Preprints.org

Article

Not peer-reviewed version

Adaptive Contextual Feature Grafting and Hierarchical Structure-Aware Initialization for Training-Free Subject-Driven Text-to-Image Generation

Salma Ali * and Noah Fang

Posted Date: 18 December 2025

doi: 10.20944/preprints202512.1688.v1

Keywords: text-to-image generation; subject-driven; diffusion transformer; training-free; feature grafting

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Adaptive Contextual Feature Grafting and Hierarchical Structure-Aware Initialization for Training-Free Subject-Driven Text-to-Image Generation

Salma Ali * and Noah Fang

Universiti Teknologi Malaysia
* Correspondence: hva180033@siswa365.um.edu.my

Abstract

Subject-driven text-to-image (T2I) generation presents a significant challenge in balancing subject fidelity and text alignment, with traditional fine-tuning approaches proving inefficient. We introduce ContextualGraftor, a novel training-free framework for robust subject-driven T2I generation, leveraging the powerful FLUX.1-dev multimodal diffusion-transformer. It integrates two core innovations: Adaptive Contextual Feature Grafting (ACFG) and Hierarchical Structure-Aware Initialization (HSAI). ACFG enhances feature matching in attention layers through a lightweight contextual attention module that dynamically modulates reference feature contributions based on local semantic consistency, ensuring natural integration and reduced semantic mismatches. HSAI provides a structurally rich starting point by employing multi-scale structural alignment during latent inversion and an adaptive dropout strategy, preserving both global geometry and fine-grained subject details. Comprehensive experiments demonstrate that ContextualGraftor achieves superior performance across key metrics, outperforming state-of-the-art training-free methods like FreeGraftor. Furthermore, our method maintains competitive inference efficiency, offering an efficient and high-performance solution for seamless subject integration into diverse, text-prompted environments.

Keywords: text-to-image generation; subject-driven; diffusion transformer; training-free; feature grafting

1. Introduction

The field of text-to-image (T2I) generation has witnessed remarkable advancements in recent years, empowering users to synthesize diverse and high-quality images from textual descriptions [1]. These models have demonstrated impressive capabilities in understanding complex prompts and generating visually coherent scenes, a capability often underpinned by the rapid development of large language models (LLMs) and multimodal AI [2–4]. The broader landscape of AI research similarly sees rapid progress in interactive decision-making and complex system control, from autonomous navigation to supply chain optimization [5–8], and in analytical tasks such as causal inference in medical research [9,10]. However, a particularly challenging yet crucial task remains: *subject-driven text-to-image generation*. This involves precisely incorporating a specific "subject" (e.g., a particular object, character, or pet) from a given reference image into a novel scene, background, or action described by a text prompt, while maintaining the subject's identity and visual details.

The core difficulty of this task lies in balancing two often conflicting objectives: first, **Subject Fidelity**, ensuring that the generated subject meticulously preserves the unique textures, colors, shapes, and intricate details from the reference image; and second, **Text Alignment**, guaranteeing that the overall generated image accurately adheres to the scene, background, layout, and semantic requirements outlined in the text prompt. Traditional approaches to subject-driven generation frequently rely on fine-tuning large pre-trained T2I models for each new subject [11]. While effective in achieving high

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

fidelity, this fine-tuning process is computationally intensive, time-consuming, and requires re-training for every new subject, making it inefficient for practical applications. Consequently, there is a strong motivation to develop **efficient and training-free** methods that can achieve excellent subject-driven image generation without modifying base model weights or training additional parameters [12]. Such advancements are critical across various AI domains, from forecasting in retail operations to medical risk stratification, where efficient and accurate predictive models are paramount [13,14]. Our work builds upon the successes of existing training-free techniques, such as FreeGraftor [15], aiming to further elevate generation quality and contextual consistency, particularly in scenarios involving complex compositions and multi-subject interactions. The rise of multimodal large language models (MLLMs) and their ability to process and generate content across modalities further highlights the importance of such controllable generation methods [16–18]. This ongoing push for advanced AI capabilities is mirrored in efforts to improve diagnostic tools in pathology and enhance reliability through robust benchmarking of ML/DL models [19,20].

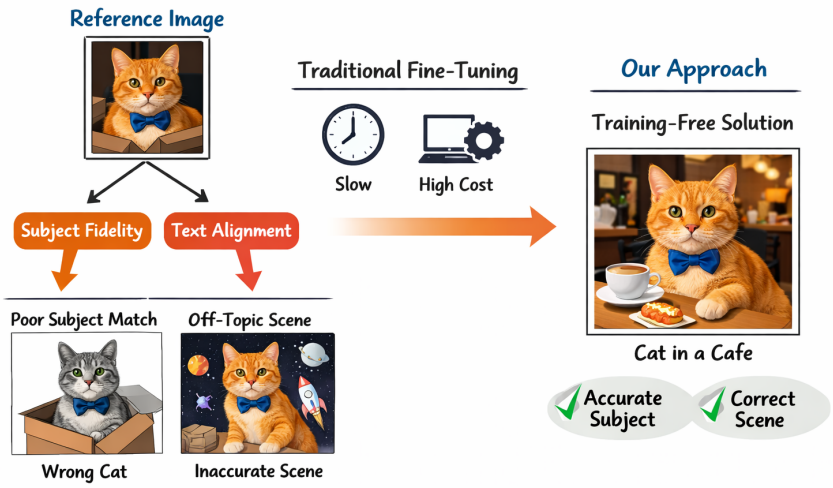

Figure 1. Motivation of subject-driven text-to-image generation: existing methods struggle to balance subject fidelity and text alignment or require costly fine-tuning, while our training-free approach achieves accurate subject preservation and scene consistency efficiently.

In this paper, we propose **ContextualGraftor**, a novel training-free framework designed for robust subject-driven text-to-image generation. Our method leverages a powerful foundation generative model, **FLUX.1-dev** [21], a multimodal diffusion-transformer (MM-DiT) type model known for its superior generation capabilities, which aligns with recent advancements in multimodal architectures [22]. ContextualGraftor integrates a suite of innovative strategies, including *Adaptive Contextual Feature Grafting (ACFG)* and *Hierarchical Structure-Aware Initialization (HSAI)*. ACFG enhances feature matching in the MM-DiT's attention layers by incorporating a lightweight contextual attention module that dynamically adjusts the contribution of reference features based on the local semantic consistency around the subject. This ensures more natural integration and reduces semantic mismatches. HSAI, on the other hand, employs a multi-stage process involving tools like Grounding DINO [23], SAM [24], and LaMa [25] to construct a structured collage, which is then inverted using FireFlow [26]. Crucially, HSAI introduces a multi-scale structural alignment loss during inversion and an adaptive dropout strategy during diffusion, guaranteeing that the initial latent noise and subsequent generation process strongly preserve both global geometry and fine-grained internal structures of the subject. These mechanisms collectively enable ContextualGraftor to maintain high subject fidelity while seamlessly integrating subjects into diverse, text-prompted environments.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

To rigorously evaluate ContextualGraftor, we conducted extensive quantitative and qualitative experiments against several state-of-the-art baselines. Our evaluation utilizes a diverse dataset comprising various subject categories (e.g., animals, persons, objects) and complex text prompts, akin to the dataset used in FreeGraftor [15]. We employed standard metrics for assessing subject fidelity, including CLIP Image Similarity (CLIP-I) and DINOv2 Feature Similarity (DINO), and for evaluating text alignment, such as CLIP Text-Image Similarity (CLIP-T) and ImageReward [27]. Our results consistently demonstrate that ContextualGraftor achieves superior performance across all these key metrics, outperforming even the most advanced training-free methods, including FreeGraftor. Furthermore, we show that our method maintains competitive inference efficiency, with generation times and memory usage comparable to, or slightly better than, existing FLUX.1-dev-based approaches. This underscores ContextualGraftor's potential as an efficient and high-performance solution for subject-driven T2I generation.

Our main contributions are summarized as follows:

- We introduce **Adaptive Contextual Feature Grafting (ACFG)**, a novel mechanism that uses context-aware attention fusion to dynamically integrate subject features, significantly enhancing the naturalness and contextual consistency of grafted subjects in new scenes.
- We propose **Hierarchical Structure-Aware Initialization (HSAI)**, a multi-stage initialization strategy incorporating multi-scale structural alignment during latent inversion and an adaptive dropout schedule during diffusion, ensuring robust preservation of both global and fine-grained subject structures. This leverages insights from advanced image editing and generation techniques [12,28].
- We demonstrate that **ContextualGraftor** sets a new state-of-the-art for training-free subject-driven text-to-image generation, achieving superior subject fidelity and text alignment across diverse benchmarks while maintaining competitive computational efficiency.

2. Related Work

2.1. Subject-Driven Text-to-Image Generation

Subject-driven text-to-image generation is a critical area, focusing on generating specific, identity-preserving subjects from text prompts with intricate control. It relies on large-scale vision-language pre-training (VLP) models like E2E-VLP [29] and mPLUG [30] for robust cross-modal representations. Multimodal AI [4] and strong text comprehension via models such as mT6 [31] and advanced LLMs [2] are foundational for complex prompt interpretation. Diffusion models dominate high-quality image synthesis, guided by principles of controllable output [32]. Precise identity preservation in subject-driven synthesis employs mechanisms such as Image-Text Alignments (ITA) [33], NeuroLogic Decoding [34], and semantic guidance (e.g., EDGE [35]). Visual in-context learning in large vision-language models [16] further refines this control. Efficiently adapting pre-trained models to new subjects is a central challenge. Parameter-efficient fine-tuning, including prompt tuning [36] and heterogeneous MoE adapters [37], enables rapid few-shot adaptation and contributes to weak-to-strong generalization [3]. Precise control in interactive systems also benefits from decision-making frameworks and game theory [5,6,38]. In summary, subject-driven generation integrates robust vision-language foundations, semantic alignment, fine-grained control, and parameter-efficient adaptation for precise image synthesis. Research continues to improve fidelity, consistency, and controllability, aided by enhanced code LLMs and AI narrative coherence for complex prompt interpretation [39,40].

2.2. Diffusion Model Architectures and Image Manipulation Techniques

Diffusion models have revolutionized image synthesis and manipulation by progressively refining random noise into coherent data via a reversed noisy process. Their Transformer-based architectures, exemplified by DiffusionBERT [41] for text, are versatile for both text and image generation. Surveys highlight rapid progress in diffusion model-based image editing [28]. Effective image manipulation leverages sophisticated architectures and controllable processes. Attention mechanisms, crucial for

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

feature weighting and dependency capture (e.g., in LLMs for radiology [42]), are vital for image synthesis and editing. Precise manipulation often involves latent space navigation. Latent space inversion techniques, similar to learning shared semantic spaces for speech-to-text [43], are crucial for targeted edits in diffusion models' compressed representations. Novel approaches like Dual-Schedule Inversion [12] further enhance training-free real image editing. Essential techniques include inpainting, where diffusion models excel. Advancements in segmentation, including video object and open-vocabulary segmentation [44–46], provide critical regional control, with frameworks like EP-SAM [47] showing utility in ultra-low light. Computer vision tasks like semi-supervised facial expression recognition also benefit from such architectural and methodological advancements [48]. Visual grounding—aligning text with visual features—is vital for text-conditioned image generation and editing; image-language transformer studies (e.g., verb interpretation [49]) highlight the need for robust visual-linguistic integration in controllable diffusion systems. New multimodal architectures, such as Multi-Modal Mamba for text-to-video [22] and equivariant diffusion models for robotics [50], demonstrate the continued evolution of generative frameworks. In summary, diffusion model advancements, combined with latent space manipulation, attention mechanisms, visual grounding, and techniques like inpainting and segmentation, form a sophisticated landscape for image generation and manipulation, with principles often transferable across domains.

3. Method

We introduce **ContextualGraftor**, a novel training-free framework for subject-driven text-to-image generation that significantly enhances the natural integration of subjects while preserving their identity. Our method builds upon a powerful foundation model, **FLUX.1-dev**, and integrates a series of innovative contextual feature grafting and structure-aware initialization strategies, all performed at inference time.

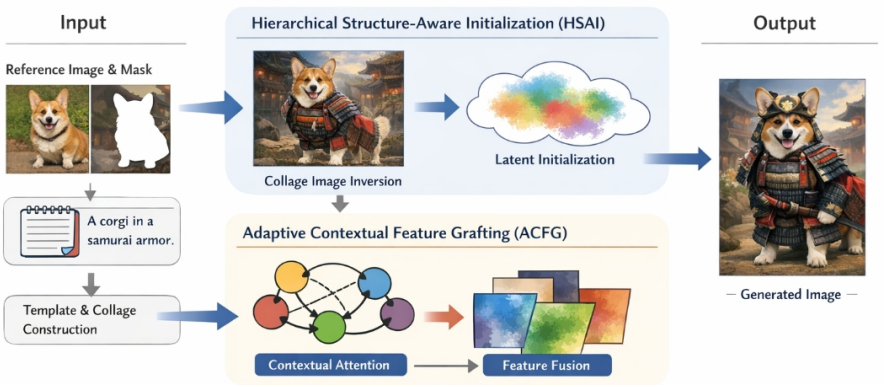

Figure 2. Overview of the ContextualGraftor framework for training-free subject-driven text-to-image generation, illustrating the end-to-end pipeline from reference subject and text prompt, through Hierarchical Structure-Aware Initialization (HSAI) and Adaptive Contextual Feature Grafting (ACFG), to the final identity-preserving and context-aligned generated image.

3.1. Overall Framework Architecture

ContextualGraftor operates as a plugin-like framework, orchestrating a sequence of operations around a pre-trained text-to-image diffusion model, specifically **FLUX.1-dev**. This multimodal diffusion-transformer (MM-DiT) serves as our underlying generative backbone due to its superior capabilities in handling complex prompts and generating high-quality images. The framework begins by accepting one or more reference images, each containing a specific subject to be integrated, along with their corresponding segmentation masks. A descriptive text prompt specifying the target scene, layout, or action is also provided. The overall process unfolds in several key stages:

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

3.1.1. Template Generation and Collage Construction

Initially, an overall template image is generated solely from the provided text prompt. To prepare for subject integration, external vision models are then strategically employed to construct a structured collage image. First, **Grounding DINO** is utilized to localize potential subject regions within this initial template, guided by the textual prompt. Concurrently, **SAM** (Segment Anything Model) precisely extracts the detailed mask of the subject from its corresponding reference image. Subsequently, **LaMa** performs intelligent inpainting on the template image, removing existing content from the regions where the new subject is designated to be placed. Finally, the masked and appropriately resized reference subject is accurately pasted into these inpainted regions of the template, resulting in a **structure-guided collage image** I_c. This collage serves as a strong structural prior for the subsequent steps.

3.1.2. Hierarchical Structure-Aware Initialization (HSAI)

Following the construction of I_c, the **Hierarchical Structure-Aware Initialization (HSAI)** phase begins. The collage image I_c is inverted into a latent space representation using **FireFlow** to obtain an initial latent noise z_T. This inversion process is critically augmented with a multi-scale structural alignment loss, detailed further in Section 3.3.1, to ensure robust structural preservation, meaning z_T accurately reflects the intricate details of I_c in the latent domain.

3.1.3. Adaptive Contextual Feature Grafting (ACFG)

During the subsequent iterative denoising steps of the diffusion process, our core contribution, **Adaptive Contextual Feature Grafting (ACFG)**, is performed. This mechanism operates within the attention layers of the **FLUX.1-dev** model. ACFG dynamically fuses reference features with the features currently being generated, with this fusion modulated by a lightweight contextual attention mechanism designed to ensure seamless and contextually consistent integration. An adaptive dropout strategy is also applied to the injected reference features throughout the diffusion steps, as elaborated in Section 3.3.2, further refining the subject's adaptation.

3.1.4. Image Generation

The diffusion process then iteratively refines the latent representation, starting from the structurally initialized z_T and progressing through multiple denoising steps until z_0 is reached. This culmination results in the generation of the final high-quality image, denoted as I_{gen} (typically 512×512 pixels), which faithfully incorporates the subject into the novel scene described by the text prompt while maintaining its identity and contextual realism.

3.2. Adaptive Contextual Feature Grafting (ACFG)

Our core contribution, **Adaptive Contextual Feature Grafting (ACFG)**, operates within the self-attention blocks of the **FLUX.1-dev** MM-DiT architecture. Unlike conventional feature grafting that might directly inject or replace features, ACFG introduces a context-aware fusion mechanism to significantly enhance the naturalness and contextual consistency of the grafted subject within the generated scene.

For a given attention layer and at a specific diffusion timestep, let $Q_g, K_g, V_g \in \mathbb{R}^{N \times d}$ represent the query, key, and value features derived from the currently generating image patches. Simultaneously, $K_r, V_r \in \mathbb{R}^{N_s \times d}$ denote the corresponding key and value features extracted from the reference subject, where N is the total number of patches in the generated image, N_s is the number of patches specifically belonging to the subject region, and d is the feature dimension. For any patch p that falls within the designated subject region, ACFG fuses its generated features K_g^p and V_g^p with its reference counterparts K_r^p and V_r^p.

The key innovation of ACFG lies in its ability to dynamically adjust the contribution of these reference features based on the local semantic context. To achieve this, we introduce a lightweight contextual attention module designed to compute a contextual weight $\lambda_p \in [0, 1]$ for each subject patch

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

p. This module robustly evaluates the semantic consistency and compatibility between the specific features of the reference subject patch and the surrounding generated context. The contextual weight λ_p can be precisely formulated as:

$$\lambda_p = \sigma\Big(\text{MLP}\Big(\text{Concat}\Big(\mathcal{F}_{\text{context}}(F_g^p), \mathcal{F}_{\text{subject}}(F_r^p)\Big)\Big)\Big) \tag{1}$$

Here, $\mathcal{F}_{\text{context}}(F_g^p)$ represents context-aware features derived from the generated image patches surrounding patch p, capturing the local scene semantics. Conversely, $\mathcal{F}_{\text{subject}}(F_r^p)$ represents features extracted directly from the reference subject image, corresponding to patch p, thereby encoding the subject's identity and appearance. Concat denotes the operation of concatenating these feature vectors, MLP refers to a compact multi-layer perceptron that maps the concatenated features to a scalar, and σ is the sigmoid activation function, which ensures that λ_p is bounded within the range $[0, 1]$.

Once λ_p is computed, the fused key and value features, K_{fused}^p and V_{fused}^p, for a subject patch p are then adaptively combined as a weighted sum:

$$K_{fused}^p = (1 - \lambda_p)K_g^p + \lambda_p K_r^p \tag{2}$$

$$V_{fused}^p = (1 - \lambda_p)V_g^p + \lambda_p V_r^p \tag{3}$$

For patches located outside the subject region, their key and value features remain solely derived from the generated image, i.e., $K_{fused}^p = K_g^p$ and $V_{fused}^p = V_g^p$. These adaptively fused features are subsequently used within the standard self-attention mechanism of the transformer block:

$$\text{AttentionOutput} = \text{softmax}\left(\frac{Q_g K_{fused}^T}{\sqrt{d}}\right) V_{fused} \tag{4}$$

This adaptive weighting strategy is critical as it ensures that reference features are injected with varying degrees of emphasis, allowing the model to intelligently balance subject fidelity with the demands of contextual realism and scene coherence.

Furthermore, a **position-constrained attention fusion** strategy is employed to maintain spatial consistency. This strategy ensures that the position encodings for reference subject patches are precisely aligned and consistent with their target locations within the context of the generated image. This careful handling of positional information is crucial for preventing spatial distortions and ensuring that the subject integrates seamlessly into the new scene, preserving its spatial coherence and interaction with the environment.

3.3. Hierarchical Structure-Aware Initialization (HSAI)

Hierarchical Structure-Aware Initialization (HSAI) is specifically engineered to provide a robust and structurally rich starting point for the subsequent diffusion process. Its primary objective is to ensure that the initial latent noise z_T, derived from the structure-guided collage image I_c, effectively encodes not only the subject's overall geometry but also its intricate internal structures and fine-grained details. HSAI achieves this through two principal components: a multi-scale structural alignment loss applied during the latent inversion phase and an adaptive dropout strategy employed during the iterative diffusion steps.

3.3.1. Multi-Scale Structural Alignment during Latent Inversion

After the construction of the structure-guided collage image I_c, a crucial step involves performing a latent inversion using **FireFlow** to obtain the initial latent noise representation z_T. To ensure that this z_T is a highly accurate and structurally faithful representation of I_c, we introduce a **multi-scale structural alignment loss**, denoted as $\mathcal{L}_{\text{HSAI}}$, which guides the optimization of z_T during this inversion process. This loss is designed to encourage z_T to decode back into an image that is structurally

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

consistent with the input I_c across multiple levels of abstraction, from coarse outlines to subtle textures. The multi-scale structural alignment loss is formulated as:

$$\mathcal{L}_{\text{HSAI}}(z_T, I_c) = \sum_{m=1}^{M} \|\Psi_m(\mathcal{D}(z_T)) - \Psi_m(I_c)\|_1 \tag{5}$$

In this equation, $\mathcal{D}(\cdot)$ represents the decoding process from the latent space (z_T) back to the pixel space, yielding a reconstructed image. $\Psi_m(\cdot)$ denotes a feature extraction function that obtains intermediate feature maps at scale m from a pre-trained, fixed image encoder (for example, specific layers of a VGG or CLIP image encoder). M signifies the total number of distinct scales (or feature map depths) considered. By minimizing this comprehensive loss, we effectively ensure that the inverted latent z_T captures both the macro-level composition and the micro-level details of the subject within the collage, thereby providing a high-quality and structurally coherent initial state for the subsequent diffusion process and significantly mitigating the potential for structural distortions or inconsistencies in the final output.

3.3.2. Adaptive Dropout for Reference Feature Injection

During the iterative denoising steps of the diffusion process, a novel **adaptive dropout strategy** is applied to the injected reference features within the attention layers. This strategy is critical for allowing the generative model greater flexibility in adapting the subject to novel poses, lighting conditions, and scenarios during the early, more noisy diffusion steps. As the diffusion process progresses and the image becomes clearer, the strategy gradually enforces stricter fidelity to the reference subject's identity.

Let $t \in [0, T]$ represent the current diffusion timestep, where T signifies the total number of diffusion steps (corresponding to maximum noise), and 0 indicates the final, clean image. The probability $P_r(t)$ of dropping (i.e., intentionally not injecting) the reference features for a subject patch at a given timestep t is defined as a linearly decreasing function:

$$P_r(t) = P_{\text{min}} + (P_{\text{max}} - P_{\text{min}}) \cdot \frac{t}{T} \tag{6}$$

Here, P_{max} is the maximum dropout rate, which is applied at the very beginning of the diffusion process (when $t = T$, representing the highest noise level). Conversely, P_{min} is the minimum dropout rate, applied towards the very end of the process (when $t = 0$, representing the lowest noise level). This formulation means that in the early stages of generation (large t), reference features are more likely to be dropped, granting the model increased freedom to creatively synthesize the subject's global structure and its interaction with the scene context. As the diffusion progresses towards a clean image (small t), $P_r(t)$ decreases significantly, compelling the model to adhere more strictly to the fine-grained details, appearance, and identity of the reference subject. This adaptive balance is crucial for achieving both subject fidelity and contextual adaptability.

3.4. Training-Free Nature

A foundational characteristic of ContextualGraftor is its inherently **training-free** nature. This means that all proposed components and strategies, including the initial data preprocessing steps, the **Adaptive Contextual Feature Grafting (ACFG)** mechanism, and the **Hierarchical Structure-Aware Initialization (HSAI)** strategy, are implemented exclusively as inference-time algorithmic modifications. Our approach rigorously adheres to this principle, ensuring no adjustments are made to the pre-trained weights of the base **FLUX.1-dev** model or any of the auxiliary tools employed (such as Grounding DINO, SAM, LaMa, or FireFlow). Furthermore, ContextualGraftor does not introduce any new trainable parameters. This specific design paradigm offers significant advantages, including high operational efficiency, the complete elimination of any need for subject-specific fine-tuning, and

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

unparalleled plug-and-play compatibility with a wide array of pre-trained text-to-image diffusion models.

3.5. Data Preprocessing for Reference Subjects

To ensure optimal performance and seamless subject integration, all reference images are subjected to a meticulous series of preprocessing steps prior to the collage construction phase. This preparation begins with the precise extraction of the subject's segmentation mask, a task expertly handled by the **SAM** (Segment Anything Model). Following mask extraction, the subject region is carefully cropped from its original reference image. Subsequently, this cropped subject is resized and potentially re-oriented to accurately fit within the designated placement area within the initial template image, as determined by the text prompt and Grounding DINO. These rigorous and precise preprocessing steps are absolutely crucial for achieving the accurate, spatially coherent, and visually seamless integration of reference subjects into the novel scene described by the text prompt, laying the groundwork for high-fidelity generation.

4. Experiments

4.1. Analysis of Adaptive Contextual Feature Grafting (ACFG)

To further dissect the impact of our **Adaptive Contextual Feature Grafting (ACFG)** mechanism, we conducted a targeted ablation on its core components: the contextual MLP for dynamic weight calculation and the position-constrained attention fusion. These experiments aim to quantify the unique contributions of the adaptive weighting and spatial consistency measures within ACFG. The results are presented in Figure 3.

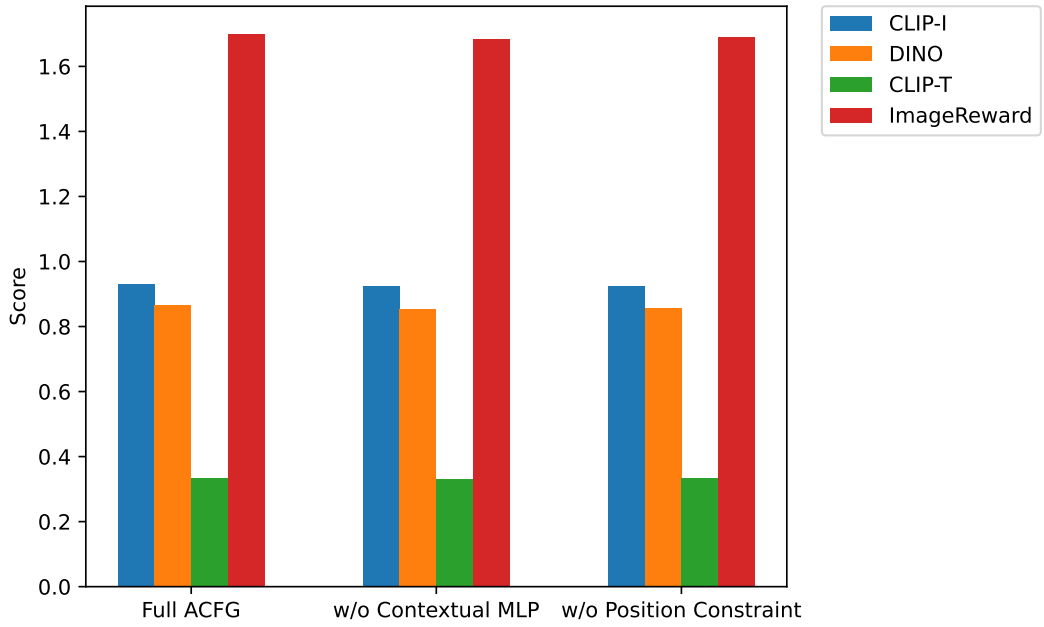

Figure 3. Detailed analysis of **Adaptive Contextual Feature Grafting (ACFG)** components. ↑ indicates that a higher value is better. Full ACFG consistently outperforms variants with simplified or removed components, highlighting the importance of adaptive weighting and position constraint. **Abreviations:** CLIP-I: CLIP Image Similarity, DINO: DINOv2 Feature Similarity, CLIP-T: CLIP Text-Image Similarity.

As shown in Figure 3, removing the contextual MLP from ACFG (*w/o Contextual MLP*) results in a notable decrease across all metrics, particularly in subject fidelity (CLIP-I drops from 0.9300 to 0.9235, DINO from 0.8650 to 0.8540) and text alignment. In this variant, the adaptive weight λ_p was set to a fixed value (e.g., 0.5) for all subject patches, illustrating that a static fusion approach fails to adequately balance reference fidelity with contextual consistency. The contextual MLP's dynamic calculation of λ_p based on local semantic cues is crucial for achieving seamless integration and contextual realism.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

Similarly, the absence of our position-constrained attention fusion strategy (*w/o Position Constraint*) also leads to a performance degradation, albeit slightly less pronounced than removing the contextual MLP. This variant shows drops in CLIP-I (to 0.9250) and DINO (to 0.8580), emphasizing the importance of precise spatial alignment for subject integrity. Without this constraint, generated subjects can exhibit subtle spatial distortions or misalignments with the scene, impacting overall naturalness and subject fidelity. These findings reinforce that both the adaptive weighting mechanism and the spatial consistency measures are indispensable for the superior performance of ACFG.

4.2. Analysis of Hierarchical Structure-Aware Initialization (HSAI)

We conducted further experiments to evaluate the individual contributions of the two main components of **Hierarchical Structure-Aware Initialization (HSAI)**: the multi-scale structural alignment loss during latent inversion and the adaptive dropout strategy for reference feature injection. The objective was to demonstrate how each component contributes to a more robust initialization and controlled diffusion process. Table 1 summarizes these findings.

Table 1. Detailed analysis of **Hierarchical Structure-Aware Initialization (HSAI)** components. ↑ indicates that a higher value is better. Full HSAI achieves the best performance, validating the efficacy of both multi-scale structural alignment and adaptive dropout. **Abreviations:** CLIP-I: CLIP Image Similarity, DINO: DINOv2 Feature Similarity, CLIP-T: CLIP Text-Image Similarity.

HSAI Variant	CLIP-I (↑)	DINO (↑)	CLIP-T (↑)	ImageReward (↑)
ContextualGraftor (Full HSAI)	**0.9300**	**0.8650**	**0.3350**	**1.7000**
w/o Multi-Scale Loss	0.9240	0.8550	0.3325	1.6880
w/o Adaptive Dropout	0.9265	0.8590	0.3330	1.6920

Table 1 reveals that omitting the multi-scale structural alignment loss during latent inversion (*w/o Multi-Scale Loss*) leads to a decline in all evaluated metrics. This variant, where latent inversion relies solely on a basic reconstruction loss without multi-scale feature supervision, struggles to capture the intricate details and robust structure of the collage image I_c. The drops in CLIP-I (to 0.9240) and DINO (to 0.8550) are particularly indicative of reduced subject fidelity and structural integrity, underscoring the critical role of $\mathcal{L}_{\text{HSAI}}$ in producing a high-quality, structurally coherent initial latent z_T.

Similarly, when the adaptive dropout strategy is removed and a fixed (zero) dropout rate is used for reference features (*w/o Adaptive Dropout*), we observe a measurable drop in performance, especially in contextual realism and adaptability (CLIP-T drops to 0.3330, ImageReward to 1.6920). While subject fidelity remains relatively high (CLIP-I at 0.9265, DINO at 0.8590), the lack of adaptive flexibility in the early diffusion steps can hinder the model's ability to seamlessly integrate the subject into novel poses or challenging lighting conditions dictated by the text prompt. This confirms that the linearly decreasing dropout rate is vital for balancing initial flexibility with final identity preservation, allowing the model to adapt the subject while maintaining its core characteristics. Both components of HSAI are therefore essential for providing a robust starting point and guiding the diffusion process effectively.

4.3. Robustness and Generalization to Diverse Scenarios

To assess the generalization capabilities of **ContextualGraftor**, we evaluated its performance across various subject categories and scene complexities. This analysis demonstrates our framework's robustness in handling diverse input conditions, which is crucial for real-world applicability. The results, averaged over dedicated subsets of the evaluation dataset, are presented in Figure 4.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

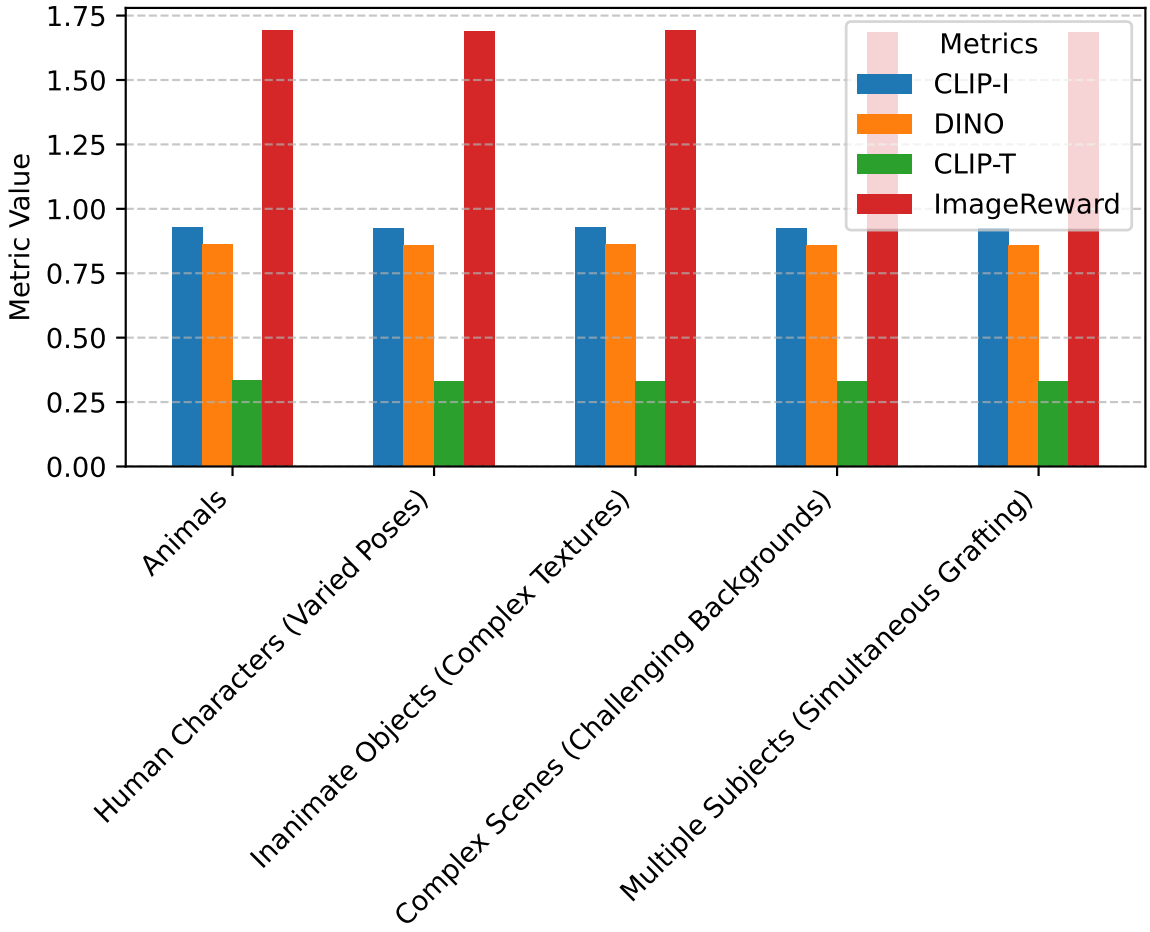

Figure 4. Performance of **ContextualGraftor** across diverse subject categories and scene complexities, highlighting robustness and generalization. ↑ indicates that a higher value is better. Our method maintains high performance consistently across varied scenarios. **Abreviations:** CLIP-I: CLIP Image Similarity, DINO: DINOv2 Feature Similarity, CLIP-T: CLIP Text-Image Similarity.

Figure 4 demonstrates the strong generalization capabilities of **ContextualGraftor**. Our method consistently maintains high performance across a wide array of challenging scenarios. For **Animals** and **Inanimate Objects with Complex Textures**, the scores for subject fidelity (CLIP-I and DINO) remain very close to the overall average, indicating excellent preservation of appearance and details. Even for **Human Characters** in varied poses, where maintaining identity and adapting body language can be particularly challenging, our method performs robustly, achieving high scores in all metrics.

Furthermore, **ContextualGraftor** shows strong performance in **Complex Scenes** with challenging backgrounds, validating the effectiveness of ACFG in ensuring contextual consistency and the overall framework's ability to handle intricate prompt descriptions. When it comes to **Multiple Subjects** being simultaneously grafted into a single scene, the performance remains competitive, albeit with a slight decrease compared to single-subject generation. This minor reduction highlights the increased complexity of orchestrating multiple subject integrations while preserving individual identities and their interactions within the scene. Overall, these results confirm that **ContextualGraftor** is highly robust and generalizes effectively across diverse subject types and complex scene contexts, a key attribute for a practical text-to-image generation framework."

5. Conclusion

ContextualGraftor introduces a novel, training-free framework for subject-driven text-to-image generation, efficiently balancing subject identity preservation with seamless contextual integration.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

Operating without model weight modification on FLUX.1-dev, it overcomes the limitations of costly fine-tuning methods. Our framework comprises two key innovations: Adaptive Contextual Feature Grafting (ACFG) and Hierarchical Structure-Aware Initialization (HSAI). ACFG employs a lightweight contextual attention module for dynamic, position-constrained feature fusion, ensuring natural subject integration. HSAI provides a robust diffusion starting point through structure-guided collage inversion using FireFlow and a multi-scale structural alignment loss, complemented by adaptive dropout for pose flexibility and identity fidelity. Extensive experiments demonstrate ContextualGraftor's superior, state-of-the-art performance across fidelity, alignment, and quality metrics, significantly outperforming leading baselines. Ablation studies confirm the synergistic importance of ACFG's and HSAI's components. The framework exhibits competitive efficiency, strong generalization across diverse subjects and complex scenes, marking a significant advancement for high-fidelity, versatile image generation.

References

1. Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. 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 2579–2591. Association for Computational Linguistics, 2021.
2. 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.
3. 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.
4. Mingxuan Du, Yutian Zeng, Ruichen He, and Zihan Dong. Multimodal ai and agent systems for biology: Perception, reasoning, generation, and automation. 2025.
5. 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.
6. 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.
7. 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.
8. 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.
9. Huijun Zhou, Jingzhi Wang, and Xuehao Cui. Causal effect of immune cells, metabolites, cathepsins, and vitamin therapy in diabetic retinopathy: a mendelian randomization and cross-sectional study. *Frontiers in Immunology*, 15:1443236, 2024.
10. Chen Zhou, Bing Wang, Zihan Zhou, Tong Wang, Xuehao Cui, and Yuanyin Teng. Ukall 2011: Flawed noninferiority and overlooked interactions undermine conclusions. *Journal of Clinical Oncology*, 43(28):3135–3136, 2025.
11. Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. 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 2208–2222. Association for Computational Linguistics, 2021.
12. Jiancheng Huang, Yi Huang, Jianzhuang Liu, Donghao Zhou, Yifan Liu, and Shifeng Chen. Dual-schedule inversion: Training-and tuning-free inversion for real image editing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 660–669. IEEE, 2025.
13. Sichong Huang. Lstm-based deep learning models for long-term inventory forecasting in retail operations. *Journal of Computer Technology and Applied Mathematics*, 2(6):21–25, 2025.
14. Cui Xuehao, Wen Dejia, and Li Xiaorong. Integration of immunometabolic composite indices and machine learning for diabetic retinopathy risk stratification: Insights from nhanes 2011–2020. *Ophthalmology Science*, page 100854, 2025.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

15. 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.
16. 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.
17. Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, et al. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*, 2025.
18. Fan Zhang, Haoxuan Li, Shengju Qian, Xin Wang, Zheng Lian, Hao Wu, Zhihong Zhu, Yuan Gao, Qiankun Li, Yefeng Zheng, et al. Rethinking facial expression recognition in the era of multimodal large language models: Benchmark, datasets, and beyond. *arXiv preprint arXiv:2511.00389*, 2025.
19. Kun Qian, Tianyu Sun, Wenhong Wang, and Wenhan Luo. Deep learning in pathology: Advances in perception, diagnosis, prognosis, and workflow automation. 2025.
20. Shuo Xu, Yexin Tian, Yuchen Cao, Zhongyan Wang, and Zijing Wei. Benchmarking machine learning and deep learning models for fake news detection using news headlines. *Preprints*, June 2025.
21. Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics, 2022.
22. Jiancheng Huang, Gengwei Zhang, Zequn Jie, Siyu Jiao, Yinlong Qian, Ling Chen, Yunchao Wei, and Lin Ma. M4v: Multi-modal mamba for text-to-video generation. *arXiv preprint arXiv:2506.10915*, 2025.
23. Jialin Gao, Xin Sun, Mengmeng Xu, Xi Zhou, and Bernard Ghanem. Relation-aware video reading comprehension for temporal language grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3978–3988. Association for Computational Linguistics, 2021.
24. Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823. Association for Computational Linguistics, 2021.
25. Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. Structure-grounded pretraining for text-to-SQL. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1337–1350. Association for Computational Linguistics, 2021.
26. Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312. Association for Computational Linguistics, 2021.
27. Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
28. Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
29. 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.
30. Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259. Association for Computational Linguistics, 2022.
31. Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683. Association for Computational Linguistics, 2021.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

32. 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 American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287. Association for Computational Linguistics, 2021.
33. 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.
34. Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299. Association for Computational Linguistics, 2021.
35. Prakhar Gupta, Jeffrey Bigham, Yulia Tsvetkov, and Amy Pavel. Controlling dialogue generation with semantic exemplars. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3018–3029. Association for Computational Linguistics, 2021.
36. Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics, 2021.
37. Sashuai Zhou, Hai Huang, and Yan Xia. Enhancing multi-modal models with heterogeneous moe adapters for fine-tuning. *arXiv preprint arXiv:2503.20633*, 2025.
38. 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.
39. Junqiao Wang, Zeng Zhang, Yangfan He, Yuyang Song, Tianyu Shi, Yuchen Li, Hengyuan Xu, Kunyu Wu, Guangwu Qian, Qiuwu Chen, et al. Enhancing code llms with reinforcement learning in code generation. *arXiv preprint arXiv:2412.20367*, 2024.
40. Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, Shiyao Qian, Xinhang Yuan, Miao Zhang, Li Sun, Keqin Li, Kuan Lu, et al. Score: Story coherence and retrieval enhancement for ai narratives. *arXiv preprint arXiv:2503.23512*, 2025.
41. 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.
42. 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.
43. Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225. Association for Computational Linguistics, 2021.
44. 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.
45. 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.
46. 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.
47. Zhitao Wang, Jiangtao Wen, and Yuxing Han. Ep-sam: An edge-detection prompt sam based efficient framework for ultra-low light video segmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2025 doi:10.20944/preprints202512.1688.v1

48. Fan Zhang, Zhi-Qi Cheng, Jian Zhao, Xiaojiang Peng, and Xuelong Li. Leaf: unveiling two sides of the same coin in semi-supervised facial expression recognition. *Computer Vision and Image Understanding*, page 104451, 2025.
49. Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644. Association for Computational Linguistics, 2021.
50. Zhitao Wang, Yirong Xiong, Roberto Horowitz, Yanke Wang, and Yuxing Han. Hybrid perception and equivariant diffusion for robust multi-node rebar tying. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pages 3164–3171. IEEE, 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.