TP-Blend: Textual-Prompt Attention Pairing for Precise Object-Style Blending in Diffusion Models

Xin Jin *felixxinjin@gmail.com*
GenPi Inc.

Yichuan Zhong *yichuanzhong27@gmail.com*
GenPi Inc.

Yapeng Tian *yapeng.tian@utdallas.edu*
The University of Texas at Dallas

Reviewed on OpenReview: *https://openreview.net/forum?id=q6M73uOBZE*

Abstract

Current text–conditioned diffusion editors handle single object replacement well but struggle when a new object and a new style must be introduced simultaneously. We present Twin-Prompt Attention Blend (TP-Blend), a lightweight training-free framework that receives two separate textual prompts, one specifying a blend object and the other defining a target style, and injects both into a single denoising trajectory. TP-Blend is driven by two complementary attention processors. Cross-Attention Object Fusion (CAOF) first averages head-wise attention to locate spatial tokens that respond strongly to either prompt, then solves an entropy-regularised optimal transport problem that reassigns complete multi-head feature vectors to those positions. CAOF updates feature vectors at the full combined dimensionality of all heads (e.g., 640 dimensions in SD-XL), preserving rich cross-head correlations while keeping memory low. Self-Attention Style Fusion (SASF) injects style at every self-attention layer through Detail-Sensitive Instance Normalization. A lightweight one-dimensional Gaussian filter separates low- and high-frequency components; only the high-frequency residual is blended back, imprinting brush-stroke-level texture without disrupting global geometry. SASF further swaps the Key and Value matrices with those derived from the style prompt, enforcing context-aware texture modulation that remains independent of object fusion. Extensive experiments show that TP-Blend produces high-resolution, photo-realistic edits with precise control over both content and appearance, surpassing recent baselines in quantitative fidelity, perceptual quality, and inference speed.

1 Introduction

Text-driven image editing with diffusion models Brack et al. (2024); Brooks et al. (2023); Sheynin et al. (2024); Mokady et al. (2023); Liu et al. (2024); Tumanyan et al. (2023); Chen et al. (2024); Avrahami et al. (2023); Ge et al. (2023); Shi et al. (2024); Deutch et al. (2024); Li et al. (2024b) has excelled at tasks like object replacement but still lacks a robust solution for object blending, where two objects must fuse seamlessly into a single coherent entity. Achieving such morphological transitions is challenging: the system must preserve each source object's defining characteristics (e.g., color, shape, texture) while synthesizing intermediate attributes that accurately reflect the intended blend. This capability is especially valuable in creative design, film production, product prototyping, and scientific or educational visualization, where smooth transitions (e.g., morphing a car into a spaceship or combining organisms to study evolutionary traits) are often essential.

Most style transfer methods still rely on reference images, limiting users to existing examples and requiring substantial effort Chung et al. (2024); Xing et al. (2024); Wang et al. (2024a); Xu et al. (2024); Li (2024); Lötzsch et al. (2022); Wang et al. (2023). By contrast, text-driven approaches Hertz et al. (2024); Zhang et al. (2023); Liu et al. (2023); Wu et al. (2024a) specify styles in natural language (e.g., "sketch-like," "art nouveau") and could offer greater flexibility, yet they remain underexplored.

Additionally, current style transfer techniques face major obstacles in achieving fine-grained, multi-scale, and region-specific control. They often fail to capture high-frequency textural details, losing subtle stylistic cues (e.g., brushstrokes, grain, intricate material features) even at high resolutions Chung et al. (2024); Xing et al. (2024), thereby compromising overall texture fidelity.

Motivated by these challenges, we propose Twin-Prompt Attention Blend (TP-Blend), a training-free framework that extends Classifier-Free Guided Text Editing (CFG-TE) Brack et al. (2024); Brooks et al. (2023); Sheynin et al. (2024); Mokady et al. (2023); Liu et al. (2024); Tumanyan et al. (2023); Chen et al. (2024); Avrahami et al. (2023); Ge et al. (2023) to support fine-grained object blending and style fusion through separate textual prompts, as illustrated in Figure 1. TP-Blend introduces two new modules: Cross-Attention Object Fusion (CAOF), which integrates features from a blend object prompt using attention maps and an Optimal Transport framework; and Self-Attention Style Fusion (SASF), which injects style via Detail-Sensitive Instance Normalization (DSIN) and replaces self-attention Key/Value matrices with those from the style prompt. Unlike prior image-based approaches, TP-Blend enables direct textual control of both content and style, offering precise and independent modulation of blending strength and texture details. By unifying object replacement, blending, and style transfer within a single denoising process, TP-Blend enhances controllability without incurring additional computational overhead.

Main Contributions. (1) Dual-Prompt Mechanism decouples object and style prompts, preventing interference and ensuring precise content representation and faithful style transfer within a unified denoising process; (2) CAOF with Optimal Transport aligns and integrates blend-object features into a replaced object by treating attention maps as distributions, enabling seamless morphological transitions and preserving semantic integrity; (3) SASF leverages DSIN to extract and transfer high-frequency style features, preserving intricate textural details without over-smoothing while allowing adaptive modulation of stylistic attributes across different spatial extents and granularities; (4) text-driven Key/Value substitution replaces self-attention Key/Value matrices with those derived from the style prompt, enforcing localized style modulation while maintaining spatial coherence and object fidelity.

Code Availability. Code is available at `https://github.com/felixxinjin1/TP-Blend`.

2 Related Work

Diffusion models have emerged as the dominant paradigm for text-guided image generation and editing, beginning with unconditional DDPMs and latent variants together with classifier-free guidance, which underpin early editing systems but leave open challenges in multi-concept disentanglement and precise regional control (Dhariwal & Nichol, 2021; Rombach et al., 2022; Ho & Salimans, 2022; Brooks et al., 2023; Brack et al., 2024). Unified generation and editing continues to broaden capability and instruction following, spanning single or dual diffusion formulations, universal dynamics-aware frameworks, adapter-based unification, and interactive or conversational editors that couple instruction tuning with in-context visual reasoning (Liu et al., 2025d; Yu et al., 2024; 2025a; Xia et al., 2024; Zhou et al., 2025b; Xia et al., 2025; Xiao et al., 2025b; Chen et al., 2025d; Duan et al., 2025; Fu et al., 2025b; Li et al., 2025h; Jia et al., 2025a; Lee et al., 2025b; Chen et al., 2025c; Xu et al., 2025; Zhou et al., 2025a; Li et al., 2025g; Wu et al., 2025d; Jia et al., 2025b; Lai et al., 2025; Mao et al., 2025a; Cao et al., 2025; Chen et al., 2025b; Luo et al., 2025a). Compositionality, multi-conditioning and regional control are advanced by balancing or transplanting representations, ordering and counting constraints, language-guided tokenization and unified tokenizers, as well as training-free or attention-repositioning guidance that improves spatial faithfulness and occlusion handling (Luo et al., 2025b; Jin et al., 2025; Cohen et al., 2024; 2025; Liang et al., 2025; Li et al., 2025d; Binyamin et al., 2025; Zeng et al., 2025; Hsiao et al., 2025; Qiu et al., 2025; Han et al., 2025c; Zhan & Liu, 2025; Wang et al., 2024b; Zha et al., 2025; Qu et al., 2025b; Wu et al., 2024b). Personalization and identity preservation leverage feature

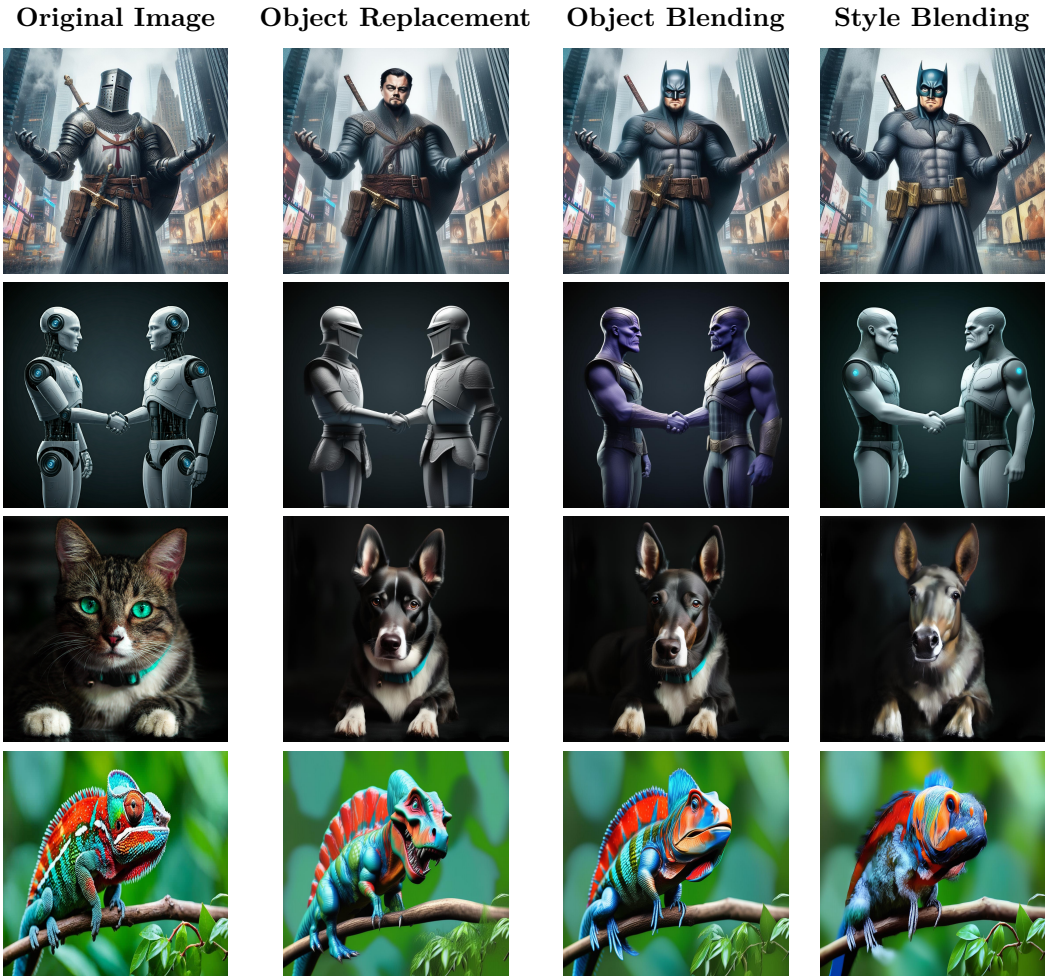

Figure 1: Demonstration of our method's capabilities. Row 1: Original object "**Knight**" is replaced by "**Leonardo DiCaprio**", blended with "**Batman**", and styled with "**Pop Art**". Row 2: Original object "**Robot**" is replaced by "**Knight**", blended with "**Thanos**", and styled with "**Cyberpunk Style**". **Row 3**: Original object "**Cat**" is replaced by "**Dog**", blended with "**Horse**", and styled with "**Oil painting**". **Row 4**: Original object "**Chameleon**" is replaced by "**Dinosaur**", blended with "**Fish**", and styled with "**Oil painting**".

caching and standardize-then-personalize pipelines, continual concept learning with mitigation of forgetting and confusion, parameter-efficient or time-aware frequency guidance, consistency from a single prompt and localized attention, LoRA merging and hypernetworks, and targeted human preference optimization (Aiello et al., 2025; Xie et al., 2025a; Guo & Jin, 2025; Li et al., 2025c; Liu et al., 2025c; Huang et al., 2025c; Liu et al., 2025e; Xiao et al., 2025a; Shenaj et al., 2024; Na et al., 2025; Zhu et al., 2025a; Li et al., 2025e; Wang et al., 2025d; Liu et al., 2025f). Style control has moved from exemplar dependence toward text-driven and detail-faithful modulation through state-space modeling, selective style element control, object-centric style editing, causal intervention and conflict-free guidance, including emotional manipulation (Liu et al., 2025a; Lei et al., 2025; Park et al., 2025b; Huang et al., 2025e; Jo et al., 2025; Yang et al., 2025d; Cai et al., 2024a; Huang et al., 2025d; Dang et al., 2025). Efficiency and scheduling improve with learned time prediction, one-step distillation and time-independent encoders, few-step or lightning editors, second-order sampling, fast inpainting and latent super-resolution, together with scale-down text encoders (Ye et al., 2024; 2025; Luo et al., 2024; Nguyen et al., 2024; 2025b;a; Chadebec et al., 2025; Wang et al., 2025a; Xie et al., 2025b; Jeong et al., 2025; Li et al., 2025e; Wang et al., 2025c). Scaling and backbone design target long-range coherence and high resolution via hierarchical transformers, frequency or attention modulation and resolution-agnostic

diffusion, while autoregressive alternatives push tokenization and decoding with large tokenizers, holistic or language-guided tokenizers, deep-compression hybrids and frequency-aware modeling (Zhang et al., 2025a; Voronov et al., 2024; Yellapragada et al., 2024; Yang et al., 2025b; Shi et al., 2025b; Wang et al., 2025b; Han et al., 2025b; Kumbong et al., 2025; Han et al., 2025a; Wu et al., 2025b; Xiong et al., 2025; Zheng et al., 2025a; Zha et al., 2025; Qu et al., 2025b; Wu et al., 2025e; Chen et al., 2025e; Zheng et al., 2025b; Qu et al., 2025a; So et al., 2025). Domain and structure-aware generation spans design and line-art images, creative and degraded layouts with cycle consistency, unified layout planning, grounded instance-level control and multi-view consistency, scene-consistent camera control, controllable street views, variable multi-layer transparency and layer decomposition, and decoupled inter and intra element conditions (Wang et al., 2025g; 2024c; Zhang et al., 2024; Cai et al., 2024b; He et al., 2025; Wang et al., 2025e; Wu et al., 2024b; Huang et al., 2024; Yuan et al., 2025; Gu et al., 2025; Pu et al., 2025; Yang et al., 2025e;c; Wu et al., 2025a). Safety, fairness and provenance are studied through typographic threats, trigger-free branding attacks, robust watermarking, anti-editing and localized concept erasure, guideline-token safety, fair mapping and debiasing with prompt optimization or entanglement-free attention, along with human-centric quality metrics (Cheng et al., 2025; Jang et al., 2025; Wang et al., 2025i;h; Lee et al., 2025a; Li et al., 2025a;b; Um & Ye, 2025; Kim et al., 2025b; Park et al., 2025a; Wang et al., 2025f; Huang et al., 2025b). Evaluation and analysis expand with enhanced compositional benchmarks, scalable human-aligned editing assessment, trade-off studies and probabilistic analyses, and work on paragraph-level grounding and attribute-centric composition (Huang et al., 2025a; Ryu et al., 2025; Zhang et al., 2025d; Yu et al., 2025b; Cong et al., 2025; Wu et al., 2025d; Xuan et al., 2025; Zhu et al., 2025b;c; Ren et al., 2025). Beyond images, image-to-video and motion-controlled generation integrate text-to-image with text-to-video models or decouple motion intensity for one-step or high-quality synthesis (Shi et al., 2025a; Mao et al., 2025b; Su et al., 2025). Practical deployment advances include mobile-oriented high-resolution generation (Chen et al., 2025a). Additional controls include quantity perception, dynamic-feedback regulation, region-aware fine-tuning, collaborative attention, proportional group representations, zero-shot multi-attribute creation, effective and diverse prompt sampling, layer-wise memory, semantic-faithful noise diffusion and spatial attention repositioning, plus zero-shot subject-driven generation by large inpainting models (Li et al., 2025f; Fu et al., 2025a; Xing et al., 2025; Yang et al., 2025a; Jung et al., 2025; Deng et al., 2025; Yun et al., 2025; Kim et al., 2025a; Miao et al., 2025; Han et al., 2025c; Shin et al., 2025). Integration with large language models supports planning, reward shaping and stepwise verification through collaborative or customized multimodal rewards, chain-of-thought reinforcement and preference optimization for alignment and controllability (Liu et al., 2025b; Tang et al., 2025; Ba et al., 2025; Zhou et al., 2025c; Jiang et al., 2025; Zhang et al., 2025b; Sun et al., 2025; Wu et al., 2025c; Guo et al., 2025; Fang et al., 2025). Complementary representations and training-free guidance broaden the modeling space with triplanes, rectified flows, dense-aligned guidance, Stable Flow and 3D-aware scene modeling, together with hyperbolic diffusion autoencoders, unified self-supervised pretraining and industrial anomaly generation (Bilecen et al., 2024; Dalva et al., 2024; Wang et al., 2025j; Avrahami et al., 2024; 2025; Zhang et al., 2025c; Li et al., 2024a; Chu et al., 2025; Dai et al., 2024; Cai et al., 2025b). Our work is situated within these trends while retaining links to foundational and domain-specific efforts such as Design-Diffusion, LineArt, Focus-N-Fix, Type-R and PreciseCam and contemporaneous advances in layout, control and training-free customization (Wang et al., 2025g; 2024c; Xing et al., 2025; Shimoda et al., 2024; Bernal-Berdun et al., 2025; Zhang et al., 2025a; Cai et al., 2025a). For completeness, we note additional concurrent directions spanning instruction-following editors, multimodal integration, compositional sliders and broader analyses (Cao et al., 2025; Chen et al., 2025b; Luo et al., 2025a; Qu et al., 2025a; So et al., 2025; Ren et al., 2025; Zhu et al., 2025c;b).

3 Proposed Method

3.1 Preliminaries

Diffusion models Dhariwal & Nichol (2021); Rombach et al. (2022) progressively denoise a latent variable to produce high-fidelity images. Classifier-Free Guidance (CFG) Ho & Salimans (2022) steers generation toward conditioning inputs by interpolating between conditional and unconditional noise predictions. Specifically, the model is trained to predict both $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$ and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c)$, where c is the conditioning. At inference, a guidance

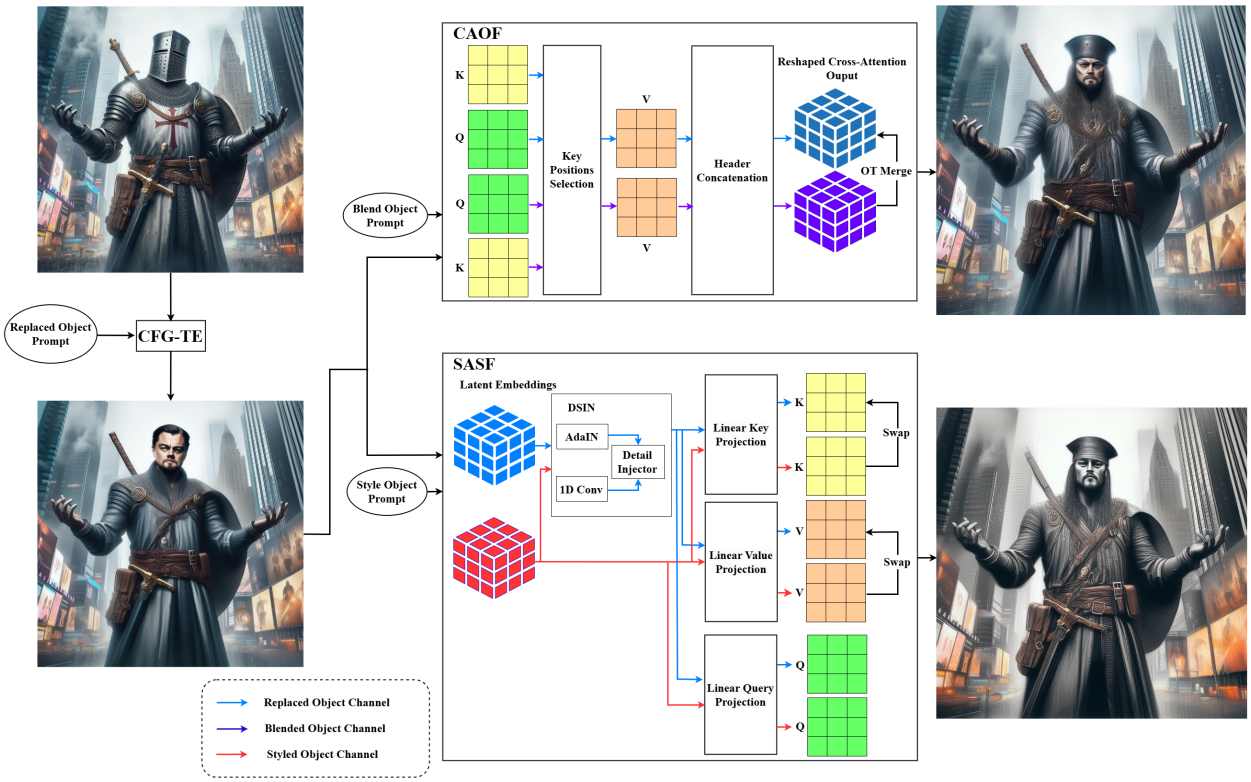

Figure 2: Flowchart of TP-Blend, integrating object replacement, blending, and style transfer within the diffusion process. In this example, the original object "**Knight**" is replaced by "**Leonardo DiCaprio**", blended with "**Captain Jack Sparrow**", and styled with a "**Charcoal Drawing**" effect.

scale s_g modifies the predicted noise:

$$\tilde{\epsilon}_\theta(\mathbf{x}_t) \;=\; \epsilon_\theta(\mathbf{x}_t) \;+\; s_g\big(\epsilon_\theta(\mathbf{x}_t, c) - \epsilon_\theta(\mathbf{x}_t)\big). \tag{1}$$

CFG-TE extends CFG to perform precise edits on an existing image $\mathbf{x}_0$. The image is inverted to a latent $\mathbf{x}_T$ via DDIM inversion Song et al. (2020), which defines a deterministic mapping between timesteps when no guidance is applied:

$$\mathbf{x}_{t-1} \;=\; \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}}\right) \;+\; \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(\mathbf{x}_t), \tag{2}$$

where α_t is the noise schedule. Once inverted, the noise prediction at each denoising step can be modified to remove or add concepts:

$$\tilde{\epsilon}_\theta^{\text{edit}}(\mathbf{x}_t) \;=\; \epsilon_\theta(\mathbf{x}_t) \;+\; s_e\,\Delta\epsilon_\theta(\mathbf{x}_t), \tag{3}$$

with

$$\Delta\epsilon_\theta(\mathbf{x}_t) = \begin{cases} \epsilon_\theta(\mathbf{x}_t, c_{\text{edit}}) - \epsilon_\theta(\mathbf{x}_t), & \text{(positive guidance)}, \\ \epsilon_\theta(\mathbf{x}_t) - \epsilon_\theta(\mathbf{x}_t, c_{\text{edit}}), & \text{(negative guidance)}, \end{cases} \tag{4}$$

where s_e is the edit guidance scale and c_{edit} is the editing prompt.

3.2 Twin-Prompt Attention Blend

CFG-TE enables object replacement by applying positive guidance to the new object prompt and negative guidance to the original. However, it lacks mechanisms for fine-grained *object blending* and *style fusion*, which require compositional mixing and textural transformations.

We introduce TP-Blend, extending CFG-TE with two additional prompts: a blend prompt and a style prompt, both assigned zero edit guidance to avoid interfering with object replacement. Cross-Attention Object Fusion (CAOF) integrates blend object features at key spatial positions using a unified attention map and an Optimal Transport framework. Self-Attention Style Fusion (SASF) modulates texture and style by locally adjusting feature statistics using DSIN and substituting the Key/Value matrices with those derived from the style prompt. By decoupling both blending and style transfer from the editing guidance scale, TP-Blend integrates seamlessly into the denoising process, enhancing CFG-TE's capabilities with high-fidelity object and style blending (Fig. 2).

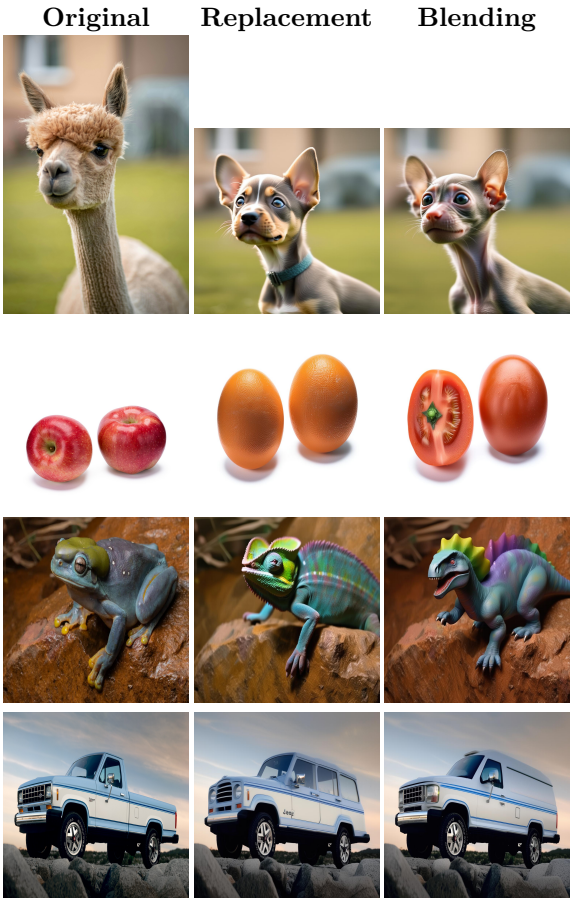

Figure 3: CAOF Object Blending across different sets. Row 1: Original object "**alpaca**" is replaced by "**puppy**" and blended with "**monkey**". Row 2: Original "**apple**" is replaced by "**orange**" and blended with "**tomato**". Row 3: Original "**frog**" is replaced by "**chameleon**" and blended with "**dinosaur**". **Row 4**: Original "**truck**" is replaced by "**jeep**" and blended with "**ambulance**".

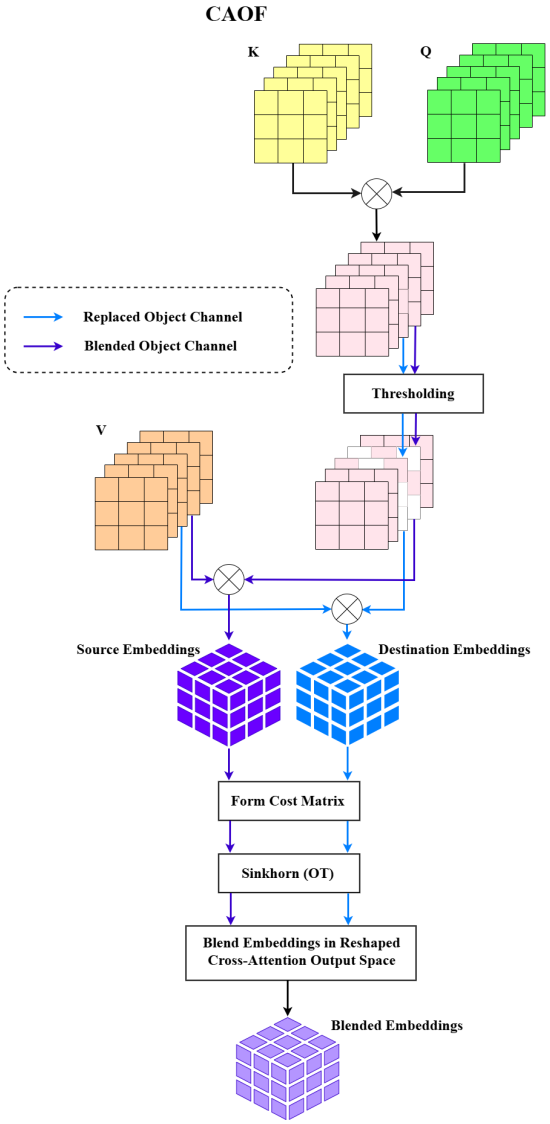

Figure 4: CAOF Flowchart: Cross-Attention Object Fusion merges the blend object's features into the replaced object by identifying key spatial positions in the attention maps and applying an optimal transport framework for coherent morphological transitions.

3.3 Cross-Attention Object Fusion

As shown in Figure 3 and summarized in Figure 4, CAOF seamlessly integrates a blend object's features into a replaced object during the diffusion process. Leveraging textual prompts for both the replaced and blend objects, CAOF locates key spatial regions in cross-attention maps and employs an Optimal Transport (OT) framework to determine blending levels.

Identifying Significant Positions in Cross-Attention Maps. In multi-head cross-attention Vaswani (2017), each head h produces attention weights

$$\mathbf{A}^{(h)} = \mathrm{softmax}\Big(\frac{\mathbf{Q}^{(h)}\,\mathbf{K}^{(h)\top}}{\sqrt{d_k}}\Big), \tag{5}$$

where $\mathbf{Q}^{(h)} \in \mathbb{R}^{N\times d_k}$ and $\mathbf{K}^{(h)} \in \mathbb{R}^{M\times d_k}$ are query/key matrices, N is the number of spatial positions, M is the number of text tokens, and d_k is the head dimension. We average over H heads and focus on the replaced and blend object tokens, t_{replaced} and t_{blend}:

$$\mathbf{a}_{\text{replaced}} = \tfrac{1}{H}\sum_{h=1}^{H}\mathbf{A}^{(h)}_{:\,,\,t_{\text{replaced}}}, \quad \mathbf{a}_{\text{blend}} = \tfrac{1}{H}\sum_{h=1}^{H}\mathbf{A}^{(h)}_{:\,,\,t_{\text{blend}}}. \tag{6}$$

To identify meaningful spatial positions, we introduce two percentile thresholds, τ_{source} and τ_{dest}. Specifically, any position i in $\mathbf{a}_{\text{blend}}$ whose attention weight exceeds the τ_{source}-percentile is included in the source set $\mathcal{S}$, and any position in $\mathbf{a}_{\text{replaced}}$ exceeding the τ_{dest}-percentile is placed in the destination set $\mathcal{D}$.

Clarification. The thresholds act on different head-averaged maps and therefore induce the two index sets independently. We make this explicit with

$$\mathcal{S} = \big\{\, i \in [N] : \mathbf{a}_{\text{blend}}[i] \geq q_{\tau_{\text{source}}}\big(\mathbf{a}_{\text{blend}}\big) \,\big\}, \qquad \mathcal{D} = \big\{\, i \in [N] : \mathbf{a}_{\text{replaced}}[i] \geq q_{\tau_{\text{dest}}}\big(\mathbf{a}_{\text{replaced}}\big) \,\big\}, \tag{7}$$

where $q_\tau(\cdot)$ denotes the τ-percentile. The sets $\mathcal{S}$ and $\mathcal{D}$ are not a partition of the image tokens. A position can be in neither set when it is below both cutoffs, in exactly one set when it responds strongly to only one prompt, or in both sets when it responds strongly to both prompts. In the fusion step (Sec. 3.3), only destination positions $d \in \mathcal{D}$ are updated and they receive transported features from sources $s \in \mathcal{S}$ under the OT plan $\mathbf{T}$ (Eq. 11 and Eq. 9). Tokens $i \notin \mathcal{D}$ pass through unchanged. If a spatial location lies in $\mathcal{S} \cap \mathcal{D}$, its destination slot taken from the replaced stream can still import features from its source slot taken from the blend stream because these embeddings originate from different prompt branches which avoids ambiguity. When an object phrase spans multiple text tokens, we pool their columns before thresholding and use the mean by default, while max pooling yields similar behavior in our experiments. Although we often tie the thresholds in a joint setting with $\tau_{\text{source}} = \tau_{\text{dest}}$ to simplify usage (see Fig. 12), they are defined separately and can be chosen differently to trade precision and coverage. The percentile formulation makes the selection scale-free and robust across layers and prompts because it depends on rank within each map rather than absolute magnitude.

Blending Feature Embeddings in Reshaped Cross-Attention Outputs. To effectively integrate features from the blend object into the replaced object, we begin by concatenating the per-head attention outputs along the feature dimension:

$$\mathbf{O} \;=\; \mathrm{Concat}_{h=1}^{H}\big(\mathbf{A}^{(h)}\mathbf{V}^{(h)}\big) \;\in\; \mathbb{R}^{N\times D}, \tag{8}$$

where $\mathbf{A}^{(h)} \in \mathbb{R}^{N\times M}$ are attention weight matrices, $\mathbf{V}^{(h)} \in \mathbb{R}^{M\times d_k}$ are the corresponding value matrices, $D = H \cdot d_k$ is the total feature dimensionality, N is the number of query positions, and M is the number of key tokens. By consolidating multi-head outputs into a single representation, we preserve all information necessary for seamless fusion, avoiding the loss that would occur from per-head embeddings.

We then blend the feature vectors of the replaced object with those of the blend object under a transport plan $\mathbf{T}$. Specifically, if $d_i \in \mathcal{D}$ and $s_j \in \mathcal{S}$ denote destination and source positions respectively, with $\mathbf{f}_{d_i}, \mathbf{f}_{s_j} \in \mathbb{R}^D$

being their respective feature vectors from $\mathbf{O}$, the updated feature vector at position d_i becomes

$$\mathbf{f}'_{d_i} \;=\; (1-w_0)\,\mathbf{f}_{d_i} \;+\; w_0 \sum_{s_j\in\mathcal{S}} \frac{T_{ij}}{\sum_{s_k\in\mathcal{S}} T_{ik}}\,\mathbf{f}_{s_j}, \tag{9}$$

where $w_0 \in [0,1]$ controls the relative influence of the blend features, and T_{ij} is obtained by solving the **OT problem**. By treating the multi-head outputs as a whole at the full dimensionality (e.g., $D = 640$), we not only preserve complex content and style cues but also obtain a more manageable OT cost matrix (e.g., 4096×4096), avoiding the significantly larger matrices (e.g., 40960×40960) that would result from per-head processing.

Formulating the Optimal Transport Problem. Let $\mathcal{S}$ and $\mathcal{D}$ denote the sets of source (blend object) and destination (replaced object) positions. The cost of transporting mass from source position $j \in \mathcal{S}$ to destination position $i \in \mathcal{D}$ is given by

$$C_{ij} = \lambda_{\text{feature}}\, D_{\text{feature}}(i,j) \;+\; \lambda_{\text{spatial}}\, D_{\text{spatial}}(i,j), \tag{10}$$

where $D_{\text{feature}}(i,j)$ is the cosine distance between feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ and $D_{\text{spatial}}(i,j)$ is the Euclidean distance between their spatial coordinates.

We solve the entropic OT problem:

$$\min_{\mathbf{T}\geq 0} \quad \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{S}} T_{ij}C_{ij} - \gamma H(\mathbf{T}), \tag{11}$$

$$\text{s.t.} \quad \sum_{j\in\mathcal{S}} T_{ij} = 1, \quad \forall i \in \mathcal{D}, \tag{12}$$

$$\sum_{i\in\mathcal{D}} T_{ij} \geq \frac{1}{|\mathcal{S}|}, \quad \forall j \in \mathcal{S}, \tag{13}$$

where $H(\mathbf{T}) = -\sum_{i,j} T_{ij} \log T_{ij}$ is the entropy term, and $\gamma > 0$ is the regularization parameter. Entropy regularization promotes smoother transport mass across source-destination pairs.

Solving the Optimal Transport Problem with the Sinkhorn Algorithm. The entropic regularization allows the problem to be efficiently solved using the Sinkhorn algorithm Cuturi (2013); Peyré et al. (2019); Genevay et al. (2016). We form the Gibbs kernel $\mathbf{K} = \exp(-\mathbf{C}/\gamma)$ and iteratively update scaling vectors $\mathbf{u}\in\mathbb{R}^{|\mathcal{D}|}$ and $\mathbf{v}\in\mathbb{R}^{|\mathcal{S}|}$:

$$\mathbf{u}^{(k+1)} \;=\; \frac{\mathbf{1}_{|\mathcal{D}|}}{\mathbf{K}\,\mathbf{v}^{(k)}}, \quad \mathbf{v}^{(k+1)} \;=\; \frac{\frac{1}{|\mathcal{S}|}\,\mathbf{1}_{|\mathcal{S}|}}{\mathbf{K}^\top\,\mathbf{u}^{(k+1)}}, \tag{14}$$

until convergence. The transport plan becomes

$$\mathbf{T} \;=\; \operatorname{diag}(\mathbf{u})\,\mathbf{K}\,\operatorname{diag}(\mathbf{v}). \tag{15}$$

Finally, we use $\mathbf{T}$ to blend each destination feature with weighted contributions from the source. Reintegrating these blended features into the cross-attention outputs yields a naturally fused object that inherits characteristics of the blend object at selected positions, with minimal overhead or artifacts.

3.4 Self-Attention Style Fusion

As illustrated in Figure 5 and outlined in Figure 6, SASF integrates style and texture into the replaced object through self-attention. Compared to previous methods Chung et al. (2024); Xing et al. (2024); Wang et al. (2024a); Xu et al. (2024); Li (2024); Lötzsch et al. (2022); Hertz et al. (2024); Zhang et al. (2023); Wang et al. (2023), SASF offers four advantages: (1) it introduces DSIN to capture HF textural details in a lightweight yet effective manner; (2) it relies on simple textual prompts rather than style images; (3) it fuses style and object features simultaneously during denoising, preserving both content fidelity and style

coherence; and (4) By translating historical idioms such as *Ukiyo-e*, *Renaissance*, and *Baroque* into their own material vocabularies, SASF can recode fabric weave, ornamentation, and weaponry; chain mail shifts to brocaded velvet, a plain sword strap becomes an obi sash, yet the figure's stance and the surrounding cityscape stay unchanged, as demonstrated in Figure 7.

Figure 5: Object blending enhanced with various artistic styles. **Style 1**: Pixel Art; **Style 2**: Chocolate; **Style 3**: Charcoal Drawing; **Style 4**: Oil Painting.

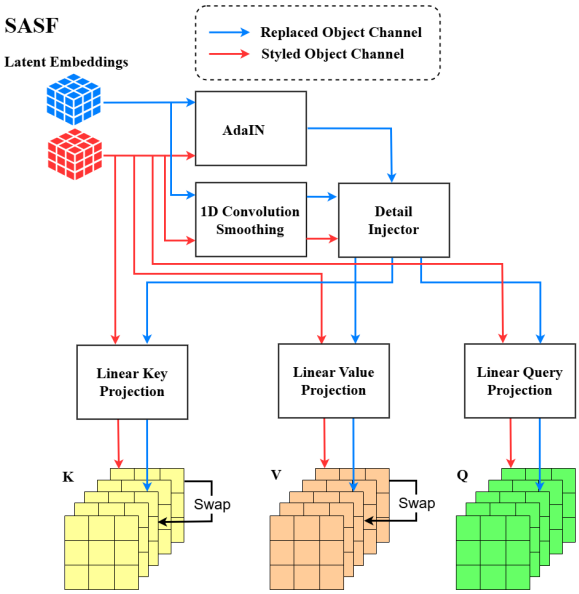

Figure 6: SASF Flowchart: Self-Attention Style Fusion incorporates style prompts by injecting high-frequency details via DSIN and substituting textual Key/Value matrices, ensuring fine-grained style modulation during the diffusion process.

Figure 7: **Stylistic renderings reshape fabric texture and accessories while pose and setting remain unchanged.** The original **knight** is replaced by **Albert Einstein**, blended with a **nobleman** concept, and then rendered in four distinct styles. Each style reinterprets the garments in a unique way: **Ukiyo-e** replaces the surcoat with a patterned kimono, complete with an obi sash and a lacquered katana; **Renaissance** introduces brocaded velvet, gilt medallions, and a scholar's cap; **Baroque** presents deep hued silk enriched with heavy gold embroidery and filigreed weaponry; **Low-Poly** abstracts every surface into planar facets and simplifies folds and metallic highlights.

Detail-Sensitive Instance Normalization. Let $\mathbf{F}_{\text{replaced}}, \mathbf{F}_{\text{style}} \in \mathbb{R}^{N \times D}$ be the latent embeddings (i.e., token-wise feature maps) of the replaced and style objects, respectively. We first perform an AdaIN step on

the replaced features:

$$\mathbf{F}'_{\text{replaced}} = \left(\frac{\mathbf{F}_{\text{replaced}} - \mu_{\text{rep}}}{\sigma_{\text{rep}}} \right) \sigma_{\text{style}} + \mu_{\text{style}}, \tag{16}$$

where $(\mu_{\text{rep}}, \sigma_{\text{rep}})$ and $(\mu_{\text{style}}, \sigma_{\text{style}})$ are the channel-wise means and standard deviations of the replaced and style embeddings. This aligns global statistics (mean and variance) to match the target style, but by itself may overlook subtle, higher-frequency stylistic cues.

Next, DSIN applies a small 1D Gaussian smoothing filter along the token dimension to decompose both $\mathbf{F}_{\text{replaced}}$ and $\mathbf{F}_{\text{style}}$ into low-frequency (LF) and high-frequency (HF) components:

$$\mathbf{F}^{\text{LF}} = \mathbf{F} * \mathbf{K}, \quad \mathbf{F}^{\text{HF}} = \mathbf{F} - \mathbf{F}^{\text{LF}}, \tag{17}$$

where $\mathbf{K}$ is a 1D Gaussian kernel of size $k = 2m + 1$ and width σ. Intuitively, $\mathbf{F}^{\text{LF}}$ captures coarse variations (slower changes across tokens), while $\mathbf{F}^{\text{HF}}$ isolates the finer details. DSIN then injects a fraction α of the style HF difference directly into the AdaIN output:

$$\mathbf{F}''_{\text{replaced}} = \mathbf{F}'_{\text{replaced}} + \alpha \left(\mathbf{F}^{\text{HF}}_{\text{style}} - \mathbf{F}^{\text{HF}}_{\text{replaced}} \right). \tag{18}$$

When DSIN applies a 1D Gaussian kernel $\mathbf{K}$ along the token dimension, it acts as a low-pass filter in the frequency domain: larger σ broadens the kernel's passband, yielding a narrower high-frequency (HF) residual $\mathbf{F}^{\text{HF}}$ and thus a subtler style injection. Conversely, smaller σ captures more mid- and high-frequency components, accentuating textural details (e.g., brushstrokes) in the final output. The injection fraction α then scales the amplitude of these style-specific HF cues. In effect, σ and α together provide a powerful mechanism for tuning the granularity and prominence of style features.

Unlike prior approaches such as Huang & Belongie (2017) or Chung et al. (2024) that apply AdaIN globally or only at the initial noise level for DDIM inversion, our DSIN is applied at *every self-attention layer* throughout the denoising process. This repeated application ensures the progressive and layer-wise infusion of fine-grained stylistic features, enabling multi-scale texture adaptation without disrupting the overall structure.

Key/Value Substitution. Following the DSIN framework, we first construct the Query, Key, and Value matrices for self-attention. We then substitute the Key and Value channels of the target (replaced) object with those of the style source:

$$\mathbf{K}_{\text{tar}} \leftarrow \mathbf{K}_{\text{sty}}, \qquad \mathbf{V}_{\text{tar}} \leftarrow \mathbf{V}_{\text{sty}}. \tag{19}$$

Since the self-attention output is computed by weighting the Value vectors using Query-Key dot products, replacing the Key and Value matrices of the replaced region with those from the style prompt allows style features to dominate the attention updates. This substitution imposes the texture and local patterns of the style onto the replaced object, leading to strong stylistic transformations.

While Chung et al. (2024) apply this substitution using Key/Value representations extracted from an image-based style encoder, our approach instead derives these from textual prompts. Specifically, we construct the Key/Value matrices from the text prompts of both the replaced object and the style source, enabling a text-driven style transfer mechanism without requiring image-based features.

Importantly, although this substitution offsets the effect of DSIN modulation in the Key/Value branches for the replaced object (since it is overwritten by style-derived features), DSIN-modified features remain intact in the Query branch. This asymmetry allows DSIN to still influence the attention outputs via its role in computing attention scores. Consequently, high-frequency stylistic cues injected through DSIN continue to impact the hidden embeddings passed to the next layer. This achieves a dual effect: the Key/Value substitution enforces stylistic consistency, while DSIN-enhanced Queries preserve the structural fidelity of the replaced object, allowing nuanced and locally-aware style transfer.

Table 1: Performance on the Object Replacement + Object Blending task.

Method	BOM↑	CLIP_R↑	CLIP_B↑	$1-\text{LPIPS}_O$↑
IP2P Brooks et al. (2023) (CVPR 2023)	0.1075	0.1819	0.2708	0.5887
StyleAligned Hertz et al. (2024) (CVPR 2024)	0.2371	0.2120	0.2866	0.5814
TurboEdit Deutch et al. (2024)	0.3199	0.1984	0.2781	0.6125
FLUX.1 Kontext Batifol et al. (2025)	0.3401	0.2013	0.2898	0.6007
LEDITS++ Brack et al. (2024) (CVPR 2024)	0.3913	0.2078	0.2834	0.6145
SeedEdit Shi et al. (2024)	0.5486	0.2096	0.2966	0.6381
Step1X-Edit Liu et al. (2025d)	0.7150	0.2120	0.2913	0.7024
Blended diffusion Avrahami et al. (2022) (CVPR 2022)	0.7241	0.2026	0.2910	0.7568
Nano Banana	0.7324	0.2159	0.2866	0.7303
CAOF	**0.8031**	0.2014	0.2937	0.8292

4 Experiments

4.1 Implementation Details

Model Architecture. All experiments employ SD-XL Podell et al. (2023) as the diffusion backbone. The source image is first inverted to a latent $\mathbf{x}_T$ via DDIM inversion, guaranteeing exact reconstruction before editing. During the forward denoising pass we apply, at every timestep: (i) TIE-CFG for object replacement (positive guidance on the target prompt, negative on the original); (ii) CAOF to transport blend-object features into attention positions selected by the joint percentile thresholds $\tau_{\text{source}} = \tau_{\text{dest}} \in \{0.6, 0.7\}$; and (iii) SASF to inject style via DSIN and key–value substitution. The Sinkhorn regulariser is fixed to $\gamma = 0.1$, with cost weights $\lambda_{\text{feature}} = 0.7$ and $\lambda_{\text{spatial}} = 0.3$ (Eq. 10).

Baseline Methods. To isolate the contribution of TP-Blend, we compare against six state-of-the-art text-driven editors and re-tune their prompts for each task so that every method receives semantically equivalent conditioning. The baselines are Step1X-Edit Liu et al. (2025d), SeedEdit Shi et al. (2024), LEDITS++ Brack et al. (2024) (CVPR 2024), StyleAligned Hertz et al. (2024) (CVPR 2024), TurboEdit Deutch et al. (2024), IP2P Brooks et al. (2023),(CVPR 2023), Blended diffusion Avrahami et al. (2022),(CVPR 2022), FLUX.1 Kontext Batifol et al. (2025), and Nano Banana.

SeedEdit and Step1X-Edit are inversion-free decoders optimised for speed, LEDITS++ and StyleAligned specialise in resolution-aware refinement, while TurboEdit and IP2P are two-stage pipelines that first predict a coarse edit mask. Blended diffusion performs training-free, CLIP-guided local edits via progressive noise-space blending for background preservation. FLUX.1 Kontext is a unified flow-matching in-context editor with fast multi-turn consistency. Nano Banana is a lightweight interactive editor focused on fast object and style edits with identity preservation. Evaluating against this diverse slate highlights TP-Blend's ability to blend rather than merely replace or stylise.

Evaluation Protocol. For our evaluation, we assembled a diverse set of high-resolution, publicly available images from Unsplash[1], following the same practice as prior work such as SLIDE Jampani et al. (2021) and Text-driven Image Editing via Learnable Regions Lin et al. (2024). The test dataset consists of 4,000 samples, created by pairing 40 base images with 20 distinct replace-blend object combinations and 5 distinct blend styles.

Evaluation Metrics. We assess alignment between generated image I_g and four textual prompts—original object P_O, replaced object P_R, blend object P_B, and style P_S—using CLIP similarity:

$$\text{CLIP}_x = \cos\big(f_{\text{vis}}(I_g),\ f_{\text{text}}(P_x)\big), \quad x \in \{O, R, B, S\}.$$

Perceptual fidelity is quantified as $1-\text{LPIPS}_O$. To ensure comparability, each score s is min–max normalized:

[1] https://unsplash.com/

Table 2: Performance on the full Object Replacement + Object & Style Blending task.

Method	BOSM↑	CLIP$_R$↑	CLIP$_B$↑	CLIP$_S$↑
IP2P Brooks et al. (2023) (CVPR 2023)	0.1277	0.1680	0.2776	0.1694
LEDITS++ Brack et al. (2024) (CVPR 2024)	0.2693	0.2039	0.2876	0.2236
TurboEdit Deutch et al. (2024)	0.3829	0.1954	0.2820	0.2090
FLUX.1 Kontext Batifol et al. (2025)	0.3955	0.2069	0.2942	0.2118
StyleAligned Hertz et al. (2024) (CVPR 2024)	0.4125	0.1888	0.2915	0.1973
SeedEdit Shi et al. (2024)	0.4650	0.1963	0.2915	0.2017
Step1X-Edit Liu et al. (2025d)	0.4652	0.2145	0.2920	0.2170
Blended diffusion Avrahami et al. (2022) (CVPR 2022)	0.4903	0.2096	0.2900	0.1805
Nano Banana	0.5849	0.2346	0.2856	0.2150
CAOF	**0.6639**	0.2014	0.2937	0.1976
CAOF+SASF	**0.8656**	0.2178	0.3022	0.2161

$$\hat{s} = \epsilon + (1-\epsilon)\,\frac{s - s_{\min}}{s_{\max} - s_{\min}}, \quad \epsilon = 0.1.$$

The normalized scores are $\hat{\text{CLIP}}_R$, $\hat{\text{CLIP}}_B$, $\hat{\text{CLIP}}_S$, and $1 - \hat{\text{LPIPS}}_O$.

BOM (Blending Object Metric) measures replacement and blending accuracy:

$$\text{BOM} = \frac{w_R + w_B + w_L}{\frac{w_R}{\hat{\text{CLIP}}_R} + \frac{w_B}{\hat{\text{CLIP}}_B} + \frac{w_L}{1-\hat{\text{LPIPS}}_O}},$$

BOSM (Blending Object Style Metric) further incorporates style fidelity:

$$\text{BOSM} = \frac{w_R + w_B + w_S}{\frac{w_R}{\hat{\text{CLIP}}_R} + \frac{w_B}{\hat{\text{CLIP}}_B} + \frac{w_S}{\hat{\text{CLIP}}_S} + \frac{w_L}{1-\hat{\text{LPIPS}}_O}}.$$

Both metrics are harmonic means where low individual scores significantly lower the final value, highlighting edits that successfully balance content fidelity and stylistic integration.

4.2 Comparisons with SOTA models

Quantitative Evaluation of Object Replacement and Blending. Table 1 presents BOM scores for 800 replacement–blend pairs. CAOF achieves the highest value (0.8388), substantially surpassing the next best method (0.7352). Its advantage does not stem from a single component: although Step1X-Edit yields the best CLIP$_R$ and SeedEdit tops CLIP$_B$, those gains are offset by weaker performance on the complementary cue and by larger perceptual drift, which the harmonic mean penalises. CAOF instead secures near-peak values on both alignment terms while also delivering the strongest image-fidelity score (1−LPIPS$_O$=0.8292). This balance arises from the cost-aware transport in CAOF, which places blend features only at semantically consistent locations, preserving global structure and avoiding the artefacts or concept omission observed in the baselines. The results confirm that effective object blending requires simultaneous optimisation of replacement accuracy, blend consistency, and photographic integrity, a trade-off that CAOF best satisfies among the methods we evaluated.

Quantitative Evaluation of Object Replacement, Blending, and Style Integration. Table 2 lists all methods in ascending BOSM order. The lower half of the table shows that aggressive stylisation or simplistic blending hurts semantic alignment, producing BOSM below 0.40. Middle-ranking approaches recover object fidelity yet still dilute style cues, so their overall balance remains limited. Pure CAOF moves into the upper tier by preserving both objects without increasing perceptual drift, yielding BOSM 0.7102. Adding SASF raises the score to 0.9244, the largest margin in the study. This improvement is not obtained by style similarity alone: CLIP$_R$ and CLIP$_B$ also climb, indicating that the high-frequency details injected by DSIN and the text-driven Key–Value substitution sharpen local structure and make both identities

Figure 8: Method comparison for the task *Knight* → *Leonardo DiCaprio*, blended with *Batman* and rendered in a *water-color* style.

more recognisable. The joint optimisation of content and texture therefore proves essential when multiple conceptual constraints must be satisfied simultaneously.

Visual Assessment. Figure 8 (fantasy portrait) and Figure 9 (celebrity street scene) illustrate the quantitative trend reported in Table 2. CAOF+SASF achieves a balanced fusion where both the replaced identity,

Figure 9: Method comparison for the task *Tom Hanks* → *Taylor Swift*, blended with *jean shorts+white shirt* and rendered in an *oil-painting* style.

the blended identity, and the target style are distinctly visible while maintaining the original scene geometry and background texture. In contrast, the baselines exhibit method-specific failure modes:

Background degradation. In Figure 8, StyleAligned Hertz et al. (2024) replaces the Times Square background with a different scene, and LEDITS++ Brack et al. (2024) and FLUX.1 Kontext Batifol et al. (2025) remove the background almost entirely. In Figure 9, StyleAligned Hertz et al. (2024), Step1X-Edit Liu et al. (2025d),

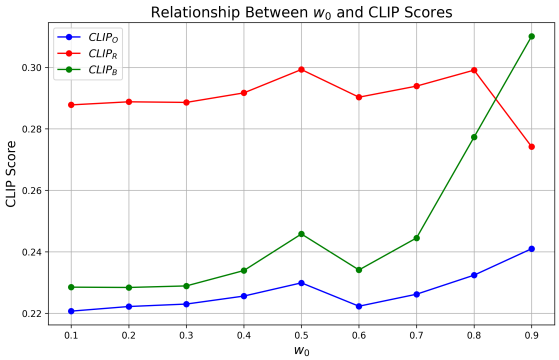

Figure 10: Variation of CLIP scores for original (P_o), replaced (P_r), and blend (P_b) object prompts as the blending coefficient w_0 changes. The curves illustrate CAOF's effectiveness in modulating blending strength, achieving the desired integration of the blend object while replacing the original object.

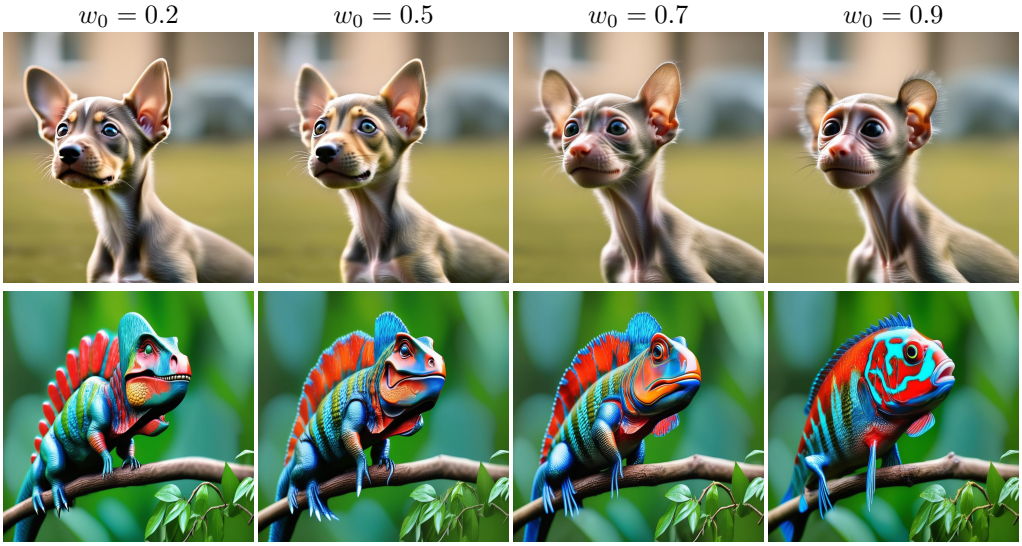

Figure 11: Object blending progression with varying blending coefficients w_0. Row 1: "**monkey**" blended into replaced "**puppy**" (originally "**alpaca**"). **Row 2**: "**fish**" blended into replaced "**dinosaur**" (originally "**chameleon**"). **Row 3**: "**ambulance**" blended into replaced "**jeep**" (originally "**truck**"). Row 4: "**Thanos**" blended into replaced "**knight**" (originally "**robot**"). Higher w_0 values correspond to increased blending intensity and finer textural details.

and FLUX.1 Kontext Batifol et al. (2025) substitute the original blue backdrop, while LEDITS++ Brack et al. (2024) partially destroys the bench. All baselines also drop secondary background details (e.g., leather shoes, socks, and the suit jacket) that are present in the source image.

Loss of replaced-object identity after blending. In Figure 8, TurboEdit Deutch et al. (2024), SeedEdit Shi et al. (2024), Step1X-Edit Liu et al. (2025d), IP2P Brooks et al. (2023), Blended diffusion Avrahami et al. (2022), and FLUX.1 Kontext Batifol et al. (2025) reduce or remove recognizable features of Leonardo DiCaprio. In Figure 9, the same methods fail to preserve identifiable traits of Taylor Swift under the jean-shorts + white-shirt blend.

Severe distortions or unintended objects. In Figure 8, SeedEdit introduces an extra arm, and IP2P generates duplicate faces. In Figure 9, SeedEdit adds an extraneous box, and IP2P produces a heavily distorted image with three faces. Across both figures, outputs from Blended diffusion Avrahami et al. (2022) are often coarse with blurred regions, obscuring small objects.

Table 3: OT ablation: CAOF vs. NoneOT.

Method	BOM↑	CLIP_R↑	CLIP_B↑	$1-\text{LPIPS}_O$↑
NoneOT	0.1429	0.1984	0.2891	0.8304
CAOF	**0.2500**	0.2014	0.2937	0.8292

Table 4: DSIN texture metrics versus α and σ.

α	σ	**LV**↑	**GC**↑	**HFS**↑
0.5	2.5	271.8709	79.9853	5.27×10^9
0.5	0.5	253.3316	79.7168	5.06×10^9
0.2	2.5	266.9800	80.0325	5.21×10^9
0.2	0.5	241.8779	76.5979	4.97×10^9
0.0	–	244.2984	68.8580	4.90×10^9

Insufficient cross-identity blending. In Figure 8, Nano Banana leaves the head largely unchanged as Leonardo DiCaprio, with little to no visible Batman attributes, indicating weak cross-identity fusion.

These issues, along with excessive denoising that washes out high-frequency details or spatial artifacts like duplicated limbs, result in lower BOSM scores for competing methods. This comparison highlights the perceptual advantage of CAOF+SASF, which maintains a coherent and natural fusion without introducing such distortions.

4.3 Ablation Study

Ablation Study on CAOF. To examine how CAOF controls the fusion strength, we vary the blending coefficient $w_0 \in [0.1, 0.9]$ (Eq. 9) and record the CLIP similarities for the original (O), replaced (R), and blend (B) prompts. Figure 10 illustrates the variation of CLIP scores with w_0. The curves clearly demonstrate CAOF's effectiveness in adjusting blending strength. As w_0 increases beyond 0.6, the influence of the blend object prompt P_b significantly rises, while the influence of the replaced object prompt P_r remains high until w_0 exceeds 0.8, after which it decreases rapidly. Concurrently, the influence of the original object prompt P_o remains consistently low throughout, aligning with our goal to replace the original object with the replaced object while blending in the blend object to the desired extent. Qualitative frames in Fig. 11 corroborate the numerical trend, showing a smooth morph from "mostly replacement" to "mostly blend" without geometric break-down.

SASF Ablation. SASF relies solely on textual prompts for style specification, prompting us to measure style blending performance through $\hat{\text{CLIP}}_S$, the normalized similarity between the generated image I_g and the style object prompt P_s. As shown in Table 2, CAOF+SASF attains a substantially higher CLIP_S of 0.2161 than CAOF's 0.1976, indicating that SASF effectively injects the desired style features.

Ablation on the joint percentile thresholds $\tau_{\text{source}}, \tau_{\text{dest}}$. CAOF builds the source set $\mathcal{S}$ and destination set $\mathcal{D}$ by thresholding head-averaged cross-attention responses of the blend and replaced prompts. Positions above the τ_{source} percentile in the blend map enter $\mathcal{S}$ and those above the τ_{dest} percentile in the replaced map enter $\mathcal{D}$. Sweeping the *joint* threshold $\tau_{\text{source}} = \tau_{\text{dest}} \in \{0, 10, \ldots, 90, 99\}$ exposes a precision vs. coverage trade-off that directly governs how CAOF redistributes features. Figure 12 plots CLIP cosine scores for the original prompt P_o, the replaced prompt P_r, and the blend prompt P_b. Three observations follow from the measured curves.

First, P_b peaks at a mid to high threshold: the best blend alignment occurs at 60% where $P_b = 0.2530$ and remains competitive at 70% with $P_b = 0.2371$. Around this regime spurious low-confidence tokens are removed yet all salient parts of the object remain in $\mathcal{S}$ and $\mathcal{D}$. The feature and spatial terms then rank candidate correspondences cleanly and the transport plan concentrates mass on semantically consistent matches, so the fused vectors carry the right identity and geometry.

Second, very high thresholds shrink coverage and favor replacement: when $\tau_{\text{source}}, \tau_{\text{dest}}$ are pushed to 80% and beyond, the sets collapse to a handful of extreme tokens. Coverage drops and CAOF touches only tiny regions, so the CFG-TE replacement signal dominates the denoising trajectory. Empirically P_r rises from 0.1719 at 60% to 0.2361 at 99%, while P_b falls sharply to 0.1592–0.1623.

Third, low thresholds dilute semantic precision: at 0%–50% the sets admit many background or off-object positions. The plan must spread mass across numerous weak matches and the averaged multi-head vectors

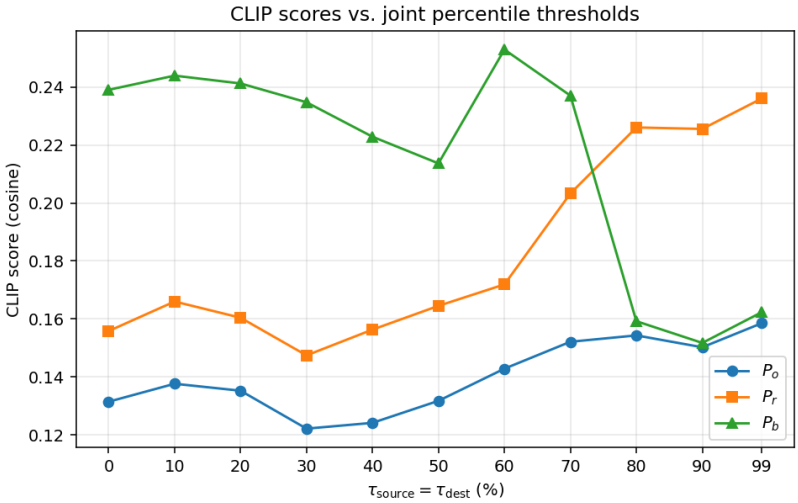

Figure 12: Joint percentile threshold ablation. CLIP cosine scores for P_O, P_R, and P_B as the joint threshold $\tau_{\text{source}} = \tau_{\text{dest}}$ varies. The blend score P_B peaks at 60%. Very high thresholds shrink $\mathcal{S}$ and $\mathcal{D}$ excessively and favor replacement (P_R) over blending, while low thresholds admit non-relevant vectors that dilute the fused signal. Moderate thresholds balance precision and coverage.

mix in non-relevant content, which reduces blend fidelity. This explains the drop from $P_b = 0.2413$ at 20% to $P_b = 0.2137$ at 50%.

Throughout the sweep the original content remains suppressed, with P_o staying low in the range 0.1221–0.1585. Taken together these trends justify the default $\tau_{\text{source}} = \tau_{\text{dest}} \in \{0.6, 0.7\}$. This window removes noise while preserving spatial coverage, maximizes P_b near its peak, keeps P_r strong enough for reliable replacement and maintains low P_o. The ablation therefore confirms that mid to high joint percentiles are critical for stable and semantically faithful blending under CAOF.

OT Ablation. To disentangle the contribution of the Sinkhorn solver, we replace it with a naïve NoneOT variant that line-up source and destination tokens by index and applies a fixed α-blend, thereby ignoring both feature similarity and spatial proximity. As summarised in Table 3, removing Optimal Transport slashes BOM from 0.2500 to 0.1429. The loss is driven almost entirely by lower alignment scores (CLIP_R and CLIP_B), while the perceptual term $1-\text{LPIPS}_O$ remains virtually unchanged. In other words, a uniform blend preserves low-level appearance but often allocates the wrong blend features to the wrong spatial regions, degrading semantic coherence. The cost-aware Sinkhorn plan redistributes those features toward geometrically and visually compatible destinations, yielding a markedly more faithful fusion without sacrificing overall image fidelity.

DSIN Ablation. Laplacian Variance (LV) Pertuz et al. (2013), GLCM Contrast (GC) Haralick et al. (1973), and FFT High–Frequency Sum (HFS) Gonzalez & Woods (2008) show that textural richness depends on the joint choice of the residual-mixing weight α and the Gaussian width σ, rather than on α alone. Raising α strengthens the amplitude of the injected high-frequency residual, but this extra energy is useful only if σ is large enough to confine the smoothing kernel to genuinely low frequencies; with $\alpha = 0.5$ the wider kernel $\sigma = 2.5$ yields the highest LV, GC, and HFS, whereas the same α combined with the narrow kernel $\sigma = 0.5$ loses mid-range structure and drops all three scores. Conversely, keeping α moderate at 0.2 still improves over pure AdaIN ($\alpha = 0$), yet the gain is larger when $\sigma = 2.5$ than when $\sigma = 0.5$. These trends confirm that α governs how much fine detail is transferred while σ sets the frequency band that will be regarded as "detail"; optimal texture emerges when both parameters are tuned together, explaining the peak at $\alpha = 0.5$, $\sigma = 2.5$ in Table 4 and the visibly crisper result in Figure 13.

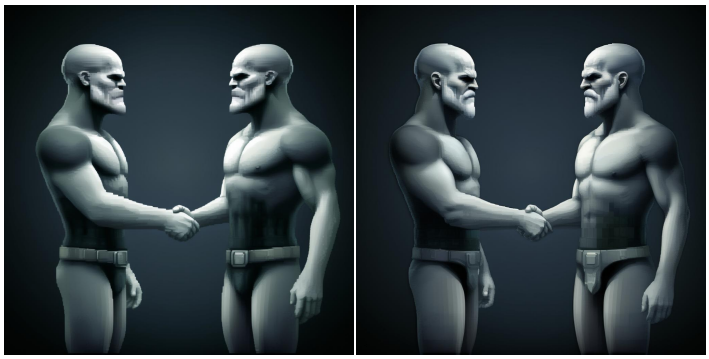

Figure 13: Pixel-art edit of "robot → knight" blended with "Thanos". Left: $\alpha = 0$, right: $\alpha = 0.5$, $\sigma = 2.5$.

5 Conclusion

We introduced TP-Blend, a training-free framework that performs object replacement, object blending, and style fusion within a single diffusion denoising run. By separating the content and style prompts, TP-Blend grants independent control over semantic structure and appearance. Cross-Attention Object Fusion employs an optimal-transport plan to place blend-object features in spatially and semantically consistent regions, while Self-Attention Style Fusion injects high-frequency texture through detail-sensitive instance normalisation and text-driven key–value substitution. Across extensive benchmarks, TP-Blend delivers sharper textures, stronger alignment with target objects and styles, and higher perceptual fidelity than recent editors, all without extra training or model fine-tuning. These results establish TP-Blend as a simple yet effective tool for precise, text-guided image editing within diffusion models.

References

Emanuele Aiello, Umberto Michieli, Diego Valsesia, Mete Ozay, and Enrico Magli. Dreamcache: Finetuning-free lightweight personalized image generation via feature caching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12480–12489, 2025.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18208–18218, 2022.

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18370–18380, 2023.

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. *arXiv preprint arXiv:2411.14430*, 2024.

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7877–7888, 2025.

Ying Ba, Tianyu Zhang, Yalong Bai, Wenyi Mo, Tao Liang, Bing Su, and Ji-Rong Wen. Enhancing reward models for high-quality image generation: Beyond text-image alignment. *arXiv preprint arXiv:2507.19002*, 2025.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. Precisecam: Precise camera control for text-to-image generation. *arXiv preprint arXiv:2501.12910*, 2025.

Bahri Batuhan Bilecen, Yigit Yalin, Ning Yu, and Aysegul Dundar. Reference-based 3d-aware image editing with triplanes. *arXiv preprint arXiv:2404.03632*, 2024.

Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13242–13251, 2025.

Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8861–8870, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Shengqu Cai, Eric Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. *arXiv preprint arXiv:2411.18616*, 2024a.

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18434–18443, 2025a.

Xinhao Cai, Qiuxia Lai, Wenguan Wang, and Yazhou Yao. Cycle-consistent learning for joint layout-to-image generation and object detection, 2024b. URL `https://openreview.net/forum?id=cHKuyeHmS9`.

Ziqi Cai, Shuchen Weng, Yifei Xia, and Boxin Shi. Phys-edit: Physics-aware semantic image editing with text description. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7867–7876. IEEE, 2025b.

Mingdeng Cao, Xuaner Zhang, Yinqiang Zheng, and Zhihao Xia. Instruction-based image manipulation by watching how things move. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2704–2713, 2025.

Clement Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *The 39th Annual AAAI Conference on Artificial Intelligence*, 2025. URL `https://openreview.net/forum?id=D8rQlCEKCT`.

Jierun Chen, Dongting Hu, Xijie Huang, Huseyin Coskun, Arpit Sahni, Aarush Gupta, Anujraaj Goyal, Dishani Lahiri, Rajesh Singh, Yerlan Idelbayev, et al. Snapgen: Taming high-resolution text-to-image models for mobile devices with efficient architectures and training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7997–8008, 2025a.

Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. *arXiv preprint arXiv:2502.20172*, 2025b.

Sherry X Chen, Misha Sra, and Pradeep Sen. Instruct-clip: Improving instruction-guided image editing with automated data refinement using contrastive learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28513–28522, 2025c.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12501–12511, 2025d.

Zhuokun Chen, Jugang Fan, Zhuowei Yu, Bohan Zhuang, and Mingkui Tan. Frequency-aware autoregressive modeling for efficient high-resolution image synthesis, 2025e. URL `https://arxiv.org/abs/2507.20454`.

Hao Cheng, Erjia Xiao, Jiayan Yang, Jiahang Cao, Qiang Zhang, Jize Zhang, Kaidi Xu, Jindong Gu, and Renjing Xu. Not just text: Uncovering vision modality typographic threats in image generation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2997–3007, 2025.

Xiangxiang Chu, Renda Li, and Yong Wang. Usp: Unified self-supervised pretraining for image generation and understanding. *arXiv preprint arXiv:2503.06132*, 2025.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8795–8805, 2024.

Nadav Z Cohen, Oron Nir, and Ariel Shamir. Conditional balance: Improving multi-conditioning trade-offs in image generation. *arXiv preprint arXiv:2412.19853*, 2024.

Nadav Z Cohen, Oron Nir, and Ariel Shamir. Conditional balance: Improving multi-conditioning trade-offs in image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2641–2650, 2025.

Yuren Cong, Martin Renqiang Min, Li Erran Li, Bodo Rosenhahn, and Michael Ying Yang. Attribute-centric compositional text-to-image generation. *International Journal of Computer Vision*, 133(7):4555–4570, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Zhewei Dai, Shilei Zeng, Haotian Liu, Xurui Li, Feng Xue, and Yu Zhou. Seas: few-shot industrial anomaly image generation with separation and sharing fine-tuning. *arXiv preprint arXiv:2410.14987*, 2024.

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow transformers. *arXiv preprint arXiv:2412.09611*, 2024.

Shengqi Dang, Yi He, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. *arXiv preprint arXiv:2501.05710*, 2025.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. Z-magic: Zero-shot multiple attributes guided image creator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18390–18400, 2025.

Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. *arXiv preprint arXiv:2408.00735*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Lunhao Duan, Shanshan Zhao, Wenjun Yan, Yinglun Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Mingming Gong, and Gui-Song Xia. Unic-adapter: Unified image-instruction adapter with multi-modal transformer for image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7963–7973, 2025.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

Fengyi Fu, Lei Zhang, Mengqi Huang, and Zhendong Mao. Feededit: Text-based image editing with dynamic feedback regulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2661–2670, 2025a.

Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. Univg: A generalist diffusion model for unified image generation and editing, 2025b. URL `https://arxiv.org/abs/2503.12652`.

Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7545–7556, 2023.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson/Prentice Hall, 3 edition, 2008.

Songen Gu, Jinming Su, Yiting Duan, Xingyue Chen, Junfeng Luo, and Hao Zhao. Text2street: Controllable text-to-image generation for street views. In *International Conference on Pattern Recognition*, pp. 130–145. Springer, 2025.

Zirun Guo and Tao Jin. Conceptguard: Continual personalized text-to-image generation with forgetting and confusion mitigation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2945–2954, 2025.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step, 2025. URL `https://arxiv.org/abs/2501.13926`.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025a.

Tao Han, Wanghan Xu, Junchao Gong, Xiaoyu Yue, Song Guo, Luping Zhou, and Lei Bai. Infgen: A resolution-agnostic paradigm for scalable image synthesis, 2025b. URL `https://arxiv.org/abs/2509.10441`.

Woojung Han, Yeonkyung Lee, Chanyoung Kim, Kwanghyun Park, and Seong Jae Hwang. Spatial transport optimization by repositioning attention map for training-free text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18401–18410, 2025c.

Robert M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.

Runze He, Bo Cheng, Yuhang Ma, Qingxiang Jia, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Liebucha Wu, Dawei Leng, and Yuhui Yin. Plangen: Towards unified layout planning and image generation in auto-regressive vision language models, 2025. URL `https://arxiv.org/abs/2503.10127`.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4775–4785, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Teng-Fang Hsiao, Bo-Kai Ruan, Yi-Lun Wu, Tzu-Ling Lin, and Hong-Han Shuai. Tf-ti2i: Training-free text-and-image-to-image generation via multi-modal implicit-context learning in text-to-image models. *arXiv preprint arXiv:2503.15283*, 2025.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Qihan Huang, Long Chan, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, and Jie Song. Patchdpo: Patch-level dpo for finetuning-free personalized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18369–18378, 2025b.

Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3707–3714, 2025c.

Shanshan Huang, Haoxuan Li, Chunyuan Zheng, Mingyuan Ge, Wei Gao, Lei Wang, and Li Liu. Text-driven fashion image editing with compositional concept learning and counterfactual abduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28726–28735, 2025d.

Shanshan Huang, Haoxuan Li, Chunyuan Zheng, Lei Wang, Guorui Liao, Zhili Gong, Huayi Yang, and Li Liu. Visual representation learning through causal intervention for controllable image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23484–23493, 2025e.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy, 2024. URL `https://arxiv.org/abs/2412.03632`.

Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T Freeman, David Salesin, Brian Curless, et al. Slide: Single image 3d photography with soft layering and depth-aware inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12518–12527, 2021.

Sangwon Jang, June Suk Choi, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Silent branding attack: Trigger-free data poisoning attack on text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8203–8212, 2025.

Jinho Jeong, Sangmin Han, Jinwoo Kim, and Seon Joo Kim. Latent space super-resolution for higher-resolution image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2355–2365, 2025.

Chengyou Jia, Changliang Xia, Zhuohang Dang, Weijia Wu, Hangwei Qian, and Minnan Luo. Chatgen: Automatic text-to-image generation from freestyle chatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13284–13293, 2025a.

Weinan Jia, Mengqi Huang, Nan Chen, Lei Zhang, and Zhendong Mao. Dˆ 2it: Dynamic diffusion transformer for accurate image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12860–12870, 2025b.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

Jian Jin, Zhenbo Yu, Yang Shen, Zhenyong Fu, and Jian Yang. Latexblend: Scaling multi-concept customized generation with latent textual blending. *arXiv preprint arXiv:2503.06956*, 2025.

Kyungmin Jo, Jooyeol Yun, and Jaegul Choo. Devil is in the detail: Towards injecting fine details of image prompt in image generation via conflict-free guidance and stratified attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23595–23603, 2025.

Sangwon Jung, Alex Oesterling, Claudio Mayrink Verdun, Sajani Vithana, Taesup Moon, and Flavio P Calmon. Multi-group proportional representations for text-to-image models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23744–23754, 2025.

Daneul Kim, Jaeah Lee, and Jaesik Park. Improving editability in image generation with layer-wise memory. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7889–7898, 2025a.

Eunji Kim, Siwon Kim, Minjun Park, Rahim Entezari, and Sungroh Yoon. Rethinking training for de-biasing text-to-image generation: Unlocking the potential of stable diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13361–13370, 2025b.

Hermann Kumbong, Xian Liu, Tsung-Yi Lin, Ming-Yu Liu, Xihui Liu, Ziwei Liu, Daniel Y Fu, Christopher Re, and David W Romero. Hmar: Efficient hierarchical masked auto-regressive image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2535–2544, 2025.

Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M Rehg, Sangmin Lee, Ning Zhang, et al. Unleashing in-context learning of autoregressive models for few-shot image manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18346–18357, 2025.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18596–18606, 2025a.

Jaerin Lee, Daniel Sungho Jung, Kanggeon Lee, and Kyoung Mu Lee. Semanticdraw: towards real-time interactive content creation from image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13021–13030, 2025b.

Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, and Chi Zhang. Stylestudio: Text-driven style transfer with selective control of style elements. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23443–23452, 2025.

Feifei Li, Mi Zhang, Yiming Sun, and Min Yang. Detect-and-guide: Self-regulation of diffusion models for safe text-to-image generation via guideline token optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13252–13262, 2025a.

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26256–26264, 2025b.

Lingxiao Li, Kaixuan Fan, Boqing Gong, and Xiangyu Yue. Hypdae: Hyperbolic diffusion autoencoders for hierarchical few-shot image generation. *arXiv preprint arXiv:2411.17784*, 2024a.

Mengtian Li, Jinshu Chen, Wanquan Feng, Bingchuan Li, Fei Dai, Songtao Zhao, and Qian He. Hyperlora: Parameter-efficient adaptive generation for portrait synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13114–13123, 2025c.

Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkang Yang, Ziyun Qian, Jiawei Chen, Jinjie Wei, Yue Jiang, Qingyao Xu, and Lihua Zhang. Mccd: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13263–13272, 2025d.

Senmao Li, Lei Wang, Kai Wang, Tao Liu, Jiehang Xie, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, and Jian Yang. One-way ticket: Time-independent unified encoder for distilling text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23563–23574, 2025e.

Shaoxu Li. Diffstyler: Diffusion-based localized image style transfer. *arXiv preprint arXiv:2403.18461*, 2024.

Yanfeng Li, Kahou Chan, Yue Sun, Chantong Lam, Tong Tong, Zitong Yu, Keren Fu, Xiaohong Liu, and Tao Tan. Moedit: On learning quantity perception for multi-object image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2683–2693, 2025f.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650, 2024b.

Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. *arXiv preprint arXiv:2504.07960*, 2025g.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2779–2790, 2025h.

Dong Liang, Jinyuan Jia, Yuhao Liu, Zhanghan Ke, Hongbo Fu, and Rynson WH Lau. Vodiff: Controlling object visibility order in text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18379–18389, 2025.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-driven image editing via learnable regions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7059–7068, 2024.

Chang Liu, Xiangtai Li, and Henghui Ding. Referring image editing: Object-level image editing via referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13128–13138, 2024.

Hongda Liu, Longguang Wang, Ye Zhang, Ziru Yu, and Yulan Guo. Samam: Style-aware state space model for arbitrary image style transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28468–28478, 2025a.

Mushui Liu, Yuhang Ma, Zhen Yang, Jun Dan, Yunlong Yu, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5523–5531, 2025b.

Mushui Liu, Dong She, Jingxuan Pang, Qihan Huang, Jiacheng Ying, Wanggui He, Yuanlei Hou, and Siming Fu. Tfcustom: Customized image generation with time-aware frequency feature guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2714–2723, 2025c.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025d.

Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. *arXiv preprint arXiv:2501.13554*, 2025e.

Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: text-guided artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3530–3534, 2023.

Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13072–13082, 2025f.

Winfried Lötzsch, Max Reimann, Martin Büssemeyer, Amir Semmo, Jürgen Döllner, and Matthias Trapp. Wise: Whitebox image stylization by example-based learning. In *European Conference on Computer Vision*, pp. 135–152. Springer, 2022.

Grace Luo, Jonathan Granskog, Aleksander Holynski, and Trevor Darrell. Dual-process image generation. *arXiv preprint arXiv:2506.01955*, 2025a.

Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Hancheng Min, Chris Callison-Burch, and René Vidal. Concept lancet: Image editing with compositional representation transplant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28502–28512, 2025b.

Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37:115377–115408, 2024.

Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025a.

Xiaofeng Mao, Zhengkai Jiang, Fu-Yun Wang, Jiangning Zhang, Hao Chen, Mingmin Chi, Yabiao Wang, and Wenhan Luo. Osv: One step is enough for high-quality image to video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12585–12594, 2025b.

Boming Miao, Chunxiao Li, Xiaoxiao Wang, Andi Zhang, Rui Sun, Zizhe Wang, and Yao Zhu. Noise diffusion for enhancing semantic faithfulness in text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23575–23584, 2025.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.

Sanghyeon Na, Yonggyu Kim, and Hyunjoon Lee. Boost your human image generation model via direct preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23551–23562, 2025.

Toan Nguyen, Kien Do, Duc Kieu, and Thin Nguyen. h-edit: Effective and flexible diffusion-based editing via doob's h-transform. *arXiv preprint arXiv:2503.02187*, 2025a.

Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. *arXiv preprint arXiv:2412.04301*, 2024.

Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21492–21501, 2025b.

Jeonghoon Park, Juyoung Lee, Chaeyeon Chung, Jaeseong Lee, Jaegul Choo, and Jindong Gu. Fair generation without unfair distortions: Debiasing text-to-image generation with entanglement-free attention. *arXiv preprint arXiv:2506.13298*, 2025a.

Jihun Park, Jongmin Gim, Kyoungmin Lee, Seunghun Lee, and Sunghoon Im. Style-editor: Text-driven object-centric style editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18281–18291, 2025b.

S. Pertuz, D. Puig, and M. A. Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, et al. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7952–7962, 2025.

Weimin Qiu, Jieke Wang, and Meng Tang. Self-cross diffusion guidance for text-to-image synthesis of similar subjects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23528–23538, 2025.

Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua. Silmm: Self-improving large multimodal models for compositional text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18497–18508, 2025a.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2545–2555, 2025b.

Yufan Ren, Zicong Jiang, Tong Zhang, Søren Forchhammer, and Sabine Süsstrunk. Fds: Frequency-aware denoising score for text-guided latent diffusion image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2651–2660, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Suho Ryu, Kihyun Kim, Eugene Baek, Dongsoo Shin, and Joonseok Lee. Towards scalable human-aligned benchmark for text-guided image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18292–18301, 2025.

Donald Shenaj, Ondrej Bohdal, Mete Ozay, Pietro Zanuttigh, and Umberto Michieli. Lora. rar: Learning to merge loras via hypernetworks for subject-style conditioned image generation. *arXiv preprint arXiv:2412.05148*, 2024.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Shuwei Shi, Biao Gong, Xi Chen, Dandan Zheng, Shuai Tan, Zizheng Yang, Yuyuan Li, Jingwen He, Kecheng Zheng, Jingdong Chen, et al. Motionstone: Decoupled motion intensity modulation with diffusion transformer for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22864–22874, 2025a.

Shuwei Shi, Wenbo Li, Yuechen Zhang, Jingwen He, Biao Gong, and Yinqiang Zheng. Resmaster: Mastering high-resolution image generation via structural and fine-grained guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6887–6895, 2025b.

Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. *arXiv preprint arXiv:2411.06686*, 2024.

Wataru Shimoda, Naoto Inoue, Daichi Haraguchi, Hayato Mitani, Seichi Uchida, and Kota Yamaguchi. Type-r: Automatically retouching typos for text-to-image generation. *arXiv preprint arXiv:2411.18159*, 2024.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7986–7996, 2025.

Junhyuk So, Juncheol Shin, Hyunho Kook, and Eunhyeok Park. Grouped speculative decoding for autoregressive image generation. *arXiv preprint arXiv:2508.07747*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Tongtong Su, Chengyu Wang, Bingyan Liu, Jun Huang, and Dongming Lu. Encapsulated composition of text-to-image and text-to-video models for high-quality video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18209–18218, 2025.

Haoyuan Sun, Bo Xia, Yongzhe Chang, and Xueqian Wang. Generalizing alignment paradigm of text-to-image generation with preferences through f-divergence minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27644–27652, 2025.

Bingda Tang, Boyang Zheng, Sayak Paul, and Saining Xie. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28586–28595, 2025.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

Soobin Um and Jong Chul Ye. Minority-focused text-to-image generation via prompt optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20926–20936, 2025.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024.

Fangyikang Wang, Hubery Yin, Lei Qian, Yinan Li, Shaobin Zhuang, Huminhao Zhu, Yilin Zhang, Yanlong Tang, Chao Zhang, Hanbin Zhao, et al. Unleashing high-quality image generation in diffusion sampling using second-order levenberg-marquardt-langevin. *arXiv preprint arXiv:2505.24222*, 2025a.

Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024a.

Jiahao Wang, Ning Kang, Lewei Yao, Mengzhao Chen, Chengyue Wu, Songyang Zhang, Shuchen Xue, Yong Liu, Taiqiang Wu, Xihui Liu, et al. Lit: Delving into a simplified linear diffusion transformer for image generation. *arXiv preprint arXiv:2501.12976*, 2025b.

Lifu Wang, Daqing Liu, Xinchen Liu, and Xiaodong He. Scaling down text encoders of text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18424–18433, 2025c.

Ruoyu Wang, Huayang Huang, Ye Zhu, Olga Russakovsky, and Yu Wu. The silent assistant: Noisequery as implicit guidance for goal-driven image generation. *arXiv preprint arXiv:2412.05101*, 2024b.

Weicheng Wang, Guoli Jia, Zhongqi Zhang, Liang Lin, and Jufeng Yang. Ps-diffusion: Photorealistic subject-driven image editing with disentangled control and attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18302–18312. IEEE Computer Society, 2025d.

Wenzhuang Wang, Yifan Zhao, Mingcan Ma, Ming Liu, Zhonglin Jiang, Yong Chen, and Jia Li. Ficgen: Frequency-inspired contextual disentanglement for layout-driven degraded image generation. *arXiv preprint arXiv:2509.01107*, 2025e.

Xi Wang, Hongzhen Li, Heng Fang, Yichen Peng, Haoran Xie, Xi Yang, and Chuntao Li. Lineart: A knowledge-guided training-free high-quality appearance transfer for design drawing with diffusion model. *arXiv preprint arXiv:2412.11519*, 2024c.

Yuxuan Wang, Tianwei Cao, Huayu Zhang, Zhongjiang He, Kongming Liang, and Zhanyu Ma. Fairhuman: Boosting hand and face quality in human image generation with minimum potential delay fairness in diffusion models. *arXiv preprint arXiv:2507.02714*, 2025f.

Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Designdiffusion: High-quality text-to-design image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20906–20915, 2025g.

Zhizhong Wang, Lei Zhao, and Wei Xing. Styledifusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7677–7689, 2023.

Zihao Wang, Yuxiang Wei, Fan Li, Renjing Pei, Hang Xu, and Wangmeng Zuo. Ace: Anti-editing concept erasure in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23505–23515, 2025h.

Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, and Zhengzhong Tu. Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8213–8224, 2025i.

Zixuan Wang, Duo Peng, Feng Chen, Yuwei Yang, and Yinjie Lei. Training-free dense-aligned diffusion guidance for modular conditional image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13135–13145, 2025j.

Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17882–17891, 2025a.

Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3944–3953. IEEE, 2025b.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation, 2025c. URL `https://arxiv.org/abs/2504.02160`.

Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, and Di Zhang. Paragraph-to-image generation with information-enriched diffusion model. *International Journal of Computer Vision*, pp. 1–22, 2025d.

Yecheng Wu, Junyu Chen, Zhuoyang Zhang, Enze Xie, Jincheng Yu, Junsong Chen, Jinyi Hu, Yao Lu, Song Han, and Han Cai. Dc-ar: Efficient masked autoregressive image generation with deep compression hybrid tokenizer. *arXiv preprint arXiv:2507.04947*, 2025e.

Yichun Wu, Huihuang Zhao, Wenhui Chen, Yunfei Yang, and Jiayi Bu. Textstyler: A clip-based approach to text-guided style transfer. *Computers & Graphics*, 119:103887, 2024a.

Yinwei Wu, Xianpan Zhou, Bing Ma, Xuefeng Su, Kai Ma, and Xinchao Wang. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240*, 2024b.

Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing. *arXiv preprint arXiv:2412.17098*, 2024.

Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28533–28543, 2025.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Fr'edo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025a.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025b.

Cong Xie, Han Zou, Ruiqi Yu, Yan Zhang, and Zhenpeng Zhan. Serialgen: Personalized image generation by first standardization then personalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2847–2856, 2025a.

Liangbin Xie, Daniil Pakhomov, Zhonghao Wang, Zongze Wu, Ziyan Chen, Yuqian Zhou, Haitian Zheng, Zhifei Zhang, Zhe Lin, Jiantao Zhou, et al. Turbofill: Adapting few-step text-to-image model for fast image inpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7613–7622, 2025b.

Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024.

Xiaoying Xing, Avinab Saha, Junfeng He, Susan Hao, Paul Vicol, Moonkyung Ryu, Gang Li, Sahil Singla, Sarah Young, Yinxiao Li, et al. Focus-n-fix: Region-aware fine-tuning for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18486–18496, 2025.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation, 2025. URL `https://arxiv.org/abs/2504.08736`.

Yingjing Xu, Jie Kong, Jiazhi Wang, Xiao Pan, Bo Lin, and Qiang Liu. Insightedit: Towards better instruction following for image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2694–2703, 2025.

Youcan Xu, Zhen Wang, Jun Xiao, Wei Liu, and Long Chen. Freetuner: Any subject in any style with training-free diffusion. *arXiv preprint arXiv:2405.14201*, 2024.

Wenjie Xuan, Jing Zhang, Juhua Liu, Bo Du, and Dacheng Tao. Rethink sparse signals for pose-guided text-to-image generation. *arXiv preprint arXiv:2506.20983*, 2025.

Han Yang, Chuanguang Yang, Qiuli Wang, Zhulin An, Weilun Feng, Libo Huang, and Yongjun Xu. Multi-party collaborative attention control for image customization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7942–7951, 2025a.

Haosen Yang, Adrian Bulat, Isma Hadji, Hai X Pham, Xiatian Zhu, Georgios Tzimiropoulos, and Brais Martinez. Fam diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2459–2468, 2025b.

Hongji Yang, Wencheng Han, Yucheng Zhou, and Jianbing Shen. Dc-controlnet: Decoupling inter-and intra-element conditions in image generation with diffusion models. *arXiv preprint arXiv:2502.14779*, 2025c.

Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoedit: Evoking emotions through image manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24690–24699, 2025d.

Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7643–7653, 2025e.

Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, and Guo-Jun Qi. Schedule on the fly: Diffusion time prediction for faster and better image generation. *arXiv preprint arXiv:2412.01243*, 2024.

Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, and Guo-Jun Qi. Schedule on the fly: Diffusion time prediction for faster and better image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23412–23422, 2025.

Srikar Yellapragada, Alexandros Graikos, Kostas Triaridis, Prateek Prasanna, Rajarsi R Gupta, Joel Saltz, and Dimitris Samaras. Zoomldm: Latent diffusion model for multi-scale image generation. *arXiv preprint arXiv:2411.16969*, 2024.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025a.

Qingtao Yu, Jaskirat Singh, Zhaoyuan Yang, Peter Henry Tu, Jing Zhang, Hongdong Li, Richard Hartley, and Dylan Campbell. Probability density geodesics in image diffusion latent space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27989–27998, 2025b.

Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7920–7930, 2025.

Taeyoung Yun, Dinghuai Zhang, Jinkyoo Park, and Ling Pan. Learning to sample effective and diverse prompts for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23625–23635, 2025.

Guanning Zeng, Xiang Zhang, Zirui Wang, Haiyang Xu, Zeyuan Chen, Bingnan Li, and Zhuowen Tu. Yolo-count: Differentiable object counting for text-to-image generation. *arXiv preprint arXiv:2508.00728*, 2025.

Kaiwen Zha, Lijun Yu, Alireza Fathi, David A Ross, Cordelia Schmid, Dina Katabi, and Xiuye Gu. Language-guided image tokenization for generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15713–15722, 2025.

Xiaohang Zhan and Dingming Liu. Larender: Training-free occlusion control in image generation via latent rendering, 2025. URL `https://arxiv.org/abs/2508.07647`.

Hui Zhang, Dexiang Hong, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. *arXiv preprint arXiv:2412.03859*, 2024.

Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23464–23473, 2025a.

Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Ziyu Guo, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Peng Gao, and Hongsheng Li. Let's verify and reinforce image generation step by step. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28662–28672, 2025b.

Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. Scene splatter: Momentum 3d scene generation from single image with video diffusion model. *arXiv preprint arXiv:2504.02764*, 2025c.

Sicheng Zhang, Binzhu Xie, Zhonghao Yan, Yuli Zhang, Donghao Zhou, Xiaofei Chen, Shi Qiu, Jiaqi Liu, Guoyang Xie, and Zhichao Lu. Trade-offs in image generation: How do different dimensions interact? *arXiv preprint arXiv:2507.22100*, 2025d.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10146–10156, 2023.

Anlin Zheng, Haochen Wang, Yucheng Zhao, Weipeng Deng, Tiancai Wang, Xiangyu Zhang, and Xiaojuan Qi. Holistic tokenizer for autoregressive image generation. *arXiv preprint arXiv:2507.02358*, 2025a.

Peng Zheng, Junke Wang, Yi Chang, Yizhou Yu, Rui Ma, and Zuxuan Wu. Rethinking discrete tokens: Treating them as conditions for continuous autoregressive image synthesis, 2025b. URL https://arxiv.org/abs/2507.01756.

Chao Zhou, Tianyi Wei, and Nenghai Yu. Scale your instructions: Enhance the instruction-following fidelity of unified image generation model by self-adaptive attention scaling. *arXiv preprint arXiv:2507.16240*, 2025a.

Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, and Xiaodan Liang. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13093–13103, 2025b.

Shijie Zhou, Ruiyi Zhang, Huaisheng Zhu, Branislav Kveton, Yufan Zhou, Jiuxiang Gu, Jian Chen, and Changyou Chen. Multimodal llms as customized reward models for text-to-image generation, 2025c. URL https://arxiv.org/abs/2507.21391.

Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. Multibooth: Towards generating all your concepts in an image from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10923–10931, 2025a.

Jiapeng Zhu, Ceyuan Yang, Kecheng Zheng, Yinghao Xu, Zifan Shi, Yifei Zhang, Qifeng Chen, and Yujun Shen. Exploring sparse moe in gans for text-conditioned image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18411–18423, 2025b.

Zixin Zhu, Kevin Duarte, Mamshad Nayeem Rizve, Chengyuan Xu, Ratheesh Kalarot, and Junsong Yuan. Compslider: Compositional slider for disentangled multiple-attribute image generation, 2025c. URL https://arxiv.org/abs/2509.01028.