A high rating (H) means that all three checks are satisfied: the resource is designed for the corresponding consistency type, uses a suitable observation unit, and provides direct diagnostic evidence. A medium rating (M) means that the resource provides useful but incomplete evidence, or that its annotations can be adapted to the consistency claim although the resource was not designed primarily for that purpose. A low rating (L) means that the resource provides only indirect evidence, or that success on the resource does not reliably imply the corresponding form of consistency.

The following cases illustrate how the rule is applied. GenEval and GenEval2 receive high coverage for prompt and compositional faithfulness because they directly target object presence, attribute binding, counting, and spatial relations, but low coverage for temporal or world-level consistency because they evaluate single generated images. MVG-Bench receives high coverage for multi-view consistency because its tasks are designed to expose cross-view incompatibility. ImageReward and HPS-style evaluators receive high coverage for preference-related normative consistency because they are trained or calibrated against human preference signals, but only medium coverage for prompt faithfulness because a preferred image may still violate a compositional prompt. Six-CD receives high coverage for safety-related normative consistency because it jointly evaluates unsafe concept suppression and benign capability retention. Tracking or segmentation datasets such as TAO, MOSE, and VSPW receive medium rather than high coverage for temporal or narrative consistency because they provide useful state-tracking annotations but are not generative consistency benchmarks by themselves.

3 Additional Related Literature Omitted from the Main Text

The main paper intentionally cites a compact subset of the literature rather than attempting to cite every relevant work inline. This choice is not meant to imply that the remaining papers are less important. It reflects a difference between two goals. The main text is designed to present the conceptual taxonomy, the representative technical mechanisms, and the evaluation logic of consistency in diffusion-based visual generation. Therefore, its citation set is used as an extitargument scaffold: each cited work is selected to support a specific claim, define a recurring mechanism, introduce a benchmark family, or anchor a major subarea. By contrast, this supplementary section is designed as a extitcoverage scaffold: it records additional relevant works that broaden the bibliography without overloading the narrative of the main paper.

The main text follows a deliberately selective citation policy for three reasons. First, the survey is organized by consistency relations rather than by a chronological or task-by-task catalog. Many recent papers differ mainly in model scale, engineering choices, input modality, or application setting, while sharing the same underlying consistency mechanism. Citing all such variants in the main text would create long citation clusters that obscure the central distinction among external, internal, and normative consistency. Second, the same paper may touch several relations simultaneously. For example, a video editing method may involve external instruction following, internal temporal preservation, and sometimes normative preference or safety filters. The main text cites such papers only when they are necessary for the local argument, instead of repeatedly citing them in every place where a secondary relation could be mentioned. Third, the main text must remain readable under page and reference-budget constraints. We therefore prioritize papers that are foundational, introduce a representative mechanism, define a benchmark or evaluation protocol, or expose a clear trade-off that is repeatedly used in the survey discussion.

The works collected here are included in the supplement for complementary reasons. Some are strong or recent papers whose role is parallel to a method already cited in the main text; including them inline would improve breadth but not change the conceptual argument. Some are specialized variants that instantiate the same optimization locus, such as additional control adapters, editing pipelines, personalization modules, multi-view systems, video generators, reward-optimization methods, safety-erasure methods, or physical-world benchmarks. Some are general-purpose backbones, foundation systems, or technical reports that provide important context for the field but do not themselves define a relation-specific consistency objective. These works are therefore not omitted because of low relevance or low quality; they are placed here because their most useful role is bibliographic completeness, comparison, and reader navigation rather than main-text argument construction.

The classification below follows a one-paper-one-primary-role rule. Each work is assigned to the category corresponding to its dominant agreement target or its dominant role in the survey logic. Cross-cutting surveys, diffusion foundations, large image/video/3D/world generation backbones, and general foundation systems are placed in the cross-cutting table, because they provide infrastructure or background rather than enforcing one specific consistency relation. External-consistency entries focus on agreement with conditions supplied outside the generated sample, such as prompts, layouts, boxes, masks, poses, reference images, editing instructions, and rendered text. Internal-consistency entries focus on compatibility among generated states, such as subject identity, style, geometry, viewpoint, temporal continuity, character persistence, story state, and multi-view or video coherence. Normative-consistency entries focus on evaluative constraints, including human preference, aesthetics, safety, concept removal, benign retention, physical plausibility, commonsense, causal consequences, and world-model validity.

This organization also avoids duplicate counting. If two BibTeX entries have the same normalized title, only one representative entry is listed in the table. When a paper could reasonably appear in multiple places, we choose the location that best matches the role for which it is most likely to be consulted by readers. For instance, a large text-to-video backbone is placed under cross-cutting video foundation systems unless the paper's central contribution is an explicit temporal-consistency mechanism or benchmark. Similarly, a physical video-generation benchmark is placed under normative consistency even if it evaluates temporal sequences, because its primary agreement target is not merely frame-to-frame smoothness but physical or causal validity. The result is a supplement that complements the main text: the main article remains focused and argumentative, while this section provides a broader, taxonomy-aligned bibliography for readers who want to trace related work in greater depth.

TABLE 1: Additional cross-cutting surveys, foundation systems, backbones, and background resources not cited in the main text.

Subcategory	Additional work	Role
Cross-cutting surveys and background	[1] *Evolution of Video Generative Foundations*. CoRR/arXiv 2026	Survey
Cross-cutting surveys and background	[2] *A Comprehensive Survey on Concept Erasure in Text-to-Image Diffusion Models*. CoRR/arXiv 2025	Survey
Cross-cutting surveys and background	[3] *Controllable Video Generation: A Survey*. CoRR/arXiv 2025	Survey
Cross-cutting surveys and background	[4] *Simulating the Real World: A Unified Survey of Multimodal Generative Models*. CoRR/arXiv 2025	Survey

Continued on the next page

Subcategory	Additional work	Role
Cross-cutting surveys and background	[5] *A Survey on Vision-Language-Action Models for Embodied AI*. CoRR/arXiv 2024	Survey
Cross-cutting surveys and background	[6] *Survey: Transformer based video-language pre-training*. AI Open 2022	Survey
Video foundation systems	[7] *CogVideoX-Fun: Text-to-Video Generation with Flexible Resolution and Duration*. 2025	Foundation system
Video foundation systems	[8] *CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer*. ICLR 2025	Foundation system
Video foundation systems	[9] *Kling-Omni Technical Report*. CoRR/arXiv 2025	Foundation system
Video foundation systems	[10] *SkyReels-A2: Compose Anything in Video Diffusion Transformers*. CoRR/arXiv 2025	Foundation system
Video foundation systems	[11] *SkyReels-Audio: Omni Audio-Conditioned Talking Portraits in Video Diffusion Transformers*. CoRR/arXiv 2025	Foundation system
Video foundation systems	[12] *SkyReels-V2: Infinite-length Film Generative Model*. CoRR/arXiv 2025	Foundation system
Video foundation systems	[13] *Step-Video-T2V Technical Report: The Practice, Challenges, and Future of Video Foundation Model*. CoRR/arXiv 2025	Foundation system
Video foundation systems	[14] *Vchitect-2.0: Parallel Transformer for Scaling Up Video Diffusion Models*. CoRR/arXiv 2025	Foundation system
Video foundation systems	[15] *Allegro: Open the Black Box of Commercial-Level Video Generation Model*. CoRR/arXiv 2024	Foundation system
Video foundation systems	[16] *HunyuanVideo: A Systematic Framework For Large Video Generative Models*. CoRR/arXiv 2024	Foundation system
Video foundation systems	[17] *Lumiere: A Space-Time Diffusion Model for Video Generation*. SIGGRAPH Asia 2024	Foundation system
Video foundation systems	[18] *Movie Gen: A Cast of Media Foundation Models*. CoRR/arXiv 2024	Foundation system
Video foundation systems	[19] *Open-Sora Plan: Open-source Large Video Generation Model*. CoRR/arXiv 2024	Foundation system
Video foundation systems	[20] *Video Generation Models as World Simulators*. 2024	Foundation system
Video foundation systems	[21] *VideoPoet: A Large Language Model for Zero-Shot Video Generation*. ICML 2024	Foundation system
Video foundation systems	[22] *Vidu: a Highly Consistent, Dynamic and Skilled Text-to-Video Generator with Diffusion Models*. CoRR/arXiv 2024	Foundation system
Video foundation systems	[23] *I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models*. CoRR/arXiv 2023	Foundation system
Video foundation systems	[24] *Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions*. ICLR 2023	Foundation system
Video foundation systems	[25] *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets*. CoRR/arXiv 2023	Foundation system
Video foundation systems	[26] *Imagen Video: High Definition Video Generation with Diffusion Models*. CoRR/arXiv 2022	Foundation system
3D and 4D foundation systems	[27] *Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details*. CoRR/arXiv 2025	Foundation system
3D and 4D foundation systems	[28] *Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation*. ACM Trans. Graph. 2025	Foundation system
3D and 4D foundation systems	[29] *DreamCraft3D: Hierarchical 3D Generation with Bootstrapped Diffusion Prior*. ICLR 2024	Foundation system
3D and 4D foundation systems	[30] *GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation*. ECCV 2024	Foundation system
3D and 4D foundation systems	[31] *Instant3D: Fast Text-to-3D with Sparse-view Generation and Large Reconstruction Model*. ICLR 2024	Foundation system
3D and 4D foundation systems	[32] *LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation*. ECCV 2024	Foundation system
3D and 4D foundation systems	[33] *MeshAnything: Artist-Created Mesh Generation with Autoregressive Transformers*. CoRR/arXiv 2024	Foundation system
3D and 4D foundation systems	[34] *OpenLRM: Open-Source Large Reconstruction Models*. 2024	Foundation system
3D and 4D foundation systems	[35] *TripoSR: Fast 3D Object Reconstruction from a Single Image*. CoRR/arXiv 2024	Foundation system
3D and 4D foundation systems	[36] *DreamFusion: Text-to-3D using 2D Diffusion*. ICLR 2023	Foundation system
3D and 4D foundation systems	[37] *Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation*. ICCV 2023	Foundation system
3D and 4D foundation systems	[38] *GaussianDreamer: Fast Generation from Text to 3D Gaussian Splatting with Point Cloud Priors*. CoRR/arXiv 2023	Foundation system
3D and 4D foundation systems	[39] *Magic3D: High-Resolution Text-to-3D Content Creation*. CVPR 2023	Foundation system
3D and 4D foundation systems	[40] *One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization*. NeurIPS 2023	Foundation system
3D and 4D foundation systems	[41] *ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation*. NeurIPS 2023	Foundation system
3D and 4D foundation systems	[42] *ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes*. CVPR 2017	Foundation system
3D and 4D foundation systems	[43] *ShapeNet: An Information-Rich 3D Model Repository*. CoRR/arXiv 2015	Foundation system
World and embodied foundation systems	[44] *Cosmos World Foundation Model Platform for Physical AI*. CoRR/arXiv 2025	Foundation system

Continued on the next page

Subcategory	Additional work	Role
World and embodied foundation systems	[45] *DriveDreamer4D: World Models Are Effective Data Machines for 4D Driving Scene Representation*. CVPR 2025	Foundation system
World and embodied foundation systems	[46] *GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving*. CoRR/arXiv 2025	Foundation system
World and embodied foundation systems	[47] *VideoVerse: Does Your T2V Generator Have World Model Capability to Synthesize Videos?*. CoRR/arXiv 2025	Foundation system
World and embodied foundation systems	[48] *DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation*. CoRR/arXiv 2024	Foundation system
World and embodied foundation systems	[49] *DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving*. CoRR/arXiv 2023	Foundation system
World and embodied foundation systems	[50] *GAIA-1: A Generative World Model for Autonomous Driving*. CoRR/arXiv 2023	Foundation system
Low-level restoration and enhancement resources	[51] *DeblurDiff: Real-World Image Deblurring with Generative Diffusion Models*. CoRR/arXiv 2025	Related method
Low-level restoration and enhancement resources	[52] *Exploiting Diffusion Prior for Real-World Image Dehazing with Unpaired Training*. AAAI 2025	Related method
Low-level restoration and enhancement resources	[53] *Learning Hazing to Dehazing: Towards Realistic Haze Generation for Real-World Image Dehazing*. CVPR 2025	Related method
Low-level restoration and enhancement resources	[54] *Proxies for Distortion and Consistency with Applications for Real-World Image Restoration*. CoRR/arXiv 2025	Related method
Low-level restoration and enhancement resources	[55] *TSD-SR: One-Step Diffusion with Target Score Distillation for Real-World Image Super-Resolution*. CVPR 2025	Related method
Low-level restoration and enhancement resources	[56] *DreamClear: High-Capacity Real-World Image Restoration with Privacy-Safe Dataset Curation*. NeurIPS 2024	Related method
Low-level restoration and enhancement resources	[57] *Zero-Reference Low-Light Enhancement via Physical Quadruple Priors*. CVPR 2024	Related method

TABLE 2: Additional external-consistency works not cited in the main text.

Subcategory	Additional work	Role
Prompt and compositional faithfulness	[58] *DesignDiffusion: High-Quality Text-to-Design Image Generation with Diffusion Models*. CVPR 2025	Method
Prompt and compositional faithfulness	[59] *Detection-Driven Object Count Optimization for Text-to-Image Diffusion Models*. arXiv 2025	Method
Prompt and compositional faithfulness	[60] *RepText: Rendering Visual Text via Replicating*. CoRR/arXiv 2025	Method
Prompt and compositional faithfulness	[61] *Direct-a-Video: Customized Video Generation with User-Directed Camera Movement and Object Motion*. SIGGRAPH 2024	Method
Prompt and compositional faithfulness	[62] *Glyph-ByT5: A Customized Text Encoder for Accurate Visual Text Rendering*. ECCV 2024	Method
Prompt and compositional faithfulness	[63] *TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering*. ECCV 2024	Method
Prompt and compositional faithfulness	[64] *U-DiffText: A Unified Framework for High-quality Text Synthesis in Arbitrary Images via Character-aware Diffusion Models*. ECCV 2024	Method
Prompt and compositional faithfulness	[65] *GlyphDraw: Learning to Draw Chinese Characters in Image Synthesis Models Coherently*. CoRR/arXiv 2023	Method
Prompt and compositional faithfulness	[66] *Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment*. NeurIPS 2023	Method
Grounded, structural, and multi-condition control	[67] *AnyI2V: Animating Any Conditional Image with Motion Control*. CoRR/arXiv 2025	Method
Grounded, structural, and multi-condition control	[68] *DivControl: Knowledge Diversion for Controllable Image Generation*. CoRR/arXiv 2025	Method
Grounded, structural, and multi-condition control	[69] *Grounding Text-to-Image Diffusion Models for Controlled High-Quality Image Generation*. CoRR/arXiv 2025	Method
Grounded, structural, and multi-condition control	[70] *Motion Prompting: Controlling Video Generation with Motion Trajectories*. CVPR 2025	Method
Grounded, structural, and multi-condition control	[71] *PosterO: Structuring Layout Trees to Enable Language Models in Generalized Content-Aware Layout Generation*. CVPR 2025	Method
Grounded, structural, and multi-condition control	[72] *UniCon: Unidirectional Information Flow for Effective Control of Large-Scale Diffusion Models*. ICLR 2025	Method
Grounded, structural, and multi-condition control	[73] *Adversarial Supervision Makes Layout-to-Image Diffusion Models Thrive*. ICLR 2024	Method
Grounded, structural, and multi-condition control	[74] *CameraCtrl: Enabling Camera Control for Text-to-Video Generation*. CoRR/arXiv 2024	Method
Grounded, structural, and multi-condition control	[75] *ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback*. ECCV 2024	Method
Grounded, structural, and multi-condition control	[76] *ControlNet-XS: Rethinking the Control of Text-to-Image Diffusion Models as Feedback-Control Systems*. ECCV 2024	Method

Continued on the next page

Subcategory	Additional work	Role
Grounded, structural, and multi-condition control	[77] *ControlNeXt: Powerful and Efficient Control for Image and Video Generation*. CoRR/arXiv 2024	Method
Grounded, structural, and multi-condition control	[78] *DogLayout: Denoising Diffusion GAN for Discrete and Continuous Layout Generation*. CoRR/arXiv 2024	Method
Grounded, structural, and multi-condition control	[79] *InstanceDiffusion: Instance-level Control for Image Generation*. CVPR 2024	Method
Grounded, structural, and multi-condition control	[80] *MotionCtrl: A Unified and Flexible Motion Controller for Video Generation*. SIGGRAPH 2024	Method
Grounded, structural, and multi-condition control	[81] *SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models*. ECCV 2024	Method
Grounded, structural, and multi-condition control	[82] *TrailBlazer: Trajectory Control for Diffusion-Based Video Generation*. SIGGRAPH Asia 2024	Method
Grounded, structural, and multi-condition control	[83] *Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models*. CoRR/arXiv 2023	Method
Grounded, structural, and multi-condition control	[84] *DenseDiffusion: Dense Text-to-Image Generation with Attention Modulation*. ICCV 2023	Method
Grounded, structural, and multi-condition control	[85] *DLT: Conditioned layout generation with Joint Discrete-Continuous Diffusion Layout Transformer*. ICCV 2023	Method
Grounded, structural, and multi-condition control	[86] *Freestyle Layout-to-Image Synthesis*. CVPR 2023	Method
Grounded, structural, and multi-condition control	[87] *LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation*. CVPR 2023	Method
Grounded, structural, and multi-condition control	[88] *LayoutDM: Transformer-based Diffusion Model for Layout Generation*. CVPR 2023	Method
Grounded, structural, and multi-condition control	[89] *LayoutGPT: Compositional Visual Planning and Generation with Large Language Models*. NeurIPS 2023	Method
Grounded, structural, and multi-condition control	[90] *LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models*. CoRR/arXiv 2023	Method
Grounded, structural, and multi-condition control	[91] *SceneComposer: Any-Level Semantic Image Synthesis*. CVPR 2023	Method
Grounded, structural, and multi-condition control	[92] *Sketch-Guided Text-to-Image Diffusion Models*. SIGGRAPH 2023	Method
Grounded, structural, and multi-condition control	[93] *VideoComposer: Compositional Video Synthesis with Motion Controllability*. NeurIPS 2023	Method
Prompt, grounding, and structural control	[94] *Guiding Text-to-Image Diffusion Model Towards Grounded Generation*. CoRR/arXiv 2023	Method
Prompt, grounding, and structural control	[95] *MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation*. ICML 2023	Method
Editing, inpainting, and local manipulation	[96] *Exploring Multimodal Diffusion Transformers for Enhanced Prompt-based Image Editing*. CoRR/arXiv 2025	Method
Editing, inpainting, and local manipulation	[97] *FlowEdit: Inversion-Free Text-Based Editing Using Pre-Trained Flow Models*. ICCV 2025	Method
Editing, inpainting, and local manipulation	[98] *GoodDrag: Towards Good Practices for Drag Editing with Diffusion Models*. ICLR 2025	Method
Editing, inpainting, and local manipulation	[99] *ILLUME+: Illuminating Unified MLLM with Dual Visual Tokenization and Diffusion Refinement*. CoRR/arXiv 2025	Method
Editing, inpainting, and local manipulation	[100] *Inpaint4Drag: Repurposing Inpainting Models for Drag-Based Image Editing via Bidirectional Warping*. CoRR/arXiv 2025	Method
Editing, inpainting, and local manipulation	[101] *MagicQuill: An Intelligent Interactive Image Editing System*. CVPR 2025	Method
Editing, inpainting, and local manipulation	[102] *OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision*. ICLR 2025	Method
Editing, inpainting, and local manipulation	[103] *One-Step Image Translation with Text-to-Image Models*. AAAI 2025	Method
Editing, inpainting, and local manipulation	[104] *PoseTraj: Pose-Aware Trajectory Control in Video Diffusion*. CVPR 2025	Method
Editing, inpainting, and local manipulation	[105] *ReFlex: Text-Guided Editing of Real Images in Rectified Flow via Mid-Step Feature Extraction and Attention Adaptation*. CoRR/arXiv 2025	Method
Editing, inpainting, and local manipulation	[106] *Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations*. ICLR 2025	Method
Editing, inpainting, and local manipulation	[107] *Training-Free Text-Guided Image Editing with Visual Autoregressive Model*. ICCV 2025	Method
Editing, inpainting, and local manipulation	[108] *VideoAnydoor: High-fidelity Video Object Insertion with Precise Motion Control*. SIGGRAPH 2025	Method
Editing, inpainting, and local manipulation	[109] *A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting*. ECCV 2024	Method
Editing, inpainting, and local manipulation	[110] *AnyText2: Visual Text Generation and Editing With Customizable Attributes*. CoRR/arXiv 2024	Method
Editing, inpainting, and local manipulation	[111] *AnyText: Multilingual Visual Text Generation and Editing*. ICLR 2024	Method
Editing, inpainting, and local manipulation	[112] *BrushEdit: All-In-One Image Inpainting and Editing*. CoRR/arXiv 2024	Method

Continued on the next page

Subcategory	Additional work	Role
Editing, inpainting, and local manipulation	[113] *BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion*. ECCV 2024	Method
Editing, inpainting, and local manipulation	[114] *Contrastive Denoising Score for Text-Guided Latent Diffusion Image Editing*. CVPR 2024	Method
Editing, inpainting, and local manipulation	[115] *Direct Inversion: Boosting Diffusion-based Editing with 3 Lines of Code*. ICLR 2024	Method
Editing, inpainting, and local manipulation	[116] *DragonDiffusion: Enabling Drag-style Manipulation on Diffusion Models*. ICLR 2024	Method
Editing, inpainting, and local manipulation	[117] *FreeDrag: Feature Dragging for Reliable Point-Based Image Editing*. CVPR 2024	Method
Editing, inpainting, and local manipulation	[118] *Guiding Instruction-based Image Editing via Multimodal Large Language Models*. ICLR 2024	Method
Editing, inpainting, and local manipulation	[119] *InfEdit: Inversion-Free Image Editing with Natural Language*. CVPR 2024	Method
Editing, inpainting, and local manipulation	[120] *Invertible Consistency Distillation for Text-Guided Image Editing in Around 7 Steps*. NeurIPS 2024	Method
Editing, inpainting, and local manipulation	[121] *LEDITS++: Limitless Image Editing using Text-to-Image Models*. CVPR 2024	Method
Editing, inpainting, and local manipulation	[122] *Readout Guidance: Learning Control from Diffusion Features*. CVPR 2024	Method
Editing, inpainting, and local manipulation	[123] *Repositioning the Subject within Image*. Trans. Mach. Learn. Res. 2024	Method
Editing, inpainting, and local manipulation	[124] *Text-Driven Image Editing via Learnable Regions*. CVPR 2024	Method
Editing, inpainting, and local manipulation	[125] *UltraEdit: Instruction-based Fine-Grained Image Editing at Scale*. NeurIPS 2024	Method
Editing, inpainting, and local manipulation	[126] *VideoTetris: Towards Compositional Text-to-Video Generation with Multi-Concept Control*. NeurIPS 2024	Method
Editing, inpainting, and local manipulation	[127] *Zero-shot Image Editing with Reference Imitation*. CoRR/arXiv 2024	Method
Editing, inpainting, and local manipulation	[128] *Collaborative Diffusion for Multi-Modal Face Generation and Editing*. CVPR 2023	Method
Editing, inpainting, and local manipulation	[129] *Collaborative Score Distillation for Consistent Visual Synthesis*. CoRR/arXiv 2023	Method
Editing, inpainting, and local manipulation	[130] *Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold*. SIGGRAPH 2023	Method
Editing, inpainting, and local manipulation	[131] *DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory*. CoRR/arXiv 2023	Method
Editing, inpainting, and local manipulation	[132] *Energy-Based Cross Attention for Bayesian Context Update in Text-to-Image Diffusion Models*. NeurIPS 2023	Method
Editing, inpainting, and local manipulation	[133] *Inpaint Anything: Segment Anything Meets Image Inpainting*. CoRR/arXiv 2023	Method
Editing, inpainting, and local manipulation	[134] *Localizing Object-level Shape Variations with Text-to-Image Diffusion Models*. ICCV 2023	Method
Editing, inpainting, and local manipulation	[135] *MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing*. ICCV 2023	Method
Editing, inpainting, and local manipulation	[136] *Visual Instruction Inversion: Image Editing via Image Prompting*. NeurIPS 2023	Method
Editing, inpainting, and local manipulation	[137] *Zero-shot Image-to-Image Translation*. SIGGRAPH 2023	Method
Editing, inpainting, and local manipulation	[138] *Blended Diffusion for Text-driven Editing of Natural Images*. CVPR 2022	Method
Editing, inpainting, and local manipulation	[139] *DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation*. CVPR 2022	Method
Editing, inpainting, and local manipulation	[140] *RePaint: Inpainting using Denoising Diffusion Probabilistic Models*. CVPR 2022	Method
Editing, inpainting, and local manipulation	[141] *Unifying Diffusion Models' Latent Space, with Applications to CycleDiffusion and Guidance*. CoRR/arXiv 2022	Method
External-consistency benchmarks and evaluators	[142] *CompAlign: Improving Compositional Text-to-Image Generation with a Complex Benchmark and Fine-Grained Feedback*. CoRR/arXiv 2025	Benchmark / evaluation
External-consistency benchmarks and evaluators	[143] *STRICT: Stress Test of Rendering Images Containing Text*. EMNLP 2025	Benchmark / evaluation
External-consistency benchmarks and evaluators	[144] *TIT-Score: Evaluating Long-Prompt Based Text-to-Image Alignment via Text-to-Image-to-Text Consistency*. CoRR/arXiv 2025	Benchmark / evaluation
External-consistency benchmarks and evaluators	[145] *GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation*. CVPR Workshop on Synthetic Data for Computer Vision 2024	Benchmark / evaluation
External-consistency benchmarks and evaluators	[146] *VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation*. ACL 2024	Benchmark / evaluation
External-consistency benchmarks and evaluators	[147] *OBJECT 3DIT: Language-guided 3D-aware Image Editing*. NeurIPS 2023	Benchmark / evaluation

TABLE 3: Additional internal-consistency works not cited in the main text.

Subcategory	Additional work	Role
Personalization, identity, and character consistency	[148] *Personalize Anything for Free with Diffusion Transformer*. AAAI 2026	Method
Personalization, identity, and character consistency	[149] *RealCustom++: Representing Images as Real Textual Word for Real-Time Customization*. IEEE Trans. Pattern Anal. Mach. Intell. 2026	Method
Personalization, identity, and character consistency	[150] *CharaConsist: Fine-Grained Consistent Character Generation*. CoRR/arXiv 2025	Method
Personalization, identity, and character consistency	[151] *CoCoIns: Consistent Subject Generation via Contrastive Instantiated Concepts*. Trans. Mach. Learn. Res. 2025	Method
Personalization, identity, and character consistency	[152] *ContextAnyone: Context-Aware Diffusion for Character-Consistent Text-to-Video Generation*. CoRR/arXiv 2025	Method
Personalization, identity, and character consistency	[153] *DiffSensei: Bridging Multi-Modal LLMs and Diffusion Models for Customized Manga Generation*. CVPR 2025	Method
Personalization, identity, and character consistency	[154] *Generating Multi-Image Synthetic Data for Text-to-Image Customization*. CoRR/arXiv 2025	Method
Personalization, identity, and character consistency	[155] *Identity-Preserving Text-to-Video Generation by Frequency Decomposition*. CVPR 2025	Method
Personalization, identity, and character consistency	[156] *IMAGDressing-v1: Customizable Virtual Dressing*. AAAI 2025	Method
Personalization, identity, and character consistency	[157] *InstantCharacter: Personalize Any Characters with a Scalable Diffusion Transformer Framework*. CoRR/arXiv 2025	Method
Personalization, identity, and character consistency	[158] *StableAnimator: High-Quality Identity-Preserving Human Image Animation*. CVPR 2025	Method
Personalization, identity, and character consistency	[159] *StyleBlend: Enhancing Style-Specific Content Creation in Text-to-Image Diffusion Models*. Comput. Graph. Forum 2025	Method
Personalization, identity, and character consistency	[160] *Tora2: Motion and Appearance Customized Diffusion Transformer for Multi-Entity Video Generation*. ACM MM 2025	Method
Personalization, identity, and character consistency	[161] *Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation*. CVPR 2024	Method
Personalization, identity, and character consistency	[162] *AnimateLCM: Accelerating the Animation of Personalized Diffusion Models and Adapters with Decoupled Consistency Learning*. SIGGRAPH Asia 2024	Method
Personalization, identity, and character consistency	[163] *AutoStudio: Crafting Consistent Subjects in Multi-turn Interactive Image Generation*. CoRR/arXiv 2024	Method
Personalization, identity, and character consistency	[164] *CustomVideo: Customizing Text-to-Video Generation with Multiple Subjects*. CoRR/arXiv 2024	Method
Personalization, identity, and character consistency	[165] *INSTASTYLE: Inversion Noise of a Stylized Image is Secretly a Style Adviser*. ECCV 2024	Method
Personalization, identity, and character consistency	[166] *Magic-Me: Identity-Specific Video Customized Diffusion*. CoRR/arXiv 2024	Method
Personalization, identity, and character consistency	[167] *MagicTailor: Component-Controllable Personalization in Text-to-Image Diffusion Models*. CoRR/arXiv 2024	Method
Personalization, identity, and character consistency	[168] *Material Palette: Extraction of Materials from a Single Image*. CVPR 2024	Method
Personalization, identity, and character consistency	[169] *OneActor: Consistent Character Generation via Cluster-Conditioned Guidance*. CoRR/arXiv 2024	Method
Personalization, identity, and character consistency	[170] *PuLID: Pure and Lightning ID Customization via Contrastive Alignment*. NeurIPS 2024	Method
Personalization, identity, and character consistency	[171] *RealCustom: Narrowing Real Text Word for Real-Time Open-Domain Text-to-Image Customization*. CVPR 2024	Method
Personalization, identity, and character consistency	[172] *Style Aligned Image Generation via Shared Attention*. CVPR 2024	Method
Personalization, identity, and character consistency	[173] *Subject-Diffusion: Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning*. SIGGRAPH 2024	Method
Personalization, identity, and character consistency	[174] *The Hidden Language of Diffusion Models*. ICLR 2024	Method
Personalization, identity, and character consistency	[175] *Total Selfie: Generating Full-Body Selfies*. CVPR 2024	Method
Personalization, identity, and character consistency	[176] *ZeST: Zero-Shot Material Transfer from a Single Image*. ECCV 2024	Method
Personalization, identity, and character consistency	[177] *DreamDistribution: Prompt Distribution Learning for Text-to-Image Diffusion Models*. CoRR/arXiv 2023	Method
Personalization, identity, and character consistency	[178] *ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation*. ICCV 2023	Method
Personalization, identity, and character consistency	[179] *ITI-Gen: Inclusive Text-to-Image Generation*. ICCV 2023	Method
Personalization, identity, and character consistency	[180] *Unsupervised Compositional Concepts Discovery with Text-to-Image Generative Models*. ICCV 2023	Method
Story and long-range narrative consistency	[181] *ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions*. CVPR 2025	Method
Story and long-range narrative consistency	[182] *Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models*. CVPR 2024	Method

Continued on the next page

Subcategory	Additional work	Role
Multi-view, 3D, and 4D consistency	[183] *3D-Consistent Multi-View Editing by Diffusion Guidance*. CoRR/arXiv 2025	Method
Multi-view, 3D, and 4D consistency	[184] *Efficient4D: Fast Dynamic 3D Object Generation from a Single-view Video*. International Journal of Computer Vision 2025	Method
Multi-view, 3D, and 4D consistency	[185] *MVRoom: Controllable 3D Indoor Scene Generation with Multi-View Diffusion Models*. CoRR/arXiv 2025	Method
Multi-view, 3D, and 4D consistency	[186] *SV4D 2.0: Enhancing Spatio-Temporal Consistency in Multi-View Video Diffusion for High-Quality 4D Generation*. CoRR/arXiv 2025	Method
Multi-view, 3D, and 4D consistency	[187] *SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency*. ICLR 2025	Method
Multi-view, 3D, and 4D consistency	[188] *ViCoDR: View-Consistent Diffusion Representations for 3D-Consistent Video Generation*. CoRR/arXiv 2025	Method
Multi-view, 3D, and 4D consistency	[189] *CAT3D: Create Anything in 3D with Multi-View Diffusion Models*. NeurIPS 2024	Method
Multi-view, 3D, and 4D consistency	[190] *Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance*. ECCV 2024	Method
Multi-view, 3D, and 4D consistency	[191] *Consistent4D: Consistent 360Dynamic Object Generation from Monocular Video*. The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024 2024	Method
Multi-view, 3D, and 4D consistency	[192] *Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention*. NeurIPS 2024	Method
Multi-view, 3D, and 4D consistency	[193] *FlexGen: Flexible Multi-View Generation from Text and Image Inputs*. CoRR/arXiv 2024	Method
Multi-view, 3D, and 4D consistency	[194] *Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors*. ICLR 2024	Method
Multi-view, 3D, and 4D consistency	[195] *MotionDreamer: Zero-Shot 3D Mesh Animation from Video Diffusion Models*. CoRR/arXiv 2024	Method
Multi-view, 3D, and 4D consistency	[196] *SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion*. ECCV 2024	Method
Multi-view, 3D, and 4D consistency	[197] *Unique3D: High-Quality and Efficient 3D Mesh Generation from a Single Image*. NeurIPS 2024	Method
Multi-view, 3D, and 4D consistency	[198] *DreamWaltz: Make a Scene with Complex 3D Animatable Avatars*. NeurIPS 2023	Method
Multi-view, 3D, and 4D consistency	[199] *STEm-Seg: Spatio-Temporal Embeddings for Instance Segmentation in Videos*. ECCV 2020	Method
Video, motion, and temporal consistency	[200] *A^{2}RD: Agentic Autoregressive Diffusion for Long Video Consistency*. CoRR/arXiv 2026	Method
Video, motion, and temporal consistency	[201] *Video is Worth a Thousand Images: Exploring the Latest Trends in Long Video Generation*. ACM Comput. Surv. 2026	Method
Video, motion, and temporal consistency	[202] *AnimateAnything: Consistent and Controllable Animation for Video Generation*. CVPR 2025	Method
Video, motion, and temporal consistency	[203] *ANYPORTAL: Zero-Shot Consistent Video Background Replacement*. CoRR/arXiv 2025	Method
Video, motion, and temporal consistency	[204] *High-Fidelity and Long-Duration Human Image Animation with Diffusion Transformer*. CoRR/arXiv 2025	Method
Video, motion, and temporal consistency	[205] *LTX-Video: Realtime Video Latent Diffusion*. CoRR/arXiv 2025	Method
Video, motion, and temporal consistency	[206] *MotionClone: Training-Free Motion Cloning for Controllable Video Generation*. ICLR 2025	Method
Video, motion, and temporal consistency	[207] *OmniHuman-1: Rethinking the Scaling-Up of One-Stage Conditioned Human Animation Models*. ICCV 2025	Method
Video, motion, and temporal consistency	[208] *Pyramidal Flow Matching for Efficient Video Generative Modeling*. ICLR 2025	Method
Video, motion, and temporal consistency	[209] *SG-I2V: Self-Guided Trajectory Control in Image-to-Video Generation*. ICLR 2025	Method
Video, motion, and temporal consistency	[210] *SketchVideo: Sketch-based Video Generation and Editing*. CVPR 2025	Method
Video, motion, and temporal consistency	[211] *SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints*. ICLR 2025	Method
Video, motion, and temporal consistency	[212] *Trajectory attention for fine-grained video motion control*. ICLR 2025	Method
Video, motion, and temporal consistency	[213] *TurboDiffusion: Accelerating Video Diffusion Models by 100-200 Times*. CoRR/arXiv 2025	Method
Video, motion, and temporal consistency	[214] *VACE: All-in-One Video Creation and Editing*. CoRR/arXiv 2025	Method
Video, motion, and temporal consistency	[215] *VideoGrain: Modulating Space-Time Attention for Multi-Grained Video Editing*. ICLR 2025	Method
Video, motion, and temporal consistency	[216] *VLOGGER: Multimodal Diffusion for Embodied Avatar Synthesis*. CVPR 2025	Method
Video, motion, and temporal consistency	[217] *Cinemo: Consistent and Controllable Image Animation with Motion Diffusion Models*. CoRR/arXiv 2024	Method

Continued on the next page

Subcategory	Additional work	Role
Video, motion, and temporal consistency	[218] *CoDeF: Content Deformation Fields for Temporally Consistent Video Processing*. CVPR 2024	Method
Video, motion, and temporal consistency	[219] *Enhancing Motion in Text-to-Video Generation with Decomposed Encoding and Conditioning*. NeurIPS 2024	Method
Video, motion, and temporal consistency	[220] *FreeInit: Bridging Initialization Gap in Video Diffusion Models*. ECCV 2024	Method
Video, motion, and temporal consistency	[221] *FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling*. ICLR 2024	Method
Video, motion, and temporal consistency	[222] *LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control*. CoRR/arXiv 2024	Method
Video, motion, and temporal consistency	[223] *MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model*. CVPR 2024	Method
Video, motion, and temporal consistency	[224] *MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance*. CoRR/arXiv 2024	Method
Video, motion, and temporal consistency	[225] *MOFA-Video: Controllable Image Animation via Generative Motion Field Adaptions in Frozen Image-to-Video Diffusion Model*. ECCV 2024	Method
Video, motion, and temporal consistency	[226] *MovieChat+: Question-aware Sparse Memory for Long Video Question Answering*. CoRR/arXiv 2024	Method
Video, motion, and temporal consistency	[227] *MovieChat: From Dense Token to Sparse Memory for Long Video Understanding*. CVPR 2024	Method
Video, motion, and temporal consistency	[228] *MusePose: A Pose-Driven Image-to-Video Framework for Virtual Human Generation*. CoRR/arXiv 2024	Method
Video, motion, and temporal consistency	[229] *Noise Calibration: Plug-and-Play Content-Preserving Video Enhancement Using Pre-trained Video Diffusion Models*. ECCV 2024	Method
Video, motion, and temporal consistency	[230] *SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction*. ICLR 2024	Method
Video, motion, and temporal consistency	[231] *Still-Moving: Customized Video Generation without Customized Video Data*. ACM Trans. Graph. 2024	Method
Video, motion, and temporal consistency	[232] *VideoBooth: Diffusion-based Video Generation with Image Prompts*. CVPR 2024	Method
Video, motion, and temporal consistency	[233] *Diffusion Video Autoencoders: Toward Temporally Consistent Face Video Editing via Disentangled Video Encoding*. CVPR 2023	Method
Video, motion, and temporal consistency	[234] *Fine-tuned CLIP Models are Efficient Video Learners*. CVPR 2023	Method
Video, motion, and temporal consistency	[235] *GLOBER: Coherent Non-autoregressive Video Generation via GLOBal Guided Video DecodER*. NeurIPS 2023	Method
Video, motion, and temporal consistency	[236] *Improving Video Instance Segmentation via Temporal Pyramid Routing*. IEEE Trans. Pattern Anal. Mach. Intell. 2023	Method
Video, motion, and temporal consistency	[237] *LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models*. CoRR/arXiv 2023	Method
Video, motion, and temporal consistency	[238] *Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models*. ICCV 2023	Method
Video, motion, and temporal consistency	[239] *Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation*. SIGGRAPH Asia 2023	Method
Video, motion, and temporal consistency	[240] *Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation*. CoRR/arXiv 2023	Method
Video, motion, and temporal consistency	[241] *StableVideo: Text-driven Consistency-aware Diffusion Video Editing*. ICCV 2023	Method
Video, motion, and temporal consistency	[242] *Video Action Recognition with Attentive Semantic Units*. ICCV 2023	Method
Video, motion, and temporal consistency	[243] *VideoLCM: Video Latent Consistency Model*. CoRR/arXiv 2023	Method
Video, motion, and temporal consistency	[244] *Expanding Language-Image Pretrained Models for General Video Recognition*. ECCV 2022	Method
Video, motion, and temporal consistency	[245] *Frozen CLIP Models are Efficient Video Learners*. ECCV 2022	Method
Video, motion, and temporal consistency	[246] *PolyphonicFormer: Unified Query Learning for Depth-Aware Video Panoptic Segmentation*. ECCV 2022	Method
Video, motion, and temporal consistency	[247] *Prompting Visual-Language Models for Efficient Video Understanding*. ECCV 2022	Method
Video, motion, and temporal consistency	[248] *Segment as Points for Efficient and Effective Online Multi-Object Tracking and Segmentation*. IEEE Trans. Pattern Anal. Mach. Intell. 2022	Method
Video, motion, and temporal consistency	[249] *Text2LIVE: Text-Driven Layered Image and Video Editing*. ECCV 2022	Method
Video, motion, and temporal consistency	[250] *Video K-Net: A Simple, Strong, and Unified Baseline for Video Segmentation*. CVPR 2022	Method
Video, motion, and temporal consistency	[251] *Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners*. CoRR/arXiv 2022	Method
Video, motion, and temporal consistency	[252] *ActionCLIP: A New Paradigm for Video Action Recognition*. CoRR/arXiv 2021	Method
Video, motion, and temporal consistency	[253] *CompFeat: Comprehensive Feature Aggregation for Video Instance Segmentation*. AAAI 2021	Method

Continued on the next page

Subcategory	Additional work	Role
Video, motion, and temporal consistency	[254] *Crossover Learning for Fast Online Video Instance Segmentation*. ICCV 2021	Method
Video, motion, and temporal consistency	[255] *End-to-End Video Instance Segmentation With Transformers*. CVPR 2021	Method
Video, motion, and temporal consistency	[256] *STEP: Segmenting and Tracking Every Pixel*. NeurIPS Datasets and Benchmarks 2021	Method
Video, motion, and temporal consistency	[257] *VIP-DeepLab: Learning Visual Perception With Depth-Aware Video Panoptic Segmentation*. CVPR 2021	Method
Video, motion, and temporal consistency	[258] *Classifying, Segmenting, and Tracking Object Instances in Video with Mask Propagation*. CVPR 2020	Method
Video, motion, and temporal consistency	[259] *SipMask: Spatial Information Preservation for Fast Image and Video Instance Segmentation*. ECCV 2020	Method
Video, motion, and temporal consistency	[260] *Towards Real-Time Multi-Object Tracking*. ECCV 2020	Method
Video, motion, and temporal consistency	[261] *Video Instance Segmentation Tracking With a Modified VAE Architecture*. CVPR 2020	Method
Video, motion, and temporal consistency	[262] *Video Panoptic Segmentation*. CVPR 2020	Method
Video, motion, and temporal consistency	[263] *FEELVOS: Fast End-To-End Embedding Learning for Video Object Segmentation*. CVPR 2019	Method
Video, motion, and temporal consistency	[264] *MOTS: Multi-Object Tracking and Segmentation*. CVPR 2019	Method
Video, motion, and temporal consistency	[265] *SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks*. CVPR 2019	Method
Video, motion, and temporal consistency	[266] *Tracking Without Bells and Whistles*. ICCV 2019	Method
Video, motion, and temporal consistency	[267] *Video Instance Segmentation*. ICCV 2019	Method
Video, motion, and temporal consistency	[268] *Video Object Segmentation Using Space-Time Memory Networks*. ICCV 2019	Method
Video, motion, and temporal consistency	[269] *Distractor-Aware Siamese Networks for Visual Object Tracking*. ECCV 2018	Method
Video, motion, and temporal consistency	[270] *Efficient Video Object Segmentation via Network Modulation*. CVPR 2018	Method
Video, motion, and temporal consistency	[271] *High Performance Visual Tracking With Siamese Region Proposal Network*. CVPR 2018	Method
Video, motion, and temporal consistency	[272] *Deep Feature Flow for Video Recognition*. CVPR 2017	Method
Video, motion, and temporal consistency	[273] *Learning Motion Patterns in Videos*. CVPR 2017	Method
Video, motion, and temporal consistency	[274] *Learning Video Object Segmentation from Static Images*. CVPR 2017	Method
Video, motion, and temporal consistency	[275] *Learning Video Object Segmentation with Visual Memory*. ICCV 2017	Method
Video, motion, and temporal consistency	[276] *Clockwork Convnets for Video Semantic Segmentation*. ECCV Workshops 2016	Method
Video, motion, and temporal consistency	[277] *Fully-Convolutional Siamese Networks for Object Tracking*. ECCV Workshops 2016	Method
Try-on and appearance persistence	[278] *Enhancing Person-to-Person Virtual Try-On with Multi-Garment Virtual Try-Off*. CoRR/arXiv 2025	Method
Try-on and appearance persistence	[279] *ITA-MDT: Image-Timestep-Adaptive Masked Diffusion Transformer Framework for Image-Based Virtual Try-On*. CVPR 2025	Method
Try-on and appearance persistence	[280] *VTON-HandFit: Virtual Try-on for Arbitrary Hand Pose Guided by Hand Priors Embedding*. CVPR 2025	Method
Try-on and appearance persistence	[281] *OutfitAnyone: Ultra-high Quality Virtual Try-On for Any Clothing and Any Person*. CoRR/arXiv 2024	Method
Internal-consistency benchmarks and datasets	[282] *LVBench-C: A Benchmark for Long-Horizon Video Consistency*. 2026	Benchmark / evaluation
Internal-consistency benchmarks and datasets	[283] *FiVE: A Fine-grained Video Editing Benchmark for Evaluating Emerging Diffusion and Rectified Flow Models*. ICCV 2025	Benchmark / evaluation
Internal-consistency benchmarks and datasets	[284] *LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding*. NeurIPS 2024	Benchmark / evaluation
Internal-consistency benchmarks and datasets	[285] *Large-scale Video Panoptic Segmentation in the Wild: A Benchmark*. CVPR 2022	Benchmark / evaluation
Internal-consistency benchmarks and datasets	[286] *Simple online and realtime tracking with a deep association metric*. ICIP 2017	Benchmark / evaluation
Internal-consistency benchmarks and datasets	[287] *The "Something Something" Video Database for Learning and Evaluating Visual Common Sense*. ICCV 2017	Benchmark / evaluation
Internal-consistency benchmarks and datasets	[288] *A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation*. CVPR 2016	Benchmark / evaluation

TABLE 4: Additional normative-consistency works not cited in the main text.

Subcategory	Additional work	Role
Preference, aesthetics, and alignment	[289] *BranchGRPO: Stable Reinforcement Learning for Diffusion Models via Branch-Based Group Relative Policy Optimization*. ICLR 2026	Method
Preference, aesthetics, and alignment	[290] *SwiftVideo: A Unified Framework for Few-Step Video Generation Through Trajectory-Distribution Alignment*. AAAI 2026	Method
Preference, aesthetics, and alignment	[291] *Towards Better Optimization For Listwise Preference in Diffusion Models*. ICLR 2026	Method
Preference, aesthetics, and alignment	[292] *Diff-Instruct++: Training One-step Text-to-image Generator Model to Align with Human Preferences*. Trans. Mach. Learn. Res. 2025	Method
Preference, aesthetics, and alignment	[293] *Diffusion-NPO: Negative Preference Optimization for Better Preference Aligned Generation of Diffusion Models*. ICLR 2025	Method
Preference, aesthetics, and alignment	[294] *Diffusion-SDPO: Safeguarded Direct Preference Optimization for Diffusion Models*. CoRR/arXiv 2025	Method
Preference, aesthetics, and alignment	[295] *DSPO: Direct Score Preference Optimization for Diffusion Model Alignment*. ICLR 2025	Method
Preference, aesthetics, and alignment	[296] *Fine-Tuning Diffusion Generative Models via Rich Preference Optimization*. CoRR/arXiv 2025	Method
Preference, aesthetics, and alignment	[297] *GalaxyDiT: Efficient Video Generation with Guidance Alignment and Adaptive Proxy in Diffusion Transformers*. CoRR/arXiv 2025	Method
Preference, aesthetics, and alignment	[298] *Improving Video Generation with Human Feedback*. NeurIPS 2025	Method
Preference, aesthetics, and alignment	[299] *McSc: Motion-Corrective Preference Alignment for Video Generation with Self-Critic Hierarchical Reasoning*. CoRR/arXiv 2025	Method
Preference, aesthetics, and alignment	[300] *Red-Teaming Text-to-Image Systems by Rule-based Preference Modeling*. arXiv 2025	Method
Preference, aesthetics, and alignment	[301] *Scalable Ranked Preference Optimization for Text-to-Image Generation*. ICCV 2025	Method
Preference, aesthetics, and alignment	[302] *T2V-Turbo-v2: Enhancing Video Generation Model Post-Training through Data, Reward, and Conditional Guidance Design*. ICLR 2025	Method
Preference, aesthetics, and alignment	[303] *VideoDPO: Omni-Preference Alignment for Video Diffusion Generation*. CVPR 2025	Method
Preference, aesthetics, and alignment	[304] *Super-resolving Real-world Image Illumination Enhancement: A New Dataset and A Conditional Diffusion Model*. CoRR/arXiv 2024	Method
Preference, aesthetics, and alignment	[305] *T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback*. NeurIPS 2024	Method
Preference, aesthetics, and alignment	[306] *Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model*. CVPR 2024	Method
Preference, aesthetics, and alignment	[307] *VideoTuna: A Powerful Toolkit for Video Generation with Model Fine-Tuning and Post-Training*. 2024	Method
Preference, aesthetics, and alignment	[308] *DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models*. NeurIPS 2023	Method
Safety, concept erasure, and unlearning	[309] *Compensation-free Machine Unlearning in Text-to-Image Diffusion Models by Eliminating the Mutual Information*. CoRR/arXiv 2026	Method
Safety, concept erasure, and unlearning	[310] *ACE: Concept Editing in Diffusion Models without Performance Degradation*. ACM Multimedia 2025	Method
Safety, concept erasure, and unlearning	[311] *Safe-Control: A Safety Patch for Mitigating Unsafe Content in Text-to-Image Generation Models*. CoRR/arXiv 2025	Method
Safety, concept erasure, and unlearning	[312] *Concept Sliders: LoRA Adaptors for Precise Control in Diffusion Models*. ECCV 2024	Method
Safety, concept erasure, and unlearning	[313] *Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models*. NeurIPS 2024	Method
Safety, concept erasure, and unlearning	[314] *One-dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications*. CVPR 2024	Method
Safety, concept erasure, and unlearning	[315] *Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models*. ECCV 2024	Method
Safety, concept erasure, and unlearning	[316] *Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models?*. ICLR 2024	Method
Safety, concept erasure, and unlearning	[317] *SafeGen: Mitigating Unsafe Content Generation in Text-to-Image Models*. CoRR/arXiv 2024	Method
Safety, concept erasure, and unlearning	[318] *SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation*. ICLR 2024	Method
Safety, concept erasure, and unlearning	[319] *SneakyPrompt: Jailbreaking Text-to-image Generative Models*. IEEE S&P 2024	Method
Safety, concept erasure, and unlearning	[320] *Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers*. CoRR/arXiv 2023	Method
Safety, concept erasure, and unlearning	[321] *Selective Amnesia: A Continual Learning Approach to Forgetting in Deep Generative Models*. CoRR/arXiv 2023	Method
Safety, concept erasure, and unlearning	[322] *Towards Safe Self-Distillation of Internet-Scale Text-to-Image Diffusion Models*. CoRR/arXiv 2023	Method
Safety and concept-erasure benchmarks	[323] *T2I-RiskyPrompt: A Benchmark for Safety Evaluation, Attack, and Defense on Text-to-Image Model*. CoRR/arXiv 2025	Benchmark / evaluation

Continued on the next page

Subcategory	Additional work	Role
Safety and concept-erasure benchmarks	[324] *UnlearnCanvas: A Stylized Image Dataset to Benchmark Machine Unlearning for Diffusion Models*. NeurIPS 2024	Benchmark / evaluation
Physical, causal, and world consistency	[325] *Do generative video models understand physical principles?*. arXiv preprint arXiv:2501.09038 2025	Benchmark / evaluation
Physical, causal, and world consistency	[326] *How Far Is Video Generation from World Model: A Physical Law Perspective*. ICML 2025	Benchmark / evaluation
Physical, causal, and world consistency	[327] *DriveArena: A Closed-loop Generative Simulation Platform for Autonomous Driving*. CoRR/arXiv 2024	Benchmark / evaluation
Physical, causal, and world-consistency benchmarks	[328] *Morpheus: Benchmarking Physical Reasoning of Video Generative Models with Real Physical Experiments*. CoRR/arXiv 2025	Benchmark / evaluation
Physical, causal, and world-consistency benchmarks	[329] *WorldModelBench: Judging Video Generation Models As World Models*. CoRR/arXiv 2025	Benchmark / evaluation
Physical, causal, and world-consistency benchmarks	[330] *WorldScore: A Unified Evaluation Benchmark for World Generation*. arXiv preprint arXiv:2504.00983 2025	Benchmark / evaluation
Physical, causal, and world-consistency benchmarks	[331] *WorldSimBench: Towards Video Generation Models as World Simulators*. CoRR/arXiv 2024	Benchmark / evaluation
Physical, causal, and world-consistency benchmarks	[332] *IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning*. IEEE Trans. Pattern Anal. Mach. Intell. 2021	Benchmark / evaluation
Normative consistency methods	[333] *DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors*. ECCV 2024	Method
Normative consistency methods	[334] *GuardT2I: Defending Text-to-Image Models from Adversarial Prompts*. CoRR/arXiv 2024	Method
Normative consistency methods	[335] *CLEVRER: CoLlision Events for Video REpresentation and Reasoning*. ICLR 2020	Method

REFERENCES

[1] T. Hu, J. Zhang, H. Huang *et al.*, "Evolution of video generative foundations," *CoRR*, vol. abs/2604.06339, 2026.
[2] C. Kim and Y. Qi, "A comprehensive survey on concept erasure in text-to-image diffusion models," *CoRR*, vol. abs/2502.14896, 2025.
[3] Y. Ma, K. Feng, Z. Hu *et al.*, "Controllable video generation: A survey," *CoRR*, vol. abs/2507.16869, 2025.
[4] Y. Hu, L. Wang, X. Liu *et al.*, "Simulating the real world: A unified survey of multimodal generative models," *CoRR*, vol. abs/2503.04641, 2025.
[5] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied AI," *CoRR*, vol. abs/2405.14093, 2024.
[6] L. Ruan and Q. Jin, "Survey: Transformer based video-language pre-training," *AI Open*, vol. 3, pp. 1–13, 2022.
[7] A. P. Team, "CogVideoX-Fun: Text-to-video generation with flexible resolution and duration," GitHub repository, 2025, https://github.com/aigc-apps/VideoX-Fun.
[8] Z. Yang, J. Teng, W. Zheng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," in *ICLR*. OpenReview.net, 2025.
[9] K. Team, "Kling-omni technical report," *CoRR*, vol. abs/2512.16776, 2025.
[10] Z. Fei, D. Li, D. Qiu *et al.*, "Skyreels-a2: Compose anything in video diffusion transformers," *CoRR*, vol. abs/2504.02436, 2025.
[11] Z. Fei, H. Jiang, D. Qiu *et al.*, "Skyreels-audio: Omni audio-conditioned talking portraits in video diffusion transformers," *CoRR*, vol. abs/2506.00830, 2025.
[12] G. Chen, D. Lin, J. Yang *et al.*, "Skyreels-v2: Infinite-length film generative model," *CoRR*, vol. abs/2504.13074, 2025.
[13] S. Team, "Step-video-t2v technical report: The practice, challenges, and future of video foundation model," *CoRR*, vol. abs/2502.10248, 2025.
[14] W. Fan, C. Si, J. Song *et al.*, "Vchitect-2.0: Parallel transformer for scaling up video diffusion models," *CoRR*, vol. abs/2501.08453, 2025.
[15] Y. Zhou, Q. Wang, Y. Cai, and H. Yang, "Allegro: Open the black box of commercial-level video generation model," *CoRR*, vol. abs/2410.15458, 2024.
[16] W. Kong, Q. Tian, Z. Zhang *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *CoRR*, vol. abs/2412.03603, 2024.
[17] O. Bar-Tal, H. Chefer, O. Tov *et al.*, "Lumiere: A space-time diffusion model for video generation," in *SIGGRAPH Asia*. ACM, 2024.
[18] T. M. G. team, "Movie gen: A cast of media foundation models," *CoRR*, vol. abs/2410.13720, 2024.
[19] B. Lin, Y. Ge, X. Cheng *et al.*, "Open-sora plan: Open-source large video generation model," *CoRR*, vol. abs/2412.00131, 2024.
[20] OpenAI, "Video generation models as world simulators," Technical report, 2024.
[21] D. Kondratyuk, L. Yu, X. Gu *et al.*, "VideoPoet: A large language model for zero-shot video generation," in *ICML*. PMLR, 2024.
[22] F. Bao, C. Xiang, G. Yue *et al.*, "Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models," *CoRR*, vol. abs/2405.04233, 2024.
[23] S. Zhang, J. Wang, Y. Zhang *et al.*, "I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models," *CoRR*, vol. abs/2311.04145, 2023.
[24] R. Villegas, M. Babaeizadeh, P. Kindermans *et al.*, "Phenaki: Variable length video generation from open domain textual descriptions," in *ICLR*. OpenReview.net, 2023.
[25] A. Blattmann, T. Dockhorn, S. Kulal *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *CoRR*, vol. abs/2311.15127, 2023.
[26] J. Ho, W. Chan, C. Saharia *et al.*, "Imagen video: High definition video generation with diffusion models," *CoRR*, vol. abs/2210.02303, 2022.
[27] Z. Lai, Y. Zhao, H. Liu *et al.*, "Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details," *CoRR*, vol. abs/2506.16504, 2025.
[28] T. Huang, W. Zheng, T. Wang *et al.*, "Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation," *ACM Trans. Graph.*, vol. 44, no. 6, pp. 245:1–245:15, 2025.
[29] J. Sun, B. Zhang, R. Shao *et al.*, "DreamCraft3D: Hierarchical 3D generation with bootstrapped diffusion prior," in *ICLR*. OpenReview.net, 2024.
[30] Y. Xu, Z. Shi, W. Yifan *et al.*, "GRM: Large gaussian reconstruction model for efficient 3D reconstruction and generation," in *ECCV*. Springer, 2024.
[31] J. Li, H. Tan, K. Zhang *et al.*, "Instant3D: Fast text-to-3D with sparse-view generation and large reconstruction model," in *ICLR*. OpenReview.net, 2024.

[32] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu, "LGM: Large multi-view gaussian model for high-resolution 3D content creation," in *ECCV*. Springer, 2024.
[33] Y. Chen, T. He, D. Huang *et al.*, "MeshAnything: Artist-created mesh generation with autoregressive transformers," *CoRR*, vol. abs/2406.10163, 2024.
[34] Z. He and T. Wang, "OpenLRM: Open-source large reconstruction models," GitHub repository, 2024, https://github.com/3DTopia/OpenLRM.
[35] D. Tochilkin, D. Pankratz, Z. Liu *et al.*, "Triposr: Fast 3d object reconstruction from a single image," *CoRR*, vol. abs/2403.02151, 2024.
[36] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *ICLR*. OpenReview.net, 2023.
[37] R. Chen, Y. Chen, N. Jiao, and K. Jia, "Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation," in *ICCV*. Computer Vision Foundation / IEEE, 2023.
[38] T. Yi, J. Fang, G. Wu *et al.*, "GaussianDreamer: Fast generation from text to 3D gaussian splatting with point cloud priors," *CoRR*, vol. abs/2310.08529, 2023.
[39] C. Lin, J. Gao, L. Tang *et al.*, "Magic3d: High-resolution text-to-3d content creation," in *CVPR*. IEEE, 2023, pp. 300–309.
[40] M. Liu, C. Xu, H. Jin *et al.*, "One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization," in *NeurIPS*. Curran Associates, Inc., 2023.
[41] Z. Wang, C. Lu, Y. Wang *et al.*, "ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation," in *NeurIPS*. Curran Associates, Inc., 2023.
[42] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*. IEEE Computer Society, 2017, pp. 2432–2443.
[43] A. X. Chang, T. A. Funkhouser, L. J. Guibas *et al.*, "Shapenet: An information-rich 3d model repository," *CoRR*, vol. abs/1512.03012, 2015.
[44] N. Agarwal, A. Ali, M. Bala *et al.*, "Cosmos world foundation model platform for physical AI," *CoRR*, vol. abs/2501.03575, 2025.
[45] G. Zhao, C. Ni, X. Wang *et al.*, "Drivedreamer4d: World models are effective data machines for 4d driving scene representation," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 12 015–12 026.
[46] L. Russell, A. Hu, L. Bertoni *et al.*, "GAIA-2: A controllable multi-view generative world model for autonomous driving," *CoRR*, vol. abs/2503.20523, 2025.
[47] Z. Wang, X. Wei, B. Li *et al.*, "VideoVerse: Does your T2V generator have world model capability to synthesize videos?" *CoRR*, vol. abs/2510.08398, 2025.
[48] G. Zhao, X. Wang, Z. Zhu *et al.*, "Drivedreamer-2: Llm-enhanced world models for diverse driving video generation," *CoRR*, vol. abs/2403.06845, 2024.
[49] X. Wang, Z. Zhu, G. Huang, X. Chen, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *CoRR*, vol. abs/2309.09777, 2023.
[50] A. Hu, L. Russell, H. Yeo *et al.*, "GAIA-1: A generative world model for autonomous driving," *CoRR*, vol. abs/2309.17080, 2023.
[51] L. Kong, J. Zhang, D. Zou *et al.*, "Deblurdiff: Real-world image deblurring with generative diffusion models," *CoRR*, vol. abs/2502.03810, 2025.
[52] Y. Lan, Z. Cui, C. Liu *et al.*, "Exploiting diffusion prior for real-world image dehazing with unpaired training," in *AAAI*, T. Walsh, J. Shah, and Z. Kolter, Eds. AAAI Press, 2025, pp. 4455–4463.
[53] R. Wang, Y. Zheng, Z. Zhang *et al.*, "Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 23 091–23 100.
[54] S. Man, G. Ohayon, R. Raphaeli, and M. Elad, "Proxies for distortion and consistency with applications for real-world image restoration," *CoRR*, vol. abs/2501.12102, 2025.
[55] L. Dong, Q. Fan, Y. Guo *et al.*, "TSD-SR: one-step diffusion with target score distillation for real-world image super-resolution," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 23 174–23 184.
[56] Y. Ai, X. Zhou, H. Huang *et al.*, "Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation," in *NeurIPS*, A. Globersons, L. Mackey, D. Belgrave *et al.*, Eds., 2024.
[57] W. Wang, H. Yang, J. Fu, and J. Liu, "Zero-reference low-light enhancement via physical quadruple priors," in *CVPR*. IEEE, 2024, pp. 26 057–26 066.
[58] Z. Wang, J. Bao, S. Gu, D. Chen, W. Zhou, and H. Li, "Designdiffusion: High-quality text-to-design image generation with diffusion models," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 20 906–20 915.
[59] O. Zafar, Y. Cohen, L. Wolf, and I. Schwartz, "Detection-driven object count optimization for text-to-image diffusion models," 2025.
[60] H. Wang, Y. Xu, Y. Li *et al.*, "RepText: Rendering visual text via replicating," *CoRR*, vol. abs/2504.19724, 2025.
[61] S. Yang, L. Hou, H. Huang *et al.*, "Direct-a-video: Customized video generation with user-directed camera movement and object motion," in *SIGGRAPH*. ACM, 2024.
[62] Z. Liu, W. Liang, Z. Liang *et al.*, "Glyph-ByT5: A customized text encoder for accurate visual text rendering," in *ECCV*. Springer, 2024.
[63] J. Chen, Y. Huang, T. Lv, L. Cui, Q. Chen, and F. Wei, "Textdiffuser-2: Unleashing the power of language models for text rendering," in *ECCV*. Springer, 2024.
[64] Y. Zhao and Z. Lian, "U-DiffText: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models," in *ECCV*. Springer, 2024.
[65] J. Ma, M. Zhao, C. Chen *et al.*, "Glyphdraw: Learning to draw chinese characters in image synthesis models coherently," *CoRR*, vol. abs/2303.17870, 2023.
[66] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik, "Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment," in *NeurIPS*, 2023.
[67] Z. Li, H. Luo, X. Shuai, and H. Ding, "Anyi2v: Animating any conditional image with motion control," *CoRR*, vol. abs/2507.02857, 2025.
[68] Y. Xie, F. Feng, R. Shi, J. Wang, Y. Rui, and X. Geng, "Divcontrol: Knowledge diversion for controllable image generation," *CoRR*, vol. abs/2507.23620, 2025.
[69] A. Suleyman and G. Biricik, "Grounding text-to-image diffusion models for controlled high-quality image generation," *CoRR*, vol. abs/2501.09194, 2025.
[70] D. Geng, C. Herrmann, J. Hur *et al.*, "Motion prompting: Controlling video generation with motion trajectories," in *CVPR*. Computer Vision Foundation / IEEE, 2025.
[71] H. Hsu and Y. Peng, "Postero: Structuring layout trees to enable language models in generalized content-aware layout generation," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 8117–8127.
[72] F. Yu, J. Gu, J. Hu, Z. Li, and C. Dong, "Unicon: Unidirectional information flow for effective control of large-scale diffusion models," in *ICLR*. OpenReview.net, 2025.
[73] Y. Li, M. Keuper, D. Zhang, and A. Khoreva, "Adversarial supervision makes layout-to-image diffusion models thrive," in *ICLR*. OpenReview.net, 2024.
[74] H. He, Y. Xu, Y. Guo *et al.*, "CameraCtrl: Enabling camera control for text-to-video generation," *CoRR*, vol. abs/2404.02101, 2024.
[75] M. Li, T. Yang, H. Kuang *et al.*, "Controlnet++: Improving conditional controls with efficient consistency feedback," in *ECCV*. Springer, 2024.

[76] D. Zavadski, J. Feiden, and C. Rother, "ControlNet-XS: Rethinking the control of text-to-image diffusion models as feedback-control systems," in *ECCV*. Springer, 2024.

[77] B. Peng, J. Wang, Y. Zhang, W. Li, M. Yang, and J. Jia, "ControlNeXt: Powerful and efficient control for image and video generation," *CoRR*, vol. abs/2408.06070, 2024.

[78] Z. Gan and G. Ye, "Doglayout: Denoising diffusion GAN for discrete and continuous layout generation," *CoRR*, vol. abs/2412.00381, 2024.

[79] X. Wang, T. Darrell, S. S. Rambhatla, R. Girdhar, and I. Misra, "Instancediffusion: Instance-level control for image generation," in *CVPR*. Computer Vision Foundation / IEEE, 2024, pp. 6232–6242.

[80] Z. Wang, Z. Yuan, X. Wang *et al.*, "MotionCtrl: A unified and flexible motion controller for video generation," in *SIGGRAPH*. ACM, 2024.

[81] Y. Guo, C. Yang, A. Rao, M. Agrawala, D. Lin, and B. Dai, "SparseCtrl: Adding sparse controls to text-to-video diffusion models," in *ECCV*. Springer, 2024.

[82] W. K. Ma, J. P. Lewis, and W. B. Kleijn, "TrailBlazer: Trajectory control for diffusion-based video generation," in *SIGGRAPH Asia*. ACM, 2024.

[83] W. Chen, J. Wu, P. Xie *et al.*, "Control-a-video: Controllable text-to-video generation with diffusion models," *CoRR*, vol. abs/2305.13840, 2023.

[84] Y. Kim, J. Lee, J. Kim, J. Ha, and J. Zhu, "DenseDiffusion: Dense text-to-image generation with attention modulation," in *ICCV*. Computer Vision Foundation / IEEE, 2023.

[85] E. Levi, E. Brosh, M. Mykhailych, and M. Perez, "DLT: conditioned layout generation with joint discrete-continuous diffusion layout transformer," in *ICCV*. IEEE, 2023, pp. 2106–2115.

[86] H. Xue, Z. Huang, Q. Sun, L. Song, and W. Zhang, "Freestyle layout-to-image synthesis," in *CVPR*. IEEE, 2023, pp. 14 256–14 266.

[87] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "LayoutDiffusion: Controllable diffusion model for layout-to-image generation," in *CVPR*. Computer Vision Foundation / IEEE, 2023.

[88] S. Chai, L. Zhuang, and F. Yan, "Layoutdm: Transformer-based diffusion model for layout generation," in *CVPR*. IEEE, 2023, pp. 18 349–18 358.

[89] W. Feng, W. Zhu, T. Fu *et al.*, "LayoutGPT: Compositional visual planning and generation with large language models," in *NeurIPS*. Curran Associates, Inc., 2023.

[90] L. Lian, B. Li, A. Yala, and T. Darrell, "Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models," *CoRR*, vol. abs/2305.13655, 2023.

[91] Y. Zeng, Z. Lin, J. Zhang *et al.*, "SceneComposer: Any-level semantic image synthesis," in *CVPR*. Computer Vision Foundation / IEEE, 2023.

[92] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," in *SIGGRAPH*, E. Brunvand, A. Sheffer, and M. Wimmer, Eds. ACM, 2023, pp. 55:1–55:11.

[93] X. Wang, H. Yuan, S. Zhang *et al.*, "VideoComposer: Compositional video synthesis with motion controllability," in *NeurIPS*. Curran Associates, Inc., 2023.

[94] Z. Li, Q. Zhou, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Guiding text-to-image diffusion model towards grounded generation," *CoRR*, vol. abs/2301.05221, 2023.

[95] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "Multidiffusion: Fusing diffusion paths for controlled image generation," in *ICML*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds. PMLR, 2023, pp. 1737–1752.

[96] J. Shin, A. Hwang, Y. Kim, D. Kim, and J. Park, "Exploring multimodal diffusion transformers for enhanced prompt-based image editing," *CoRR*, vol. abs/2508.07519, 2025.

[97] V. Kulikov, M. Kleiner, I. Huberman-Spiegelglas, and T. Michaeli, "Flowedit: Inversion-free text-based editing using pre-trained flow models," in *ICCV*. Computer Vision Foundation / IEEE, 2025.

[98] Z. Zhang, H. Liu, J. Chen, and X. Xu, "Gooddrag: Towards good practices for drag editing with diffusion models," in *ICLR*. OpenReview.net, 2025.

[99] R. Huang, C. Wang, J. Yang *et al.*, "ILLUME+: illuminating unified MLLM with dual visual tokenization and diffusion refinement," *CoRR*, vol. abs/2504.01934, 2025.

[100] J. Lu and K. Han, "Inpaint4drag: Repurposing inpainting models for drag-based image editing via bidirectional warping," *CoRR*, vol. abs/2509.04582, 2025.

[101] Z. Liu, Y. Yu, H. Ouyang *et al.*, "MagicQuill: An intelligent interactive image editing system," in *CVPR*. Computer Vision Foundation / IEEE, 2025.

[102] C. Wei, Z. Xiong, W. Ren, X. Du, G. Zhang, and W. Chen, "OmniEdit: Building image editing generalist models through specialist supervision," in *ICLR*. OpenReview.net, 2025.

[103] G. Parmar, T. Park, S. Narasimhan, and J. Zhu, "One-step image translation with text-to-image models," in *AAAI*. AAAI Press, 2025.

[104] L. Ji, L. Zhong, P. Wei, and C. Li, "Posetraj: Pose-aware trajectory control in video diffusion," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 22 776–22 785.

[105] J. Kim, J. Park, Y. Song, N. Kwak, and W. Rhee, "Reflex: Text-guided editing of real images in rectified flow via mid-step feature extraction and attention adaptation," *CoRR*, vol. abs/2507.01496, 2025.

[106] L. Rout, Y. Chen, N. Ruiz, C. Caramanis, S. Shakkottai, and W. Chu, "Semantic image inversion and editing using rectified stochastic differential equations," in *ICLR*. OpenReview.net, 2025.

[107] Y. Wang, L. Guo, Z. Li *et al.*, "Training-free text-guided image editing with visual autoregressive model," in *ICCV*. Computer Vision Foundation / IEEE, 2025.

[108] Y. Tu, H. Luo, X. Chen, S. Ji, X. Bai, and H. Zhao, "VideoAnydoor: High-fidelity video object insertion with precise motion control," in *SIGGRAPH*. ACM, 2025.

[109] J. Zhuang, Y. Zeng, W. Liu, C. Yuan, and K. Chen, "A task is worth one word: Learning with task prompts for high-quality versatile image inpainting," in *ECCV*. Springer, 2024.

[110] Y. Tuo, Y. Geng, and L. Bo, "AnyText2: Visual text generation and editing with customizable attributes," *CoRR*, vol. abs/2411.15245, 2024.

[111] Y. Tuo, W. Xiang, J. He, Y. Geng, and X. Xie, "Anytext: Multilingual visual text generation and editing," in *ICLR*. OpenReview.net, 2024.

[112] Y. Li, Y. Bian, X. Ju *et al.*, "BrushEdit: All-in-one image inpainting and editing," *CoRR*, vol. abs/2412.10316, 2024.

[113] X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, and Q. Xu, "Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion," in *ECCV*. Springer, 2024.

[114] H. Nam, G. Kwon, G. Y. Park, and J. C. Ye, "Contrastive denoising score for text-guided latent diffusion image editing," in *CVPR*. IEEE, 2024, pp. 9192–9201.

[115] X. Ju, A. Zeng, Y. Bian, S. Liu, and Q. Xu, "Direct inversion: Boosting diffusion-based editing with 3 lines of code," in *ICLR*. OpenReview.net, 2024.

[116] C. Mou, X. Wang, J. Song, Y. Shan, and J. Zhang, "Dragondiffusion: Enabling drag-style manipulation on diffusion models," in *ICLR*. OpenReview.net, 2024.

[117] P. Ling, L. Chen, P. Zhang, H. Chen, Y. Jin, and J. Zheng, "Freedrag: Feature dragging for reliable point-based image editing," in *CVPR*. IEEE, 2024, pp. 6860–6870.

[118] T. Fu, W. Hu, X. Du, W. Y. Wang, Y. Yang, and Z. Gan, "Guiding instruction-based image editing via multimodal large language models," in *ICLR*. OpenReview.net, 2024.

[119] S. Xu, Y. Huang, J. Pan, Z. Ma, and J. Chai, "Infedit: Inversion-free image editing with natural language," in *CVPR*. Computer Vision Foundation / IEEE, 2024, pp. 9452–9461.

[120] N. Starodubcev, M. Khoroshikh, A. Babenko, and D. Baranchuk, "Invertible consistency distillation for text-guided image editing in around 7 steps," in *NeurIPS*, A. Globersons, L. Mackey, D. Belgrave *et al.*, Eds., 2024.

[121] M. Brack, F. Friedrich, K. Kornmeier *et al.*, "LEDITS++: Limitless image editing using text-to-image models," in *CVPR*. Computer Vision Foundation / IEEE, 2024.

[122] G. Luo, T. Darrell, O. Wang, D. B. Goldman, and A. Holynski, "Readout guidance: Learning control from diffusion features," in *CVPR*. IEEE, 2024, pp. 8217–8227.

[123] Y. Wang, C. Cao, K. Fan *et al.*, "Repositioning the subject within image," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

[124] Y. Lin, Y. Chen, Y. Tsai, L. Jiang, and M. Yang, "Text-driven image editing via learnable regions," in *CVPR*. IEEE, 2024, pp. 7059–7068.

[125] H. Zhao, X. Ma, L. Chen *et al.*, "UltraEdit: Instruction-based fine-grained image editing at scale," in *NeurIPS*. Curran Associates, Inc., 2024.

[126] Y. Tian, L. Yang, H. Yang *et al.*, "VideoTetris: Towards compositional text-to-video generation with multi-concept control," in *NeurIPS*. Curran Associates, Inc., 2024.

[127] X. Chen, Y. Feng, M. Chen *et al.*, "Zero-shot image editing with reference imitation," *CoRR*, vol. abs/2406.07547, 2024.

[128] Z. Huang, K. C. K. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *CVPR*. IEEE, 2023, pp. 6080–6090.

[129] S. Kim, K. Lee, J. S. Choi, J. Jeong, K. Sohn, and J. Shin, "Collaborative score distillation for consistent visual synthesis," *CoRR*, vol. abs/2307.04787, 2023.

[130] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your GAN: interactive point-based manipulation on the generative image manifold," in *SIGGRAPH*, E. Brunvand, A. Sheffer, and M. Wimmer, Eds. ACM, 2023, pp. 78:1–78:11.

[131] S. Yin, C. Wu, J. Liang *et al.*, "DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory," *CoRR*, vol. abs/2308.08089, 2023.

[132] G. Y. Park, J. Kim, B. Kim, S. W. Lee, and J. C. Ye, "Energy-based cross attention for bayesian context update in text-to-image diffusion models," in *NeurIPS*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[133] T. Yu, R. Feng, R. Feng *et al.*, "Inpaint anything: Segment anything meets image inpainting," *CoRR*, vol. abs/2304.06790, 2023.

[134] O. Patashnik, D. Garibi, I. Azuri, H. Averbuch-Elor, and D. Cohen-Or, "Localizing object-level shape variations with text-to-image diffusion models," in *ICCV*. IEEE, 2023, pp. 22 994–23 004.

[135] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *ICCV*. IEEE, 2023, pp. 22 503–22 513.

[136] T. Nguyen, Y. Li, U. Ojha, and Y. J. Lee, "Visual instruction inversion: Image editing via image prompting," in *NeurIPS*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[137] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J. Zhu, "Zero-shot image-to-image translation," in *SIGGRAPH*, E. Brunvand, A. Sheffer, and M. Wimmer, Eds. ACM, 2023, pp. 11:1–11:11.

[138] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *CVPR*. Computer Vision Foundation / IEEE, 2022.

[139] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *CVPR*. Computer Vision Foundation / IEEE, 2022.

[140] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool, "RePaint: Inpainting using denoising diffusion probabilistic models," in *CVPR*. Computer Vision Foundation / IEEE, 2022.

[141] C. H. Wu and F. D. la Torre, "Unifying diffusion models' latent space, with applications to cyclediffusion and guidance," *CoRR*, vol. abs/2210.05559, 2022.

[142] Y. Wan and K. Chang, "CompAlign: Improving compositional text-to-image generation with a complex benchmark and fine-grained feedback," *CoRR*, vol. abs/2505.11178, 2025.

[143] T. Zhang, X. Wang, L. Li *et al.*, "STRICT: Stress test of rendering images containing text," in *EMNLP*. Association for Computational Linguistics, 2025.

[144] J. Wang, H. Duan, J. Wang, Z. Jia, G. Zhai, and X. Min, "TIT-Score: Evaluating long-prompt based text-to-image alignment via text-to-image-to-text consistency," *CoRR*, vol. abs/2510.02987, 2025.

[145] B. Li, Z. Lin, D. Pathak *et al.*, "GenAI-Bench: Evaluating and improving compositional text-to-visual generation," in *CVPR Workshop on Synthetic Data for Computer Vision*, 2024.

[146] M. Ku, D. Jiang, C. Wei, X. Yue, and W. Chen, "VIEScore: Towards explainable metrics for conditional image synthesis evaluation," in *ACL*. Association for Computational Linguistics, 2024.

[147] O. Michel, A. Bhattad, E. VanderBilt, R. Krishna, A. Kembhavi, and T. Gupta, "OBJECT 3DIT: Language-guided 3d-aware image editing," in *NeurIPS*. Curran Associates, Inc., 2023.

[148] H. Feng, Z. Huang, L. Li, and L. Sheng, "Personalize anything for free with diffusion transformer," in *AAAI*, S. Koenig, C. Jenkins, and M. E. Taylor, Eds. AAAI Press, 2026, pp. 3921–3929.

[149] Z. Mao, M. Huang, F. Ding, M. Liu, Q. He, and Y. Zhang, "Realcustom++: Representing images as real textual word for real-time customization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 2, pp. 2078–2095, 2026.

[150] M. Wang, H. Ding, J. Peng, Y. Zhao, Y. Chen, and Y. Wei, "Characonsist: Fine-grained consistent character generation," *CoRR*, vol. abs/2507.11533, 2025.

[151] H. Lee, K. C. K. Chan, and M. Yang, "Cocoins: Consistent subject generation via contrastive instantiated concepts," *Trans. Mach. Learn. Res.*, vol. 2025, 2025.

[152] Z. Mai and Y. Tai, "ContextAnyone: Context-aware diffusion for character-consistent text-to-video generation," *CoRR*, vol. abs/2512.07328, 2025.

[153] J. Wu, C. Tang, J. Wang, Y. Zeng, X. Li, and Y. Tong, "Diffsensei: Bridging multi-modal llms and diffusion models for customized manga generation," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 28 684–28 693.

[154] N. Kumari, X. Yin, J. Zhu, I. Misra, and S. Azadi, "Generating multi-image synthetic data for text-to-image customization," *CoRR*, vol. abs/2502.01720, 2025.

[155] S. Yuan, J. Huang, X. He *et al.*, "Identity-preserving text-to-video generation by frequency decomposition," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 12 978–12 988.

[156] F. Shen, X. Jiang, X. He *et al.*, "Imagdressing-v1: Customizable virtual dressing," in *AAAI*, T. Walsh, J. Shah, and Z. Kolter, Eds. AAAI Press, 2025, pp. 6795–6804.

[157] J. Tao, Y. Zhang, Q. Wang *et al.*, "InstantCharacter: Personalize any characters with a scalable diffusion transformer framework," *CoRR*, vol. abs/2504.12395, 2025.

[158] S. Tu, Z. Xing, X. Han *et al.*, "Stableanimator: High-quality identity-preserving human image animation," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 21 096–21 106.
[159] Z. Chen, S. Wang, and Y. Zhou, "Styleblend: Enhancing style-specific content creation in text-to-image diffusion models," *Comput. Graph. Forum*, vol. 44, no. 2, 2025.
[160] Z. Zhang, J. Liao, X. Meng, L. Qin, and W. Wang, "Tora2: Motion and appearance customized diffusion transformer for multi-entity video generation," in *ACM MM*, C. Gurrin, K. Schoeffmann, M. Zhang *et al.*, Eds. ACM, 2025, pp. 9434–9443.
[161] L. Hu, X. Gao, P. Zhang, K. Sun, B. Zhang, and L. Bo, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *CVPR*. Computer Vision Foundation / IEEE, 2024, pp. 8153–8163.
[162] F. Wang, Z. Huang, X. Shi *et al.*, "AnimateLCM: Accelerating the animation of personalized diffusion models and adapters with decoupled consistency learning," in *SIGGRAPH Asia*. ACM, 2024.
[163] J. Cheng, X. Lu, H. Li *et al.*, "Autostudio: Crafting consistent subjects in multi-turn interactive image generation," *CoRR*, vol. abs/2406.01388, 2024.
[164] Z. Wang, A. Li, L. Zhu, Y. Guo, Q. Dou, and Z. Li, "CustomVideo: Customizing text-to-video generation with multiple subjects," *CoRR*, vol. abs/2401.09962, 2024.
[165] X. Cui, Z. Li, P. Li, H. Huang, X. Liu, and Z. He, "INSTASTYLE: inversion noise of a stylized image is secretly a style adviser," in *ECCV*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Springer, 2024, pp. 455–472.
[166] Z. Ma, D. Zhou, C. Yeh *et al.*, "Magic-me: Identity-specific video customized diffusion," *CoRR*, vol. abs/2402.09368, 2024.
[167] D. Zhou, J. Huang, J. Bai *et al.*, "Magictailor: Component-controllable personalization in text-to-image diffusion models," *CoRR*, vol. abs/2410.13370, 2024.
[168] I. Lopes, F. Pizzati, and R. de Charette, "Material palette: Extraction of materials from a single image," in *CVPR*. IEEE, 2024, pp. 4379–4388.
[169] J. Wang, C. Yan, H. Lin, and W. Zhang, "Oneactor: Consistent character generation via cluster-conditioned guidance," *CoRR*, vol. abs/2404.10267, 2024.
[170] Z. Guo, Y. Wu, Z. Chen, L. Chen, P. Zhang, and Q. He, "PuLID: Pure and lightning ID customization via contrastive alignment," in *NeurIPS*. Curran Associates, Inc., 2024.
[171] M. Huang, Z. Mao, M. Liu, Q. He, and Y. Zhang, "Realcustom: Narrowing real text word for real-time open-domain text-to-image customization," in *CVPR*. IEEE, 2024, pp. 7476–7485.
[172] A. Hertz, A. Voynov, S. Fruchter, and D. Cohen-Or, "Style aligned image generation via shared attention," in *CVPR*. IEEE, 2024, pp. 4775–4785.
[173] J. Ma, J. Liang, C. Chen, and H. Lu, "Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning," in *SIGGRAPH*. ACM, 2024.
[174] H. Chefer, O. Lang, M. Geva *et al.*, "The hidden language of diffusion models," in *ICLR*. OpenReview.net, 2024.
[175] B. Chen, B. Curless, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Total selfie: Generating full-body selfies," in *CVPR*. IEEE, 2024, pp. 6701–6711.
[176] T. Y. Cheng, P. Sharma, A. Markham, N. Trigoni, and V. Jampani, "Zest: Zero-shot material transfer from a single image," in *ECCV*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Springer, 2024, pp. 370–386.
[177] B. N. Zhao, Y. Xiao, J. Xu *et al.*, "Dreamdistribution: Prompt distribution learning for text-to-image diffusion models," *CoRR*, vol. abs/2312.14216, 2023.
[178] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo, "ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation," in *ICCV*. Computer Vision Foundation / IEEE, 2023, pp. 15 943–15 953.
[179] C. Zhang, X. Chen, S. Chai *et al.*, "Iti-gen: Inclusive text-to-image generation," in *ICCV*. IEEE, 2023, pp. 3946–3957.
[180] N. Liu, Y. Du, S. Li, J. B. Tenenbaum, and A. Torralba, "Unsupervised compositional concepts discovery with text-to-image generative models," in *ICCV*. IEEE, 2023, pp. 2085–2095.
[181] T. Soucek, P. Gatti, M. Wray, I. Laptev, D. Damen, and J. Sivic, "Showhowto: Generating scene-conditioned step-by-step visual instructions," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 27 435–27 445.
[182] C. Liu, H. Wu, Y. Zhong, X. Zhang, Y. Wang, and W. Xie, "Intelligent grimm - open-ended visual storytelling via latent diffusion models," in *CVPR*. IEEE, 2024, pp. 6190–6200.
[183] J. Bengtson, D. Nilsson, D. I. Lee, and F. Kahl, "3d-consistent multi-view editing by diffusion guidance," *CoRR*, vol. abs/2511.22228, 2025.
[184] Z. Pan, Z. Yang, X. Zhu, and L. Zhang, "Efficient4d: Fast dynamic 3d object generation from a single-view video," *International Journal of Computer Vision*, vol. 134, no. 1, 2025.
[185] S. Fang, C. Yu, F. Wang, and Q. Huang, "MVRoom: Controllable 3D indoor scene generation with multi-view diffusion models," *CoRR*, vol. abs/2512.04248, 2025.
[186] Y. Xie, C. Yao, V. Voleti, H. Jiang, and V. Jampani, "SV4D 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4D generation," *CoRR*, vol. abs/2503.16396, 2025.
[187] ——, "SV4D: Dynamic 3D content generation with multi-frame and multi-view consistency," in *ICLR*. OpenReview.net, 2025.
[188] D. Danier, G. Gao, S. McDonagh, C. Li, H. Bilen, and O. M. Aodha, "ViCoDR: View-consistent diffusion representations for 3d-consistent video generation," *CoRR*, vol. abs/2511.18991, 2025.
[189] R. Gao, A. Holynski, P. Henzler *et al.*, "CAT3D: Create anything in 3D with multi-view diffusion models," in *NeurIPS*. Curran Associates, Inc., 2024.
[190] S. Zhu, J. L. Chen, Z. Dai *et al.*, "Champ: Controllable and consistent human image animation with 3D parametric guidance," in *ECCV*. Springer, 2024.
[191] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao, "Consistent4d: Consistent 360° dynamic object generation from monocular video," in *ICLR*. OpenReview.net, 2024.
[192] P. Li, Y. Liu, X. Long *et al.*, "Era3D: High-resolution multiview diffusion using efficient row-wise attention," in *NeurIPS*. Curran Associates, Inc., 2024.
[193] X. Xu, W. Ge, J. Lin *et al.*, "Flexgen: Flexible multi-view generation from text and image inputs," *CoRR*, vol. abs/2410.10745, 2024.
[194] G. Qian, J. Mai, A. Hamdi *et al.*, "Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors," in *ICLR*. OpenReview.net, 2024.
[195] L. Uzolas, E. Eisemann, and P. Kellnhofer, "Motiondreamer: Zero-shot 3d mesh animation from video diffusion models," *CoRR*, vol. abs/2405.20155, 2024.
[196] V. Voleti, C. Yao, M. Boss *et al.*, "SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion," in *ECCV*. Springer, 2024.
[197] K. Wu, F. Liu, Z. Cai *et al.*, "Unique3D: High-quality and efficient 3D mesh generation from a single image," in *NeurIPS*. Curran Associates, Inc., 2024.
[198] Y. Huang, J. Wang, A. Zeng *et al.*, "DreamWaltz: Make a scene with complex 3d animatable avatars," in *NeurIPS*. Curran Associates, Inc., 2023.

[199] A. Athar, S. Mahadevan, A. Osep, L. Leal-Taixé, and B. Leibe, "Stem-seg: Spatio-temporal embeddings for instance segmentation in videos," in *ECCV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Springer, 2020, pp. 158–177.

[200] D. X. Long, Y. Song, M. Kan, T. Pfister, and L. T. Le, "A^2RD: Agentic autoregressive diffusion for long video consistency," *CoRR*, vol. abs/2605.06924, 2026.

[201] F. Waseem and M. Shahzad, "Video is worth a thousand images: Exploring the latest trends in long video generation," *ACM Comput. Surv.*, vol. 58, no. 6, pp. 154:1–154:35, 2026.

[202] G. Lei, C. Wang, R. Zhang, Y. Wang, H. Li, and W. Xu, "Animateanything: Consistent and controllable animation for video generation," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 27 946–27 956.

[203] W. Gao, X. Lan, and S. Yang, "ANYPORTAL: zero-shot consistent video background replacement," *CoRR*, vol. abs/2509.07472, 2025.

[204] S. Zheng, J. Cai, Y. Guan *et al.*, "High-fidelity and long-duration human image animation with diffusion transformer," *CoRR*, vol. abs/2512.21905, 2025.

[205] Y. HaCohen, N. Chiprut, B. Brazowski *et al.*, "LTX-Video: Realtime video latent diffusion," *CoRR*, vol. abs/2501.00103, 2025.

[206] P. Ling, J. Bu, P. Zhang *et al.*, "Motionclone: Training-free motion cloning for controllable video generation," in *ICLR*. OpenReview.net, 2025.

[207] G. Lin, J. Jiang, J. Yang *et al.*, "OmniHuman-1: Rethinking the scaling-up of one-stage conditioned human animation models," in *ICCV*. Computer Vision Foundation / IEEE, 2025.

[208] Y. Jin, Z. Sun, N. Li *et al.*, "Pyramidal flow matching for efficient video generative modeling," in *ICLR*. OpenReview.net, 2025.

[209] K. Namekata, S. Bahmani, Z. Wu, Y. Kant, I. Gilitschenski, and D. B. Lindell, "SG-I2V: self-guided trajectory control in image-to-video generation," in *ICLR*. OpenReview.net, 2025.

[210] F. Liu, H. Fu, X. Wang *et al.*, "Sketchvideo: Sketch-based video generation and editing," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 23 379–23 390.

[211] J. Bai, M. Xia, X. Wang *et al.*, "Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints," in *ICLR*. OpenReview.net, 2025.

[212] Z. Xiao, W. Ouyang, Y. Zhou *et al.*, "Trajectory attention for fine-grained video motion control," in *ICLR*. OpenReview.net, 2025.

[213] J. Zhang, K. Zheng, K. Jiang *et al.*, "TurboDiffusion: Accelerating video diffusion models by 100-200 times," *CoRR*, vol. abs/2512.16093, 2025.

[214] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu, "VACE: all-in-one video creation and editing," *CoRR*, vol. abs/2503.07598, 2025.

[215] X. Yang, L. Zhu, H. Fan, and Y. Yang, "Videograin: Modulating space-time attention for multi-grained video editing," in *ICLR*. OpenReview.net, 2025.

[216] E. Corona, A. Zanfir, E. G. Bazavan, N. Kolotouros, T. Alldieck, and C. Sminchisescu, "VLOGGER: Multimodal diffusion for embodied avatar synthesis," in *CVPR*. Computer Vision Foundation / IEEE, 2025.

[217] X. Ma, Y. Wang, G. Jia *et al.*, "Cinemo: Consistent and controllable image animation with motion diffusion models," *CoRR*, vol. abs/2407.15642, 2024.

[218] H. Ouyang, Q. Wang, Y. Xiao *et al.*, "Codef: Content deformation fields for temporally consistent video processing," in *CVPR*. IEEE, 2024, pp. 8089–8099.

[219] P. Ruan, P. Wang, D. Saxena, J. Cao, and Y. Shi, "Enhancing motion in text-to-video generation with decomposed encoding and conditioning," in *NeurIPS*, A. Globersons, L. Mackey, D. Belgrave *et al.*, Eds., 2024.

[220] T. Wu, C. Si, Y. Jiang, Z. Huang, and Z. Liu, "FreeInit: Bridging initialization gap in video diffusion models," in *ECCV*. Springer, 2024.

[221] H. Qiu, M. Xia, Y. Zhang *et al.*, "FreeNoise: Tuning-free longer video diffusion via noise rescheduling," in *ICLR*. OpenReview.net, 2024.

[222] J. Guo, D. Zhang, X. Liu *et al.*, "LivePortrait: Efficient portrait animation with stitching and retargeting control," *CoRR*, vol. abs/2407.03168, 2024.

[223] Z. Xu, J. Zhang, J. H. Liew *et al.*, "MagicAnimate: Temporally consistent human image animation using diffusion model," in *CVPR*. Computer Vision Foundation / IEEE, 2024.

[224] Y. Zhang, J. Gu, L. Wang *et al.*, "MimicMotion: High-quality human motion video generation with confidence-aware pose guidance," *CoRR*, vol. abs/2406.19680, 2024.

[225] M. Niu, X. Cun, X. Wang, Y. Zhang, Y. Shan, and Y. Zheng, "Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model," in *ECCV*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Springer, 2024, pp. 111–128.

[226] W. Chai, E. Song, Y. Zhang *et al.*, "MovieChat+: Question-aware sparse memory for long video question answering," *CoRR*, vol. abs/2404.17176, 2024.

[227] E. Song, W. Chai, G. Wang *et al.*, "MovieChat: From dense token to sparse memory for long video understanding," in *CVPR*. Computer Vision Foundation / IEEE, 2024.

[228] Z. Tong, C. Li, Z. Chen, B. Wu, and W. Zhou, "MusePose: A pose-driven image-to-video framework for virtual human generation," *CoRR*, 2024.

[229] Q. Yang, H. Chen, Y. Zhang *et al.*, "Noise calibration: Plug-and-play content-preserving video enhancement using pre-trained video diffusion models," in *ECCV*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Springer, 2024, pp. 307–326.

[230] X. Chen, Y. Wang, L. Zhang *et al.*, "SEINE: Short-to-long video diffusion model for generative transition and prediction," in *ICLR*. OpenReview.net, 2024.

[231] H. Chefer, S. Zada, R. Paiss *et al.*, "Still-moving: Customized video generation without customized video data," *ACM Trans. Graph.*, vol. 43, no. 6, pp. 244:1–244:11, 2024.

[232] Y. Jiang, T. Wu, S. Yang *et al.*, "VideoBooth: Diffusion-based video generation with image prompts," in *CVPR*. Computer Vision Foundation / IEEE, 2024.

[233] G. Kim, H. Shim, H. Kim, Y. Choi, J. Kim, and E. Yang, "Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding," in *CVPR*. IEEE, 2023, pp. 6091–6100.

[234] H. A. Rasheed, M. U. Khattak, M. Maaz, S. H. Khan, and F. S. Khan, "Fine-tuned CLIP models are efficient video learners," in *CVPR*. IEEE, 2023, pp. 6545–6554.

[235] M. Sun, W. Wang, Z. Qin, J. Sun, S. Chen, and J. Liu, "GLOBER: coherent non-autoregressive video generation via global guided video decoder," in *NeurIPS*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[236] X. Li, H. He, Y. Yang *et al.*, "Improving video instance segmentation via temporal pyramid routing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6594–6601, 2023.

[237] Y. Wang, X. Chen, X. Ma *et al.*, "LaVie: High-quality video generation with cascaded latent diffusion models," *CoRR*, vol. abs/2309.15103, 2023.

[238] S. Ge, S. Nah, G. Liu *et al.*, "Preserve your own correlation: A noise prior for video diffusion models," in *ICCV*. Computer Vision Foundation / IEEE, 2023.

[239] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, "Rerender A video: Zero-shot text-guided video-to-video translation," in *SIGGRAPH Asia*. ACM, 2023.

[240] D. J. Zhang, J. Z. Wu, J. Liu *et al.*, "Show-1: Marrying pixel and latent diffusion models for text-to-video generation," *CoRR*, vol. abs/2309.15818, 2023.
[241] W. Chai, X. Guo, G. Wang, and Y. Lu, "Stablevideo: Text-driven consistency-aware diffusion video editing," in *ICCV*. IEEE, 2023, pp. 22 983–22 993.
[242] Y. Chen, D. Chen, R. Liu, H. Li, and W. Peng, "Video action recognition with attentive semantic units," in *ICCV*. IEEE, 2023, pp. 10 136–10 146.
[243] X. Wang, S. Zhang, H. Zhang *et al.*, "VideoLCM: Video latent consistency model," *CoRR*, vol. abs/2312.09109, 2023.
[244] B. Ni, H. Peng, M. Chen *et al.*, "Expanding language-image pretrained models for general video recognition," in *ECCV*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Springer, 2022, pp. 1–18.
[245] Z. Lin, S. Geng, R. Zhang *et al.*, "Frozen CLIP models are efficient video learners," in *ECCV*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Springer, 2022, pp. 388–404.
[246] H. Yuan, X. Li, Y. Yang *et al.*, "Polyphonicformer: Unified query learning for depth-aware video panoptic segmentation," in *ECCV*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Springer, 2022, pp. 582–599.
[247] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *ECCV*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Springer, 2022, pp. 105–124.
[248] Z. Xu, W. Yang, W. Zhang, X. Tan, H. Huang, and L. Huang, "Segment as points for efficient and effective online multi-object tracking and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6424–6437, 2022.
[249] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2live: Text-driven layered image and video editing," in *ECCV*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Springer, 2022, pp. 707–723.
[250] X. Li, W. Zhang, J. Pang *et al.*, "Video k-net: A simple, strong, and unified baseline for video segmentation," in *CVPR*. IEEE, 2022, pp. 18 825–18 835.
[251] S. Yan, T. Zhu, Z. Wang *et al.*, "Video-text modeling with zero-shot transfer from contrastive captioners," *CoRR*, vol. abs/2212.04979, 2022.
[252] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *CoRR*, vol. abs/2109.08472, 2021.
[253] Y. Fu, L. Yang, D. Liu, T. S. Huang, and H. Shi, "Compfeat: Comprehensive feature aggregation for video instance segmentation," in *AAAI*. AAAI Press, 2021, pp. 1361–1369.
[254] S. Yang, Y. Fang, X. Wang *et al.*, "Crossover learning for fast online video instance segmentation," in *ICCV*. IEEE, 2021, pp. 8023–8032.
[255] Y. Wang, Z. Xu, X. Wang *et al.*, "End-to-end video instance segmentation with transformers," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 8741–8750.
[256] M. Weber, J. Xie, M. D. Collins *et al.*, "STEP: segmenting and tracking every pixel," in *NeurIPS Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., 2021.
[257] S. Qiao, Y. Zhu, H. Adam, A. L. Yuille, and L. Chen, "Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 3997–4008.
[258] G. Bertasius and L. Torresani, "Classifying, segmenting, and tracking object instances in video with mask propagation," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 9736–9745.
[259] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Sipmask: Spatial information preservation for fast image and video instance segmentation," in *ECCV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Springer, 2020, pp. 1–18.
[260] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *ECCV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Springer, 2020, pp. 107–122.
[261] C. Lin, Y. Hung, R. Feris, and L. He, "Video instance segmentation tracking with a modified VAE architecture," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 13 144–13 154.
[262] D. Kim, S. Woo, J. Lee, and I. S. Kweon, "Video panoptic segmentation," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 9856–9865.
[263] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L. Chen, "FEELVOS: fast end-to-end embedding learning for video object segmentation," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 9481–9490.
[264] P. Voigtlaender, M. Krause, A. Osep *et al.*, "MOTS: multi-object tracking and segmentation," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 7942–7951.
[265] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 4282–4291.
[266] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," in *ICCV*. IEEE, 2019, pp. 941–951.
[267] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *ICCV*. IEEE, 2019, pp. 5187–5196.
[268] S. W. Oh, J. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *ICCV*. IEEE, 2019, pp. 9225–9234.
[269] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *ECCV*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer, 2018, pp. 103–119.
[270] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, "Efficient video object segmentation via network modulation," in *CVPR*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6499–6507.
[271] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *CVPR*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 8971–8980.
[272] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *CVPR*. IEEE Computer Society, 2017, pp. 4141–4150.
[273] P. Tokmakov, K. Alahari, and C. Schmid, "Learning motion patterns in videos," in *CVPR*. IEEE Computer Society, 2017, pp. 531–539.
[274] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *CVPR*. IEEE Computer Society, 2017, pp. 3491–3500.
[275] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *ICCV*. IEEE Computer Society, 2017, pp. 4491–4500.
[276] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork convnets for video semantic segmentation," in *ECCV Workshops*, ser. Lecture Notes in Computer Science, G. Hua and H. Jégou, Eds., 2016, pp. 852–868.
[277] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *ECCV Workshops*, ser. Lecture Notes in Computer Science, G. Hua and H. Jégou, Eds., 2016, pp. 850–865.
[278] R. Velioglu, P. Bevandic, R. Chan, and B. Hammer, "Enhancing person-to-person virtual try-on with multi-garment virtual try-off," *CoRR*, vol. abs/2504.13078, 2025.
[279] J. W. Hong, T. Ton, T. X. Pham, G. Koo, S. Yoon, and C. D. Yoo, "ITA-MDT: image-timestep-adaptive masked diffusion transformer framework for image-based virtual try-on," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 28 284–28 294.
[280] Y. Liang, X. Hu, B. Jiang *et al.*, "Vton-handfit: Virtual try-on for arbitrary hand pose guided by hand priors embedding," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 22 616–22 626.
[281] K. Sun, J. Cao, Q. Wang *et al.*, "Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person," *CoRR*, vol. abs/2407.16224, 2024.

[282] "LVBench-C: A benchmark for long-horizon video consistency," 2026.
[283] M. Li, C. Xie, Y. Wu, L. Zhang, and M. Wang, "FiVE: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models," in *ICCV*. Computer Vision Foundation / IEEE, 2025.
[284] H. Wu, D. Li, B. Chen, and J. Li, "LongVideoBench: A benchmark for long-context interleaved video-language understanding," in *NeurIPS*. Curran Associates, Inc., 2024.
[285] J. Miao, X. Wang, Y. Wu *et al.*, "Large-scale video panoptic segmentation in the wild: A benchmark," in *CVPR*. IEEE, 2022, pp. 21 001–21 011.
[286] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *ICIP*. IEEE, 2017, pp. 3645–3649.
[287] R. Goyal, S. E. Kahou, V. Michalski *et al.*, "The "something something" video database for learning and evaluating visual common sense," in *ICCV*. Computer Vision Foundation / IEEE, 2017.
[288] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. H. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*. IEEE Computer Society, 2016, pp. 724–732.
[289] Y. Li, Y. Wang, Y. Zhu *et al.*, "BranchGRPO: Stable reinforcement learning for diffusion models via branch-based group relative policy optimization," in *ICLR*. OpenReview.net, 2026.
[290] Y. Sun, J. Wu, Y. Cao *et al.*, "Swiftvideo: A unified framework for few-step video generation through trajectory-distribution alignment," in *AAAI*, S. Koenig, C. Jenkins, and M. E. Taylor, Eds. AAAI Press, 2026, pp. 9233–9241.
[291] J. Bai, X. Yu, M. Xu *et al.*, "Towards better optimization for listwise preference in diffusion models," in *ICLR*. OpenReview.net, 2026.
[292] W. Luo, "Diff-instruct++: Training one-step text-to-image generator model to align with human preferences," *Trans. Mach. Learn. Res.*, vol. 2025, 2025.
[293] F. Wang, Y. Shui, J. Piao, K. Sun, and H. Li, "Diffusion-NPO: Negative preference optimization for better preference aligned generation of diffusion models," in *ICLR*. OpenReview.net, 2025.
[294] M. Fu, G. Wang, T. Cui *et al.*, "Diffusion-SDPO: Safeguarded direct preference optimization for diffusion models," *CoRR*, vol. abs/2511.03317, 2025.
[295] H. Zhu, T. Xiao, and V. G. Honavar, "DSPO: Direct score preference optimization for diffusion model alignment," in *ICLR*. OpenReview.net, 2025.
[296] H. Zhao, H. Chen, Y. Guo *et al.*, "Fine-tuning diffusion generative models via rich preference optimization," *CoRR*, vol. abs/2503.11720, 2025.
[297] Z. Song, S. Dai, B. Keller, and B. Khailany, "Galaxydit: Efficient video generation with guidance alignment and adaptive proxy in diffusion transformers," *CoRR*, vol. abs/2512.03451, 2025.
[298] J. Liu, G. Liu, J. Liang *et al.*, "Improving video generation with human feedback," in *NeurIPS*. Curran Associates, Inc., 2025.
[299] Q. Yang, Z. Wang, Y. Dong, Z. Lian, and H. Su, "McSc: Motion-corrective preference alignment for video generation with self-critic hierarchical reasoning," *CoRR*, vol. abs/2511.22974, 2025.
[300] Y. Cao, Y. Miao, X.-S. Gao, and Y. Dong, "Red-teaming text-to-image systems by rule-based preference modeling," 2025.
[301] S. Karthik, H. Coskun, Z. Akata, S. Tulyakov, J. Ren, and A. Kag, "Scalable ranked preference optimization for text-to-image generation," in *ICCV*. Computer Vision Foundation / IEEE, 2025, pp. 18 399–18 410.
[302] J. Li, Q. Long, J. Zheng *et al.*, "T2V-Turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design," in *ICLR*. OpenReview.net, 2025.
[303] R. Liu, H. Wu, Z. Zheng *et al.*, "VideoDPO: Omni-preference alignment for video diffusion generation," in *CVPR*. Computer Vision Foundation / IEEE, 2025.
[304] Y. Liu, Y. Liu, J. Pan *et al.*, "Super-resolving real-world image illumination enhancement: A new dataset and A conditional diffusion model," *CoRR*, vol. abs/2410.12961, 2024.
[305] J. Li, W. Feng, T. Fu *et al.*, "T2V-Turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback," in *NeurIPS*. Curran Associates, Inc., 2024.
[306] K. Yang, J. Tao, J. Lyu *et al.*, "Using human feedback to fine-tune diffusion models without any reward model," in *CVPR*. Computer Vision Foundation / IEEE, 2024.
[307] Y. He, Y. Xing, Z. Rao *et al.*, "VideoTuna: A powerful toolkit for video generation with model fine-tuning and post-training," GitHub repository, 2024, https://github.com/VideoVerses/VideoTuna-dev.
[308] Y. Fan, O. Watkins, Y. Du *et al.*, "DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models," in *NeurIPS*. Curran Associates, Inc., 2023.
[309] X. Cheng, J. Zhang, Z. Huang, Y. Wu, and X. Huang, "Compensation-free machine unlearning in text-to-image diffusion models by eliminating the mutual information," *CoRR*, vol. abs/2603.00992, 2026.
[310] R. Wang, J. Fang, J. Li *et al.*, "ACE: Concept editing in diffusion models without performance degradation," in *ACM Multimedia*. ACM, 2025.
[311] X. Meng, Y. Dong, N. Yu, L. Wang, Z. Li, and S. Guo, "Safe-control: A safety patch for mitigating unsafe content in text-to-image generation models," *CoRR*, vol. abs/2508.21099, 2025.
[312] R. Gandikota, J. Materzyńska, T. Zhou, A. Torralba, and D. Bau, "Concept sliders: LoRA adaptors for precise control in diffusion models," in *ECCV*. Springer, 2024.
[313] Y. Zhang, X. Chen, J. Jia *et al.*, "Defensive unlearning with adversarial training for robust concept erasure in diffusion models," in *NeurIPS*. Curran Associates, Inc., 2024.
[314] M. Lyu, Y. Yang, H. Hong *et al.*, "One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications," in *CVPR*. IEEE, 2024, pp. 7559–7568.
[315] C. Gong, K. Chen, Z. Wei, J. Chen, and Y. Jiang, "Reliable and efficient concept erasure of text-to-image diffusion models," in *ECCV*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Springer, 2024, pp. 73–88.
[316] Y. Tsai, C. Hsu, C. Xie *et al.*, "Ring-a-bell! how reliable are concept removal methods for diffusion models?" in *ICLR*. OpenReview.net, 2024.
[317] X. Li, Y. Yang, J. Deng *et al.*, "Safegen: Mitigating unsafe content generation in text-to-image models," *CoRR*, vol. abs/2404.06666, 2024.
[318] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu, "SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation," in *ICLR*. OpenReview.net, 2024.
[319] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "SneakyPrompt: Jailbreaking text-to-image generative models," in *IEEE S&P*. IEEE, 2024.
[320] C. Huang, K. Chang, C. Tsai, Y. Lai, and Y. F. Wang, "Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers," *CoRR*, vol. abs/2311.17717, 2023.
[321] A. Heng and H. Soh, "Selective amnesia: A continual learning approach to forgetting in deep generative models," *CoRR*, vol. abs/2305.10120, 2023.
[322] S. Kim, S. Jung, B. Kim, M. Choi, J. Shin, and J. Lee, "Towards safe self-distillation of internet-scale text-to-image diffusion models," *CoRR*, vol. abs/2307.05977, 2023.
[323] C. Zhang, T. Zhang, L. Wang, R. Chen, W. Li, and A. Liu, "T2i-riskyprompt: A benchmark for safety evaluation, attack, and defense on text-to-image model," *CoRR*, vol. abs/2510.22300, 2025.

[324] Y. Zhang, C. Fan, Y. Zhang *et al.*, "UnlearnCanvas: A stylized image dataset to benchmark machine unlearning for diffusion models," in *NeurIPS*. Curran Associates, Inc., 2024.
[325] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos, "Do generative video models understand physical principles?" *arXiv preprint arXiv:2501.09038*, 2025.
[326] B. Kang, Y. Yue, R. Lu *et al.*, "How far is video generation from world model: A physical law perspective," in *ICML*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu *et al.*, Eds. PMLR / OpenReview.net, 2025.
[327] X. Yang, L. Wen, Y. Ma *et al.*, "Drivearena: A closed-loop generative simulation platform for autonomous driving," *CoRR*, vol. abs/2408.00415, 2024.
[328] C. Zhang, D. Cherniavskii, A. Zadaianchuk *et al.*, "Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments," *CoRR*, vol. abs/2504.02918, 2025.
[329] D. Li, Y. Fang, Y. Chen *et al.*, "Worldmodelbench: Judging video generation models as world models," *CoRR*, vol. abs/2502.20694, 2025.
[330] H. Duan, H.-X. Yu, S. Chen, L. Fei-Fei, and J. Wu, "Worldscore: A unified evaluation benchmark for world generation," *arXiv preprint arXiv:2504.00983*, 2025.
[331] Y. Qin, Z. Shi, J. Yu *et al.*, "Worldsimbench: Towards video generation models as world simulators," *CoRR*, vol. abs/2410.18072, 2024.
[332] R. Riochet, M. Y. Castro, M. Bernard *et al.*, "IntPhys: A framework and benchmark for visual intuitive physics reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3140–3153, 2021.
[333] J. Xing, M. Xia, Y. Zhang *et al.*, "DynamiCrafter: Animating open-domain images with video diffusion priors," in *ECCV*. Springer, 2024.
[334] Y. Yang, R. Gao, X. Yang, J. Zhong, and Q. Xu, "Guardt2i: Defending text-to-image models from adversarial prompts," *CoRR*, vol. abs/2403.01446, 2024.
[335] K. Yi, C. Gan, Y. Li *et al.*, "CLEVRER: Collision events for video representation and reasoning," in *ICLR*. OpenReview.net, 2020.