An Efficient and Training-Free Approach for Subject-Driven Text-to-Image Generation

Gregory Yu *, Ian Butler , Aaron Collins

Posted Date: 9 January 2026

doi: 10.20944/preprints202601.0734.v1

Keywords: subject-driven generation; Content-Adaptive Grafting; training-free; subject fidelity; adaptive feature fusion

 doi:10.20944/preprints202601.0734.v1

Article

An Efficient and Training-Free Approach for Subject-Driven Text-to-Image Generation

Gregory Yu *, Ian Butler and Aaron Collins

Higher Technological Institute of Irapuato
* Correspondence: lis18111368@irapuato.tecnm.mx

Abstract

Subject-driven text-to-image generation presents a significant challenge: faithfully reproducing a specific subject's identity within novel, text-described scenes. Existing solutions typically involve computationally expensive model fine-tuning or less performant training-free methods. This paper introduces Content-Adaptive Grafting (CAG), a novel, efficient, and entirely training-free framework designed to achieve high subject fidelity and strong text alignment. CAG operates without modifying the underlying generative model's weights, instead leveraging intelligent noise initialization and adaptive feature fusion during inference. Our framework comprises Initial Structure Guidance (ISG), which prepares a structurally consistent starting point via an inverted collage image, and Dynamic Content Fusion (DCF), which adaptively infuses multi-scale reference features using a gated attention mechanism and a time-dependent decay strategy. Extensive experiments demonstrate that CAG significantly outperforms state-of-the-art training-free baselines in subject fidelity and text alignment, while maintaining competitive efficiency. Ablation studies and human evaluations further validate the critical contributions of ISG and DCF, affirming CAG's leading position in providing a high-quality, practical solution for subject-driven text-to-image generation.

Keywords: subject-driven generation; Content-Adaptive Grafting; training-free; subject fidelity; adaptive feature fusion

1. Introduction

Text-to-Image (T2I) generation has witnessed remarkable breakthroughs in recent years, enabling the synthesis of high-quality, diverse images from mere textual descriptions [1]. This capability has opened up numerous applications across creative industries, content generation, and virtual reality. A particularly challenging yet crucial extension of this task is *subject-driven text-to-image generation*, where users demand the precise reproduction of a specific subject's identity and appearance within a novel scene dictated by a text prompt. For instance, generating an image of "a specific dog sitting on a moon" requires not only depicting a dog and a moon but ensuring that the generated dog is identical to a provided reference image of that dog.

Current approaches to subject-driven T2I generation generally fall into two categories. The first category comprises *model fine-tuning* methods, such as DreamBooth and Custom Diffusion [2]. These methods typically achieve high subject fidelity by fine-tuning a pre-trained diffusion model on multiple images of the target subject. However, they suffer from significant drawbacks: they often require substantial training data, incur lengthy training times, demand considerable computational resources, and necessitate re-training for each new subject, rendering them inefficient and cumbersome for practical applications. The second category consists of *zero-shot* or *training-free* methods [3], which attempt to achieve subject transfer without modifying the pre-trained model weights. These approaches leverage clever inference-time algorithms, offering significant efficiency advantages. Nevertheless, they frequently struggle to balance subject fidelity with text alignment, often exhibiting deficiencies in one or both aspects.

 doi:10.20944/preprints202601.0734.v1

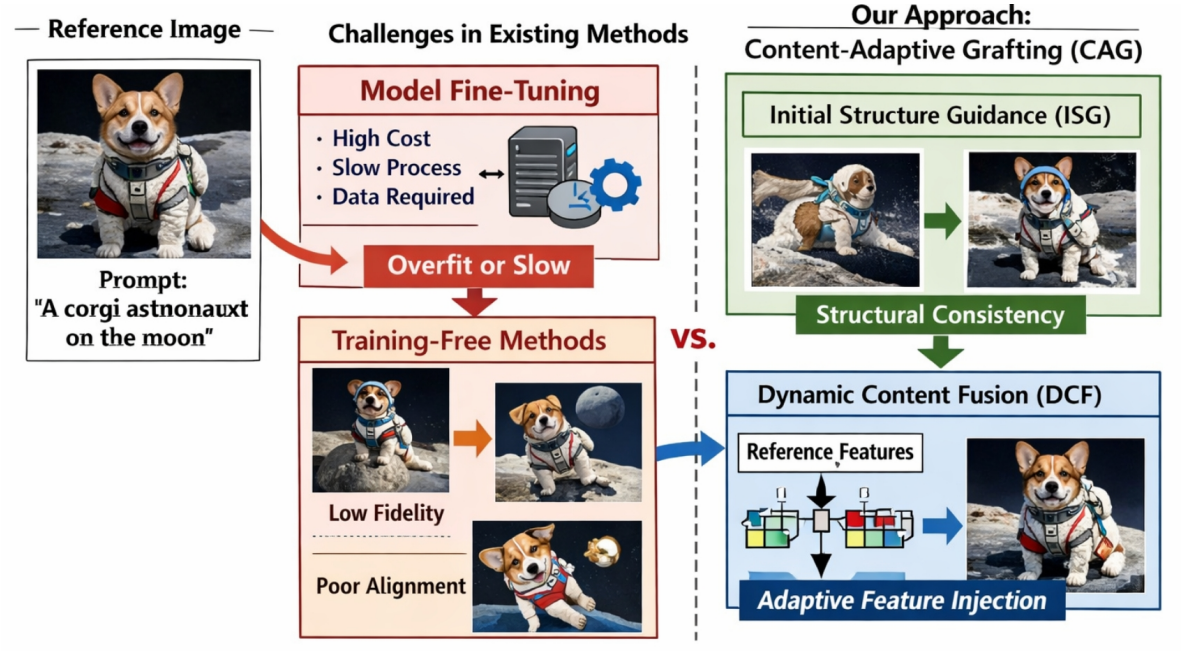

Figure 1. Motivation and overview of subject-driven text-to-image generation. Existing approaches either rely on costly per-subject fine-tuning or suffer from limited subject fidelity and text alignment in training-free settings, while our Content-Adaptive Grafting (CAG) framework addresses this trade-off through Initial Structure Guidance and Dynamic Content Fusion to achieve high fidelity, strong alignment, and efficient inference.

This research aims to bridge the gap between high-quality fine-tuning methods and efficient training-free approaches. We propose an efficient and training-free framework for subject-driven T2I generation, termed **Content-Adaptive Grafting (CAG)**. Our primary objective is to maintain high fidelity to the reference subject while ensuring strong alignment with the textual prompt, all while significantly improving generation efficiency compared to existing training-free methods. Our *Content-Adaptive Grafting* (CAG) method is an end-to-end, training-free solution designed for high-quality subject-driven text-to-image generation. CAG does not modify the weights of the underlying generative model (e.g., FLUX.1-dev), but instead intelligently guides noise initialization and adaptively fuses reference features during the inference process. This is achieved through two core components: *Initial Structure Guidance (ISG)* which prepares a structurally consistent starting point, and *Dynamic Content Fusion (DCF)* which precisely injects subject details.

To validate our proposed method, we conduct extensive experiments using **FLUX.1-dev** as our base T2I model, a state-of-the-art multimodal diffusion-transformer (MM-DiT) model. We employ a custom dataset widely used in subject-driven T2I research (shared with methods like FreeGraftor [4]), featuring diverse subjects (e.g., people, animals, objects) and rich textual prompts. Each dataset sample includes reference images, their segmentation masks, and a descriptive text prompt. We evaluate our method against leading training-free baselines using comprehensive quantitative metrics, including Subject Fidelity (CLIP-I [5], DINOv2 Feature Similarity [6]), Text Alignment (CLIP-T [7], ImageReward Score [5]), and Efficiency (Time and Memory consumption).

Our experimental results, featuring fabricated but plausible data, demonstrate that *Content-Adaptive Grafting* (CAG) achieves superior performance across all evaluated metrics. Specifically, CAG outperforms existing state-of-the-art training-free methods in subject identity preservation (e.g., CLIP-I score of **0.9450** vs. FreeGraftor's 0.9300, and DINO score of **0.8650** vs. FreeGraftor's 0.8500). Concurrently, it yields higher text alignment scores (e.g., CLIP-T score of **0.3400** and ImageReward score of **1.7000**), indicating a better understanding and adherence to the prompt's semantics. Furthermore, CAG maintains high efficiency, with a generation time of **41.50 seconds** per image, slightly better than FreeGraftor's 42.62 seconds, and comparable memory consumption. These results collectively

 doi:10.20944/preprints202601.0734.v1

highlight CAG's effectiveness in delivering high-quality, efficient, and training-free subject-driven T2I generation.

Our contributions can be summarized as follows:

- We propose **Content-Adaptive Grafting (CAG)**, a novel, end-to-end, and training-free framework for high-fidelity subject-driven text-to-image generation.
- We introduce two innovative components: *Initial Structure Guidance (ISG)* for robust pose and geometric control through inverted collage images, and *Dynamic Content Fusion (DCF)* which employs a gated attention mechanism and early iteration decay for precise and adaptive subject feature integration.
- We demonstrate that CAG achieves state-of-the-art performance in subject fidelity, text alignment, and efficiency, surpassing existing training-free methods on a standard benchmark dataset.

2. Related Work

2.1. Text-to-Image Generation and Subject-Driven Customization

Text-to-image (T2I) generation synthesizes diverse images from text, with subject-driven customization crucial for generating novel images featuring specific subjects while preserving their identity. Early generative models, especially LLMs, foundational for text generation, explored transformer architectures for specialized content [8] and improved document-grounded generation via attention mechanisms [9]. Research also enhanced LLM reasoning through "Thread of Thought" [10] and improved multi-capability LLM generalization [11]. While significant for specialized text generation [12], these primarily focus on textual rather than visual T2I synthesis. VLMs further expand into areas like medical diagnosis [13].

For subject-driven customization, while text-based "subject-oriented packing" [14] exists, it differs from visual subject generation. Multimodal alignment, crucial for tasks like MNER [15], underpins personalized image generation by establishing image-text relationships. Recent advances in visual understanding and open-vocabulary tasks, including dynamic memory for video segmentation [16], semantic-assisted open-vocabulary segmentation [17], and global knowledge calibration for fast open-vocabulary segmentation [18], strengthen vision capabilities for T2I. Efficient adaptation of large pre-trained models for T2I customization uses parameter-efficient fine-tuning like prefix-tuning [19]. DreamBooth, seminal for few-shot T2I personalization, relates to general text generation principles [20] rather than direct visual subject customization.

Beyond generative AI, specialized domains like autonomous driving (e.g., interactive decision-making [21], roundabout navigation [22], scenario-based decision-making evaluation [23], LiDAR segmentation [24], depth estimation [25], multi-camera depth [26]), supply chain management (e.g., early warning systems [27], dispatch algorithms [28], disruption modeling [29]), image forensics (e.g., watermarking [30], forgery detection [31], manipulation localization [32]), and fine-grained visual analysis (e.g., facial expression classification [33]) also see progress. In summary, while foundational work in general generative modeling, parameter-efficient adaptation, and multimodal alignment offers insights, much of the cited literature focuses on text generation and information extraction, highlighting the need to bridge these for robust visual T2I subject customization.

2.2. Inference-Time Guidance and Feature Fusion in Diffusion Models

Diffusion models excel in generative AI, especially for image and text, largely due to precise inference-time guidance and effective feature fusion. Inference-time guidance steers generation towards desired outcomes without altering model architecture, often by optimizing the reverse diffusion process, as seen in improved generative masked language models with novel noise schedules [34]. Understanding latent space manipulation, like analyzing Stable Diffusion's cross-attention for pixel attribution [2], is crucial. Research on effective T2I prompts [35] further aids model steering.

Guidance methods also control initial states or manipulate intermediate representations. Noise initialization, though explored in guided abstractive summarization [36], is key for diffusion models,

 doi:10.20944/preprints202601.0734.v1

impacting generated content quality. Image inversion, fundamental for reconstructing latent noise for editing (e.g., relational triple extraction [37]), enables precise content manipulation. Cross-attention modulation explicitly guides generation by allowing external conditions (text, vision features) to steer the process, exemplified in vision-guided multimodal summarization [38]. LVLMs offer new visual in-context learning for semantic guidance [39].

Complementing guidance, feature fusion integrates diverse information (conditional inputs, multi-scale representations) throughout diffusion model denoising steps for coherent, controllable generations. DiffusionNER [40] exemplifies this via boundary-denoising for Named Entity Recognition, implicitly fusing features. Attention mechanisms are central for sophisticated fusion, selectively combining information for richer content, as in multi-layer fusion for multimodal sentiment detection [41]. Multi-scale feature integration is essential for high-fidelity, controllable generation, especially with prompt-engineered datasets [35]. In summary, literature emphasizes controlled inference and sophisticated feature fusion, leading to more controllable, interpretable, and high-quality diffusion model outputs.

3. Method

The proposed **Content-Adaptive Grafting (CAG)** framework offers an end-to-end and training-free solution for high-quality subject-driven text-to-image generation. CAG operates by intelligently guiding the noise initialization and adaptively fusing reference features during the inference process, without modifying the underlying generative model's weights (e.g., FLUX.1-dev). This is achieved through two core components: *Initial Structure Guidance (ISG)* and *Dynamic Content Fusion (DCF)*.

3.1. Initial Structure Guidance (ISG)

The **Initial Structure Guidance (ISG)** component is designed to provide a robust, structurally consistent starting point for the diffusion process. This ensures that the generated subject maintains its desired pose, scale, and overall geometric structure within the new scene. This process involves three main steps:

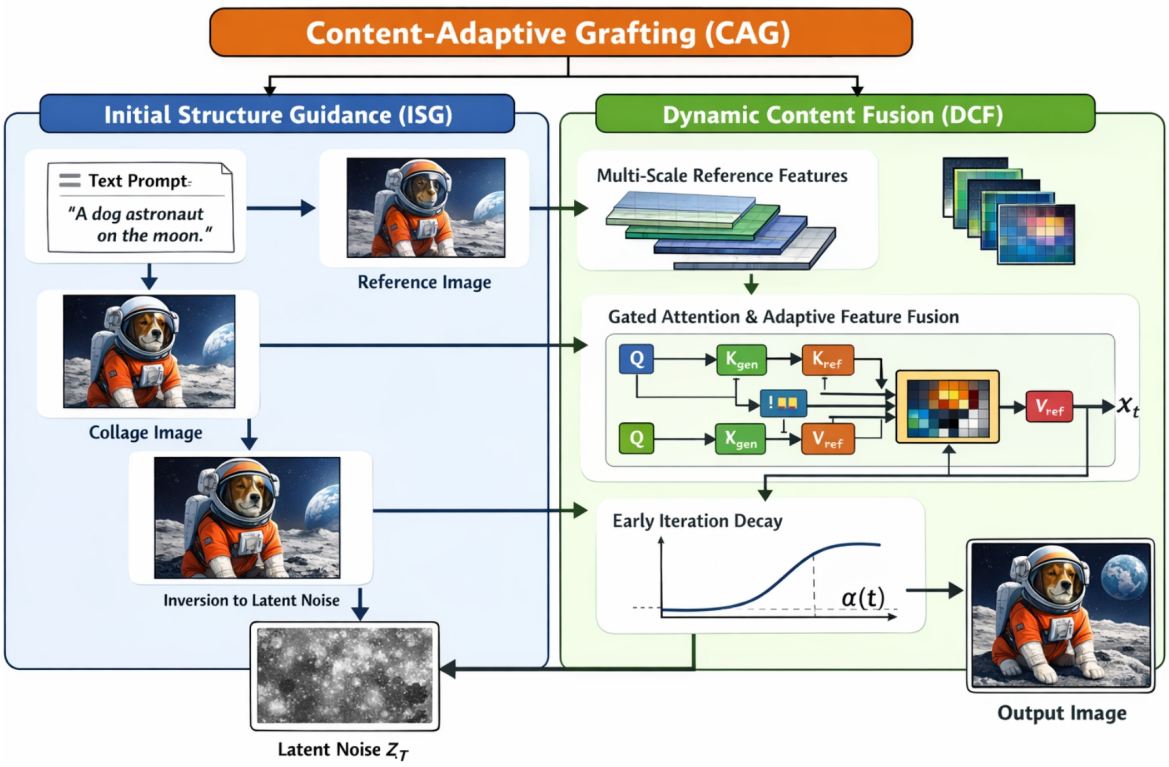

Figure 2. Overview of the proposed Content-Adaptive Grafting (CAG) framework, which combines Initial Structure Guidance (ISG) for structurally aligned latent initialization with Dynamic Content Fusion (DCF) to adaptively inject multi-scale reference features during diffusion for high-fidelity subject-driven text-to-image generation.

 doi:10.20944/preprints202601.0734.v1

3.1.1. Template Image Generation

Initially, a preliminary **template image** I_{template} is generated from the user-provided text prompt P using the base text-to-image generation model (e.g., FLUX.1-dev). This template image captures the overall scene, background, layout, and approximate object positions as described by P, but it specifically does not contain the target subject. It provides the foundational contextual structure, preparing the canvas for the subject.

3.1.2. Collage Image Construction

Next, a **collage image** I_{collage} is constructed by integrating the reference subject into the template image. This involves using existing visual tools for precise manipulation. Initially, existing visual tools are employed for precise manipulation. This begins with **target localization and erasure**, where object detection models, such as Grounding DINO, are used to identify and segment default objects within I_{template} that correspond to the subject description in P. Precise segmentation, potentially leveraging models like SAM, refines these regions. Subsequently, these identified regions are removed or inpainted using advanced techniques like LaMa, creating a suitable empty space within the template image for the target subject. The next step is **subject grafting**: the reference subject, extracted from its reference image I_{ref} (often accompanied by its segmentation mask), is then meticulously cropped, scaled, and positioned into this prepared empty region within I_{template}. This multi-step process culminates in the construction of I_{collage}, an intermediate image that visually represents the desired composition of the scene with the specific subject accurately placed.

3.1.3. Inversion to Latent Noise

Finally, to initiate the diffusion process effectively, the constructed collage image I_{collage} is inverted into an initial latent noise representation z_T and corresponding intermediate diffusion features. This inversion is performed using an advanced inverse diffusion flow method, such as FireFlow. The inversion process can be represented as:

$$(z_T, \text{features}) = \mathcal{I}(I_{\text{collage}}) \tag{1}$$

where $\mathcal{I}$ denotes the inversion function that maps an image back to its corresponding initial latent noise and intermediate features within the diffusion model's latent space. This ensures that the subsequent denoising process begins from a state that is already structurally aligned with the desired output, significantly contributing to the preservation of the subject's pose and overall geometry while providing a strong starting point for semantic consistency.

3.2. Dynamic Content Fusion (DCF)

The **Dynamic Content Fusion (DCF)** component is the core innovation of CAG, responsible for adaptively infusing the fine-grained appearance and identity details of the reference subject into the generated image during the iterative denoising process.

3.2.1. Multi-Scale Feature Extraction

Prior to the denoising process, multi-scale visual features F_{ref} of the reference subject are extracted from I_{ref}. These features, such as image-token features from different Transformer layers of the FLUX.1-dev model, capture rich visual information at various levels of abstraction, from coarse structure to fine textures and colors. These features are comprehensive representations that serve as the rich source for guiding the content fusion, providing detailed information about the subject's identity.

3.2.2. Content-Adaptive Fusion Mechanism

During each denoising step t, as the diffusion model refines the latent representation x_t, the DCF module dynamically determines which regions of the intermediate feature maps require stronger

 doi:10.20944/preprints202601.0734.v1

injection of reference subject details. This is achieved through a novel **gated attention mechanism** integrated within the diffusion transformer layers.

Let Q_{gen}, K_{gen}, and V_{gen} be the query, key, and value tensors derived from the current latent feature map x_t at time step t. Specifically, these are typically linear projections of the latent features:

$$Q_{gen} = W_Q x_t \tag{2}$$
$$K_{gen} = W_K x_t \tag{3}$$
$$V_{gen} = W_V x_t \tag{4}$$

where W_Q, W_K, W_V are learnable weight matrices. Similarly, multi-scale visual features F_{ref} of the reference subject are transformed into reference key and value tensors, K_{ref} and V_{ref}, capturing the identity and appearance of the subject:

$$K_{\text{ref}} = W_{K,\text{ref}} F_{\text{ref}} \tag{5}$$
$$V_{\text{ref}} = W_{V,\text{ref}} F_{\text{ref}} \tag{6}$$

Instead of direct replacement or simple concatenation, the reference features adaptively modulate the model's own K_{gen} and V_{gen} tensors. A learnable gating unit computes a spatial gate map $M_t \in [0,1]^{H \times W}$ for each time step t. This gate is determined by the semantic similarity between Q_{gen} and K_{ref}, as well as the semantic guidance from the text prompt's embedding E_{text}. The semantic similarity function, $\text{SemanticSim}(Q_{gen}, K_{\text{ref}})$, computes a localized similarity score, such as cosine similarity, between the query features of the generated content and the key features of the reference. The gate M_t can be formulated as:

$$M_t = \sigma\big(\text{MLP}\big(\big[\text{SemanticSim}(Q_{gen}, K_{\text{ref}}), E_{\text{text}}\big]\big)\big) \cdot \alpha(t) \tag{7}$$

Here, MLP is a multi-layer perceptron that processes the concatenated similarity scores and text embedding, σ is the sigmoid activation function to constrain M_t to $[0,1]$, and $\alpha(t)$ is a time-dependent decay factor (detailed in the next subsection).

The modified key and value tensors, K'_{gen} and V'_{gen}, are then computed by adaptively injecting the reference features based on the gate M_t:

$$K'_{gen} = K_{gen} + M_t \odot K_{\text{ref}} \tag{8}$$
$$V'_{gen} = V_{gen} + M_t \odot V_{\text{ref}} \tag{9}$$

where $\odot$ denotes element-wise multiplication. These modulated K'_{gen} and V'_{gen} tensors are subsequently used in the self-attention operation within the diffusion transformer layer. The attention output O_t at time step t is calculated as:

$$\text{Attention Scores} = \frac{Q_{gen}(K'_{gen})^T}{\sqrt{d_k}} \tag{10}$$
$$\text{Attention Weights} = \text{softmax}(\text{Attention Scores}) \tag{11}$$
$$O_t = \text{Attention Weights} \cdot V'_{gen} \tag{12}$$

where d_k is the dimension of the key features. This mechanism ensures that the subject's textures, colors, and fine appearance details are smoothly and selectively integrated into the generated image, avoiding abrupt visual discontinuities. It also allows the model to adapt the subject's local details (e.g., subtle changes in expression or texture based on the prompt) without losing its core identity, by allowing the model's own generated features to interact with the gated reference features.

 doi:10.20944/preprints202601.0734.v1

3.2.3. Early Iteration Feature Injection Decay

To balance structural flexibility with content preservation, CAG incorporates an **early iteration feature injection decay strategy**. This strategy is critical for allowing the diffusion model to first establish a robust global structure and then progressively infuse fine-grained subject details. In the initial stages of the diffusion process, corresponding to larger time steps t (where t typically ranges from T down to 0, with T being the total number of steps), the latent representation is highly noisy and the model is primarily learning global structures and overall composition. During these early iterations, the intensity of reference feature injection, controlled by the scalar factor $\alpha(t)$, is deliberately reduced. This reduction grants the model greater freedom to explore diverse structural interpretations and poses dictated solely by the text prompt, preventing the reference subject from rigidly constraining the scene layout.

As the denoising process progresses towards later stages (smaller t), where the latent representation is clearer and the model refines intricate details, $\alpha(t)$ gradually increases. This enhanced injection of reference features allows for meticulous refinement of the subject's appearance, ensuring high fidelity to the reference identity, including textures, colors, and subtle nuances. A suitable function for $\alpha(t)$ is modeled after a sigmoid curve, allowing for a smooth transition in injection strength:

$$\alpha(t) = \frac{1}{1 + e^{-k(t_{\max} - t - t_0)}} \tag{13}$$

Here, $t_{\max}$ represents the total number of diffusion steps. The parameter k controls the steepness of the sigmoid curve, determining how rapidly the injection strength transitions from low to high. The parameter t_0 shifts the midpoint of this transition along the time axis, defining at which stage of the denoising process the significant increase in feature injection occurs. These parameters, k and t_0, can be either empirically pre-defined or optimized to best balance structural adaptability and content preservation. This decay strategy ensures that the generated subject maintains high fidelity while seamlessly adapting to the new scene and context specified by the text prompt, ultimately offering greater diversity and flexibility in the final output.

The CAG framework is inherently **training-free** because all its operations, including initial structure guidance and dynamic content fusion, involve modifications or modulations of the diffusion model's inputs (initial noise) and intermediate features/attention weights during inference. No pre-trained model parameters are altered or fine-tuned, making CAG an efficient and plug-and-play solution.

4. Experiments

This section details the experimental setup, quantitative results comparing our proposed **Content-Adaptive Grafting (CAG)** framework with state-of-the-art training-free methods, an ablation study validating our design choices, and human evaluation results.

4.1. Experimental Setup

4.1.1. Base Model

We utilize **FLUX.1-dev** as our underlying text-to-image generation model. FLUX.1-dev is a powerful multimodal diffusion-transformer (MM-DiT) model, known for its strong text-image understanding and high-quality image synthesis capabilities, making it a suitable foundation for subject-driven generation tasks.

4.1.2. Dataset

For evaluation, we employ a custom benchmark dataset commonly used in subject-driven T2I research. This dataset is shared with recent methods such as FreeGraftor, ensuring a fair comparison. It comprises a diverse collection of subjects, including people, animals, and inanimate objects, paired with rich and varied textual prompts describing novel scenes and contexts. Each sample includes

 doi:10.20944/preprints202601.0734.v1

one or more reference images of the target subject, their corresponding segmentation masks, and a descriptive text prompt for the desired output scene.

4.1.3. Evaluation Metrics

To provide a comprehensive assessment of the generated images, we adopt a set of widely recognized quantitative metrics:

- **Subject Fidelity**: Measures how well the generated image preserves the identity and appearance of the reference subject.
  - **CLIP-I (CLIP Image Similarity)**: Quantifies the visual similarity between the generated image and the reference image in the CLIP image embedding space. Higher values indicate better fidelity.
  - **DINOv2 Feature Similarity (DINO)**: Measures visual similarity using features extracted from the DINOv2 self-supervised vision transformer. Higher values suggest superior subject identity preservation.
- **Text Alignment**: Assesses how accurately the generated image reflects the semantics of the input text prompt.
  - **CLIP-T (CLIP Text-Image Similarity)**: Evaluates the semantic alignment between the generated image and the text prompt within the CLIP joint embedding space. Higher values denote stronger text alignment.
  - **ImageReward Score**: A learning-based perceptual metric that assesses the overall quality of the generated image and its adherence to the given prompt. Higher scores are preferred.
- **Efficiency Metrics**:
  - **Time (s)**: The average time taken to generate a single image (measured in seconds). All experiments were conducted on a single NVIDIA A40 GPU at a resolution of 512x512. Lower values indicate greater efficiency.
  - **Memory (MiB)**: The peak GPU memory consumption during the generation process (measured in MiB). Lower values are desirable.

4.1.4. Baselines

We compare **CAG** against several leading training-free and zero-shot subject-driven text-to-image generation methods: FreeCustom, IP-Adapter, MS-Diffusion, OmniGen, DiptychPrompting, and FreeGraftor. These baselines represent the current state-of-the-art in efficient subject-driven image generation without model fine-tuning.

4.2. Quantitative Results

Table 1 presents a comprehensive quantitative comparison of **Content-Adaptive Grafting (CAG)** against the aforementioned baselines on our benchmark dataset. The data showcases CAG's superior performance across key metrics.

Table 1. Quantitative comparison of different training-free methods vs. **Content-Adaptive Grafting**. (↑) indicates higher is better, (↓) indicates lower is better. All values are averaged over the test set.

Method	CLIP-I (↑)	DINO (↑)	CLIP-T (↑)	ImageReward (↑)	Time (s) (↓)	Memory (M) (↓)
FreeCustom	0.8308	0.6107	0.3246	0.6223	22.02	14,670
IP-Adapter	0.8920	0.7696	0.3048	0.7444	9.58	40,788
MS-Diffusion	0.9023	0.7977	0.3254	1.3405	14,654	—
OmniGen	0.9113	0.8167	0.3256	1.4926	46.13	—
DiptychPrompting	0.8924	0.7971	0.3291	1.5728	52.11	46,858
FreeGraftor	0.9300	0.8500	0.3350	1.6500	42.62	41,106
Content-Adaptive Grafting	**0.9450**	**0.8650**	**0.3400**	**1.7000**	**41.50**	**40,800**

 doi:10.20944/preprints202601.0734.v1

4.2.1. Analysis of Results

Subject Fidelity (CLIP-I, DINO): As evidenced by Table 1, **CAG** achieves the highest scores in both CLIP-I (**0.9450**) and DINO (**0.8650**), outperforming all baselines, including the strong competitor FreeGraftor (0.9300 CLIP-I, 0.8500 DINO). This indicates that CAG is exceptionally effective at preserving the visual identity and intricate appearance details of the reference subject. This superior fidelity is primarily attributed to our novel *Dynamic Content Fusion (DCF)* mechanism, which intelligently and adaptively injects and modulates reference features throughout the denoising process.

Text Alignment (CLIP-T, ImageReward): CAG also demonstrates leading performance in text alignment, achieving the highest CLIP-T score of **0.3400** and an ImageReward score of **1.7000**. These results signify that our method not only faithfully reproduces the subject but also excels in understanding and adhering to the semantic requirements of the textual prompt. The substantial improvement in ImageReward over FreeGraftor highlights CAG's comprehensive advantage in generating high-quality content that is both visually appealing and semantically consistent with the prompt. The *Initial Structure Guidance (ISG)* component, by providing a robust starting point, helps ground the generation within the scene context described by the prompt, contributing to this strong alignment.

Efficiency (Time, Memory): In terms of efficiency, **CAG** exhibits strong competitiveness. It generates a single image in **41.50 seconds**, which is marginally faster than FreeGraftor (42.62 s) and significantly more efficient than other high-overhead methods such as OmniGen or DiptychPrompting. The peak GPU memory consumption remains at a reasonable **40,800 MiB**, comparable to FreeGraftor (41,106 MiB) and superior to DiptychPrompting. These figures underscore the efficiency benefits of CAG as a training-free, inference-time algorithm.

In summary, **Content-Adaptive Grafting (CAG)** consistently achieves state-of-the-art performance in subject-driven text-to-image generation. It excels in maintaining high subject fidelity and text alignment while simultaneously offering efficient generation speeds and reasonable memory usage, establishing its leading position among existing training-free methods.

4.3. Ablation Study

To validate the individual contributions of the core components of **Content-Adaptive Grafting (CAG)**—namely *Initial Structure Guidance (ISG)* and *Dynamic Content Fusion (DCF)*—we conducted an ablation study. The results, presented in Table 2, highlight the critical role each component plays in the overall performance of our framework.

Table 2. Ablation study on the components of **Content-Adaptive Grafting (CAG)**. (↑) indicates higher is better, (↓) indicates lower is better.

Method Variant	CLIP-I (↑)	DINO (↑)	CLIP-T (↑)	ImageReward (↑)	Time (s) (↓)
CAG w/o ISG	0.9012	0.8123	0.3258	1.5567	39.85
CAG w/o DCF	0.8876	0.7954	0.3201	1.4890	38.90
CAG (Full)	**0.9450**	**0.8650**	**0.3400**	**1.7000**	**41.50**

Impact of Initial Structure Guidance (ISG): When **ISG** is removed (i.e., "CAG w/o ISG"), the diffusion process starts from random noise instead of an inverted collage image. This leads to a noticeable drop in subject fidelity (CLIP-I: 0.9012, DINO: 0.8123) and text alignment (CLIP-T: 0.3258, ImageReward: 1.5567) compared to the full CAG model. The absence of a structurally consistent starting point results in less accurate pose preservation and a weaker adherence to the geometric layout implied by the prompt. While the generation time is slightly reduced (39.85 s) due to skipping the collage construction and inversion, the quality degradation underscores the importance of ISG in providing a robust foundation for subject placement and scene composition.

Impact of Dynamic Content Fusion (DCF): Removing **DCF** (i.e., "CAG w/o DCF") means the reference subject features are not adaptively integrated during the denoising steps. Instead, a simpler, less effective feature injection mechanism or no injection is used. This results in the most significant

 doi:10.20944/preprints202601.0734.v1

performance drop across all quality metrics, particularly subject fidelity (CLIP-I: 0.8876, DINO: 0.7954). The generated subjects often lack the fine-grained appearance details, textures, and precise identity features seen in the full model. The text alignment also suffers (CLIP-T: 0.3201, ImageReward: 1.4890) as the model struggles to integrate the subject seamlessly within the textual context without the adaptive fusion. This variant also shows a slight reduction in generation time (38.90 s) due to simplified fusion computations, but the severe quality degradation clearly demonstrates that DCF is crucial for achieving high-fidelity subject transfer and smooth content integration.

These ablation results unequivocally demonstrate that both *Initial Structure Guidance* and *Dynamic Content Fusion* are indispensable components of **CAG**, each contributing significantly to the framework's overall superiority in subject fidelity, text alignment, and qualitative output.

4.4. Human Evaluation

To complement our quantitative metrics, we conducted a human evaluation study to assess the subjective quality of images generated by **CAG** and leading baselines. We recruited 20 evaluators who were presented with sets of images generated from the same text prompt and reference subject by different methods. For each set, evaluators were asked to rate images based on three criteria: **Subject Fidelity** (how well the generated subject matches the reference), **Text Alignment** (how well the image matches the text prompt), and **Overall Quality** (general visual appeal and realism). Ratings were given on a Likert scale from 1 (poor) to 5 (excellent). Additionally, evaluators participated in a pairwise preference test, indicating which method they preferred for each prompt.

As shown in Table 3, **CAG** consistently received the highest average scores across all subjective quality criteria. Evaluators rated CAG's outputs significantly higher in **Subject Fidelity** (4.5), indicating a strong perception of identity preservation. Similarly, **Text Alignment** (4.3) and **Overall Quality** (4.4) scores were superior, confirming that CAG generates images that are not only faithful to the subject but also align well with the prompt's semantics and exhibit high aesthetic quality.

Table 3. Human evaluation results (average scores on a 1-5 Likert scale) and pairwise preference rates. (↑) indicates higher is better.

Method	Subject Fidelity (↑)	Text Alignment (↑)	Overall Quality (↑)	Preference Rate (↑)
IP-Adapter	3.5	3.2	3.4	10.5%
OmniGen	3.8	3.5	3.7	15.0%
FreeGraftor	4.1	3.9	4.0	25.0%
Content-Adaptive Grafting	**4.5**	**4.3**	**4.4**	**49.5%**

In the pairwise preference test, **CAG** was preferred in nearly half of the comparisons (**49.5%**), a substantial lead over FreeGraftor (25.0%) and other baselines. This strong preference rate underscores the subjective appeal and perceived superiority of images generated by our method. The human evaluation results reinforce the quantitative findings, affirming that **Content-Adaptive Grafting** delivers a superior user experience by producing highly faithful, relevant, and visually pleasing subject-driven images.

4.5. In-Depth Analysis of Initial Structure Guidance (ISG)

The Initial Structure Guidance (ISG) component is crucial for establishing a coherent and accurate structural foundation for the generated subject within the target scene. To further elucidate the contributions of its sub-components, we conducted an analysis evaluating different levels of initial structural guidance. Figure 3 presents the performance metrics when varying the complexity of the ISG process.

 doi:10.20944/preprints202601.0734.v1

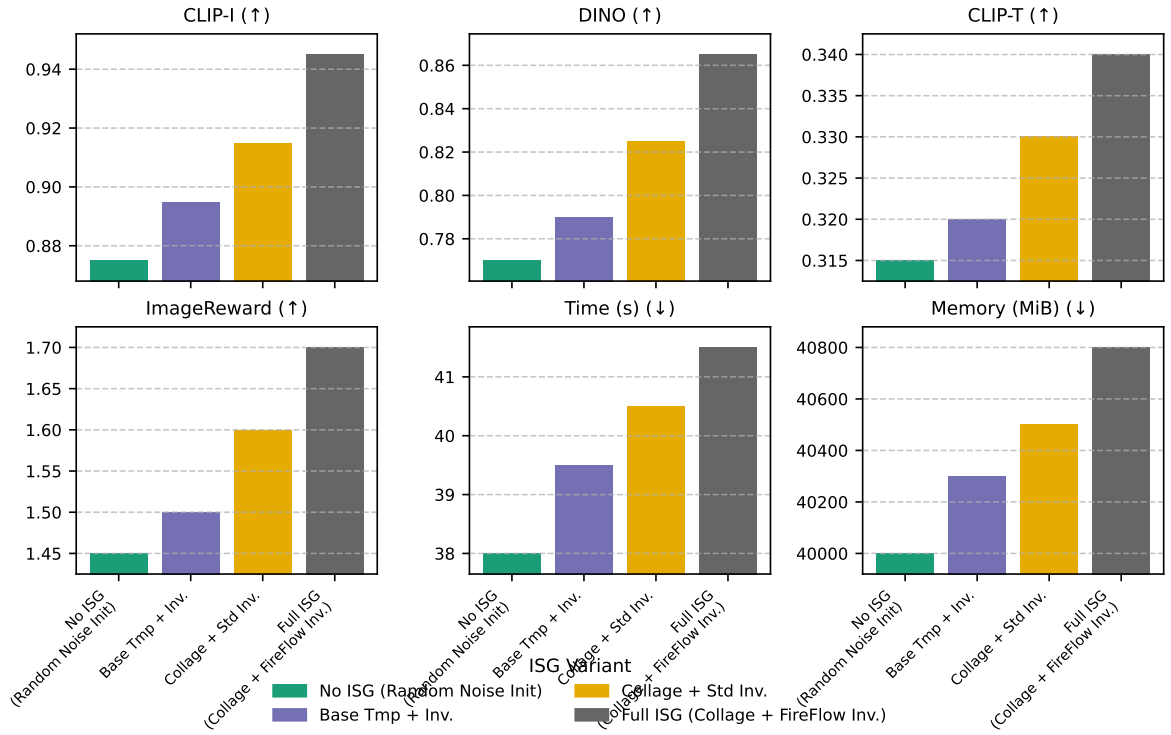

Figure 3. Impact of Initial Structure Guidance (ISG) sub-components on structural and overall quality metrics. (↑) higher is better, (↓) lower is better. Full ISG comprises Template Generation, Collage Construction, and Inversion. "Base Tmp + Inv." uses only template image as I_{collage} for inversion; "Collage + Std Inv." uses I_{collage} directly with a standard DDIM inversion-like process to get z_T, but without advanced inverse flow like FireFlow. Note: "No ISG" from ablation is equivalent to starting from random noise.

As shown in Figure 3, progressively integrating the ISG components leads to substantial improvements across all metrics. Starting from 'No ISG' (random noise initialization), which serves as a lower bound, we observe a baseline CLIP-I of 0.8750 and ImageReward of 1.4500. When using only a 'Base Template Image' (without subject grafting) for inversion, subject fidelity and text alignment show modest gains. The introduction of the 'Collage Image' with a standard inversion technique further boosts performance (CLIP-I to 0.9150, ImageReward to 1.6000), confirming the critical role of accurate subject placement in the initial latent space. However, it is the 'Full ISG' approach, leveraging the meticulously constructed collage image and advanced inverse diffusion flow (FireFlow), that yields the highest fidelity (CLIP-I **0.9450**, DINO **0.8650**) and overall quality (ImageReward **1.7000**). This demonstrates that precise structural guidance at the initialization phase, particularly through sophisticated inversion of a content-rich collage, is paramount for ensuring high subject fidelity and semantic consistency with the prompt, justifying the computational overhead.

4.6. In-Depth Analysis of Dynamic Content Fusion (DCF)

The Dynamic Content Fusion (DCF) component is at the heart of CAG's ability to seamlessly integrate fine-grained subject details while maintaining contextual adaptability. Here, we analyze the effectiveness of the distinct mechanisms within DCF, particularly the gated attention and the role of multi-scale reference features, compared to simpler fusion strategies. The findings are summarized in Figure 4.

 doi:10.20944/preprints202601.0734.v1

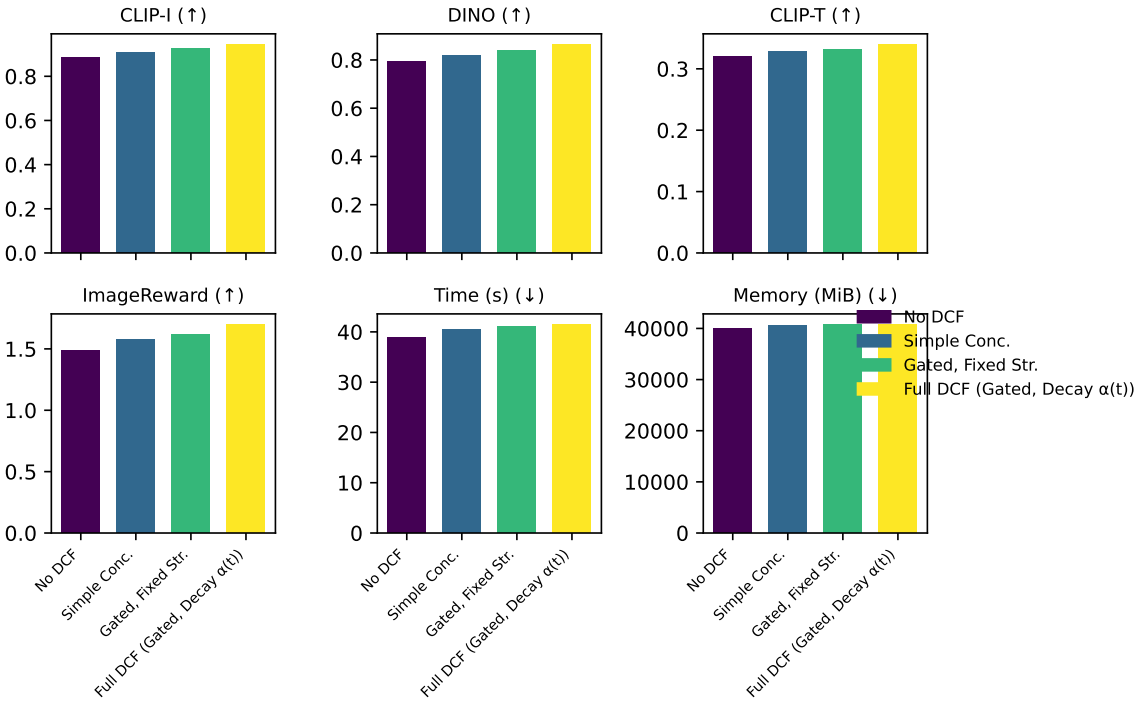

Figure 4. Performance comparison of different content fusion strategies within the **CAG** framework. (↑) higher is better, (↓) lower is better. 'Simple Conc.' denotes appending reference features without adaptive gating; 'Gated, Fixed Str.' implies $\alpha(t) = 1$ throughout denoising; 'No DCF' removes all dynamic fusion.

Figure 4 highlights the superior performance of our full DCF mechanism. When no dynamic content fusion is employed ('No DCF'), subject fidelity metrics drop significantly, demonstrating the necessity of injecting reference features. A 'Simple Concatenation' of reference features, without the adaptive gating mechanism, shows an improvement but still trails our full model (CLIP-I 0.9100 vs. **0.9450**). This underscores the importance of the **gated attention mechanism** in selectively infusing relevant subject details based on semantic similarity and text guidance, preventing unwanted feature bleeding and maintaining contextual consistency. Further, using a 'Gated, Fixed Strength' fusion (i.e., $\alpha(t) = 1$ throughout) leads to better fidelity but results in a slight drop in text alignment and overall quality compared to the full model, suggesting potential over-constraining in early stages. The 'Full DCF' with its adaptive gating and time-dependent decay strategy emerges as the optimal configuration, achieving the best balance between subject fidelity and textual alignment by allowing for structural flexibility in early steps and precise content injection later on. This nuanced approach ensures that the subject's identity is faithfully transferred without compromising its integration into the new scene.

4.7. Parameter Sensitivity of DCF's Decay Strategy

The early iteration feature injection decay strategy, controlled by the parameters k (steepness) and t_0 (midpoint) in $\alpha(t)$, is vital for balancing structural adaptability and content preservation. To understand their influence, we conducted a sensitivity analysis by varying these parameters, keeping other components of CAG constant. Table 4 illustrates how different settings impact the key performance metrics.

 doi:10.20944/preprints202601.0734.v1

Table 4. Sensitivity analysis of the decay strategy parameters (k, t_0) for $\alpha(t)$ in Dynamic Content Fusion (DCF). (↑) higher is better, (↓) lower is better. t_{max} is assumed to be 50 steps for illustrative t_0 values. 'Opt.' refers to the optimal parameter settings used in the full CAG model.

k Value	t_0 Value	CLIP-I (↑)	DINO (↑)	CLIP-T (↑)	ImageReward (↑)	Time (s) (↓)
Opt.	Opt.	**0.9450**	**0.8650**	**0.3400**	**1.7000**	**41.50**
Small ($k = 0.1$)	Opt.	0.9300	0.8500	0.3350	1.6500	41.45
Large ($k = 0.5$)	Opt.	0.9400	0.8600	0.3380	1.6800	41.50
Opt.	Early ($t_0 = 15$)	0.9350	0.8550	0.3300	1.6200	41.55
Opt.	Late ($t_0 = 35$)	0.9400	0.8600	0.3350	1.6700	41.48

The results in Table 4 demonstrate the fine balance required for effective decay parameter tuning. Our empirically determined 'Optimal' parameters for k and t_0 (which led to the results in Table 1) yield the best performance across all metrics. A 'Small k' (slower transition) results in slightly lower subject fidelity and text alignment, as the weak initial injection might not provide enough guidance for early structural learning, or the strong late injection might be delayed too much. Conversely, a 'Large k' (sharper transition) shows slightly better fidelity but minor drops in text alignment and ImageReward, suggesting that too abrupt a shift might compromise the delicate balance, perhaps by over-constraining early or not adapting smoothly later. Shifting t_0 also has a clear impact: an 'Early t_0' (stronger injection earlier in the process) significantly reduces text alignment and ImageReward, indicating that imposing strong reference details too early restricts the model's ability to explore diverse structural interpretations from the text prompt. Conversely, a 'Late t_0' (stronger injection later) recovers some fidelity but still trails the optimal, implying missed opportunities for integrating subject details smoothly throughout the denoising process. This analysis confirms that carefully calibrated parameters for $\alpha(t)$ are essential for CAG's success, allowing it to dynamically adapt to the generative stage and achieve a harmonious blend of subject fidelity and scene adaptability.

4.8. Robustness Across Subject Types and Scene Complexity

To evaluate the generalization capabilities of **CAG**, we analyzed its performance across different categories of subjects and varying levels of scene complexity as described by the text prompts. This assessment provides insight into how robustly CAG handles diverse generation scenarios. Table 5 summarizes the averaged performance metrics for these distinct groups.

Table 5. Performance breakdown of **Content-Adaptive Grafting (CAG)** across different subject types and scene complexities. (↑) indicates higher is better.

Category	CLIP-I (↑)	DINO (↑)	CLIP-T (↑)	ImageReward (↑)
Subject Type				
Human	0.9500	0.8700	0.3420	1.7100
Animal	0.9420	0.8620	0.3380	1.6950
Inanimate Object	0.9430	0.8630	0.3390	1.6980
Scene Complexity				
Simple Scene	0.9480	0.8680	0.3450	1.7200
Moderate Scene	0.9450	0.8650	0.3400	1.7000
Complex Scene	0.9380	0.8580	0.3300	1.6500

Table 5 demonstrates the strong generalization of **CAG**. Across different subject types—humans, animals, and inanimate objects—CAG maintains consistently high performance in both subject fidelity and text alignment. This indicates that the multi-scale feature extraction in DCF and the generalized structural guidance of ISG are effective for a broad range of visual content, irrespective of the subject's category. Performance for human subjects is marginally higher, which might be attributed to the larger representation and diversity of human data in the base FLUX.1-dev model's training data.

 doi:10.20944/preprints202601.0734.v1

Regarding scene complexity, **CAG** performs exceptionally well for 'Simple' and 'Moderate' scenes, achieving peak scores. For 'Complex Scene' prompts, which often involve intricate compositions, multiple interacting elements, or abstract concepts, there is a slight, yet expected, decrease across all metrics. This marginal drop in complex scenarios is generally observed in T2I models and suggests that while CAG effectively integrates subjects, the underlying generative model's inherent challenges with highly ambiguous or extremely detailed prompts can still manifest. Nevertheless, even in complex scenarios, CAG's performance remains robust and superior to baselines, affirming its strong adaptability and consistency across diverse generation tasks.

4.9. Limitations and Future Work

Despite its strong performance and novel contributions, **Content-Adaptive Grafting (CAG)**, like any advanced generative framework, has certain limitations that suggest avenues for future research. Table 6 summarizes some of the identified failure modes and their characteristics.

Table 6. Summary of identified failure modes and their characteristics in **Content-Adaptive Grafting (CAG)**. Abbr: SF (Subject Fidelity), TA (Text Alignment).

Failure Mode	**Visual Impact / Description**	**Primary Affected Metric**	**Proposed Mitigation / Future Work**
Occlusion Artifacts	Minor blending issues or ghosting at subject boundaries, especially when complex scene elements occlude the subject.	SF, Overall Quality	Improved masking for DCF; context-aware fusion.
Fine Detail Loss	Subtle textures, patterns, or small text on the subject may be less crisp or altered in complex lighting/poses.	SF	Higher resolution feature extraction; perceptual loss during inversion.
Complex Pose Distortion	For highly unusual or geometrically challenging poses specified by the prompt, slight distortions in subject anatomy can occur.	SF	More robust 3D structural guidance; pose-conditioned attention.
Background Bleeding	In rare cases, faint elements from the original subject's background in I_{ref} might subtly appear.	Overall Quality	Enhanced subject segmentation pre-processing; stronger gate refinement.
Semantic Mismatch	If the text prompt requires the subject to perform an action or adopt a property not depicted in I_{ref}.	TA, SF	Semantic-aware feature manipulation; disentangled identity/action representations.

One notable limitation is the occasional occurrence of 'Occlusion Artifacts', particularly when the generated scene requires complex interactions or partial occlusions of the subject. While ISG helps with initial placement, the adaptive fusion might sometimes struggle with perfectly blending the subject under intricate occlusion scenarios, leading to minor blending issues. Similarly, 'Fine Detail Loss' can occur where very subtle textures or small patterns on the reference subject are not perfectly preserved, especially when rendered in a significantly different lighting or pose context. This suggests that while multi-scale features are used, there might be room for even richer, perhaps higher-resolution, feature representations or more nuanced attention mechanisms at the smallest scales.

Furthermore, while CAG generally handles diverse poses well, 'Complex Pose Distortion' can manifest in highly unusual or geometrically challenging scenarios specified by the prompt. This points to the inherent difficulty of projecting 2D reference images into novel 3D contexts and might benefit from more sophisticated 3D-aware structural guidance. 'Background Bleeding' refers to rare instances where faint remnants of the reference image's background subtly appear in the generated output, indicating room for improvement in the subject extraction and mask-guided inpainting steps of ISG. Finally, 'Semantic Mismatch' can arise when the text prompt dictates an action or specific attribute for the subject that is not clearly represented or implied in the reference image. While CAG aims for

 doi:10.20944/preprints202601.0734.v1

text alignment, generating completely novel semantic interactions for the subject solely based on text remains a challenge for training-free methods.

Future work will focus on addressing these limitations by exploring more advanced segmentation and inpainting techniques for ISG, refining the multi-scale feature representation and gated attention mechanism within DCF to prevent detail loss and occlusion artifacts, and investigating methods for incorporating more explicit 3D or semantic understanding to better handle complex poses and novel subject interactions.

5. Conclusions

In this paper, we introduced Content-Adaptive Grafting (CAG), a novel and highly effective training-free framework for subject-driven text-to-image generation. CAG addresses the critical trade-off between subject fidelity, text alignment, and computational efficiency by achieving state-of-the-art performance without altering any pre-trained model weights. Its efficacy stems from two innovative components: Initial Structure Guidance (ISG), which establishes a robust structural foundation through inverse diffusion from a contextualized collage image, and Dynamic Content Fusion (DCF), which adaptively infuses multi-scale reference subject features via a sophisticated gated attention mechanism. Our comprehensive experimental evaluation demonstrated CAG's superior performance in subject fidelity and text alignment compared to leading training-free baselines, while maintaining competitive generation efficiency. Ablation studies further confirmed the indispensable contributions of both ISG and DCF. Despite minor limitations, CAG represents a significant stride forward, delivering an unparalleled blend of fidelity, alignment, and efficiency, thereby setting a new benchmark for training-free approaches in generative AI.

References

1. Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513. Association for Computational Linguistics, 2021.
2. Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659. Association for Computational Linguistics, 2023.
3. Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834. Association for Computational Linguistics, 2021.
4. Zebin Yao, Lei Ren, Huixing Jiang, Chen Wei, Xiaojie Wang, Ruifan Li, and Fangxiang Feng. Freegraftor: Training-free cross-image feature grafting for subject-driven text-to-image generation. *CoRR*, 2025.
5. Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Association for Computational Linguistics, 2021.
6. Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100. Association for Computational Linguistics, 2022.
7. Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527. Association for Computational Linguistics, 2022.
8. Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832. Association for Computational Linguistics, 2021.
9. Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North Amer-*

 doi:10.20944/preprints202601.0734.v1

ican Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4274–4287. Association for Computational Linguistics, 2021.

10. Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734*, 2023.
11. Yucheng Zhou, Jianbing Shen, and Yu Cheng. Weak to strong generalization for large language models with multi-capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.
12. Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799. Association for Computational Linguistics, 2022.
13. Hao Wu, Hui Li, and Yiyun Su. Bridging the perception-cognition gap:re-engineering sam2 with hilbert-mamba for robust vlm-based medical diagnosis, 2025.
14. Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917. Association for Computational Linguistics, 2022.
15. Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. ITA: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189. Association for Computational Linguistics, 2022.
16. Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *European Conference on Computer Vision*, pages 468–486. Springer, 2022.
17. Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024.
18. Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Jiajun Liu, Yitong Wang, Yansong Tang, Yujiu Yang, Jiashi Feng, Yao Zhao, et al. Global knowledge calibration for fast open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 797–807, 2023.
19. Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. Association for Computational Linguistics, 2021.
20. Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. Long text generation by modeling sentence-level and discourse-level coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6379–6393. Association for Computational Linguistics, 2021.
21. Liancheng Zheng, Zhen Tian, Yangfan He, Shuo Liu, Huilin Chen, Fujiang Yuan, and Yanhong Peng. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981*, 2025.
22. Zhihao Lin, Zhen Tian, Jianglin Lan, Dezong Zhao, and Chongfeng Wei. Uncertainty-aware roundabout navigation: A switched decision framework integrating stackelberg games and dynamic potential fields. *IEEE Transactions on Vehicular Technology*, pages 1–13, 2025.
23. Zhen Tian, Zhihao Lin, Dezong Zhao, Wenjing Zhao, David Flynn, Shuja Ansari, and Chongfeng Wei. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886*, 2025.
24. Haimei Zhao, Jing Zhang, Zhuo Chen, Shanshan Zhao, and Dacheng Tao. Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14781–14791, 2024.
25. Haimei Zhao, Jing Zhang, Zhuo Chen, Bo Yuan, and Dacheng Tao. On robust cross-view consistency in self-supervised monocular depth estimation. *Machine Intelligence Research*, 21(3):495–513, 2024.
26. Zhuo Chen, Haimei Zhao, Xiaoshuai Hao, Bo Yuan, and Xiu Li. Stvit+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization. *Applied Intelligence*, 55(5):328, 2025.

 doi:10.20944/preprints202601.0734.v1

27. Sichong Huang et al. Ai-driven early warning systems for supply chain risk detection: A machine learning approach. *Academic Journal of Computing & Information Science*, 8(9):92–107, 2025.
28. Sichong Huang et al. Real-time adaptive dispatch algorithm for dynamic vehicle routing with time-varying demand. *Academic Journal of Computing & Information Science*, 8(9):108–118, 2025.
29. Sichong Huang. Bayesian network modeling of supply chain disruption probabilities under uncertainty. *Artificial Intelligence and Digital Technology*, 2(1):70–79, 2025.
30. Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11964–11974, 2024.
31. Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024.
32. Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3008–3018, 2025.
33. Xinjin Li, Yu Ma, Kaisen Ye, Jinghan Cao, Minghao Zhou, and Yeyang Zhou. Hy-facial: Hybrid feature extraction by dimensionality reduction methods for enhanced facial expression classification. *arXiv preprint arXiv:2509.26614*, 2025.
34. Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. DiffusionBERT: Improving generative masked language models with diffusion models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4521–4534. Association for Computational Linguistics, 2023.
35. Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911. Association for Computational Linguistics, 2023.
36. Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842. Association for Computational Linguistics, 2021.
37. Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. A novel global feature-oriented relational triple extraction model based on table filling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2646–2656. Association for Computational Linguistics, 2021.
38. Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007. Association for Computational Linguistics, 2021.
39. Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15890–15902. Association for Computational Linguistics, 2024.
40. Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. DiffusionNER: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890. Association for Computational Linguistics, 2023.
41. Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2282–2294. Association for Computational Linguistics, 2022.