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

Anonymous authors
Paper under double-blind review

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. (2024) 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 depend on reference images, limiting users to existing examples and demanding 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. (2024)—where styles are specified by natural language (e.g., "sketch-like," "art nouveau")—could offer greater flexibility but 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.

2 Related Work

Diffusion models have become the de-facto backbone for text-guided image generation and editing, beginning with unconditional DDPMs Dhariwal & Nichol (2021) and latent variants such as SD-XL Rombach et al. (2022). Guidance strategies based on classifier-free gradients Ho & Salimans (2022) underpin early editing systems including IP2P Brooks et al. (2023) and LEDITS++ Brack et al. (2024), yet these frameworks struggle with multi-concept entanglement and fine-grained regional control. Recent work seeks broader functionality: Step1X-Edit Liu et al. (2025), AnyEdit Yu et al. (2024), DreamOmni Xia et al. (2024) and FireEdit Zhou et al. (2025) pursue unified pipelines that merge generation and editing, whereas Concept Lancet Luo et al. (2025), LaTexBlend Jin et al. (2025) and Conditional Balance Cohen et al. (2024) probe the trade-off space of multi-conditioning. Acceleration is addressed by SwiftEdit Nguyen et al. (2024), h-Edit Nguyen et al. (2025), Schedule-on-the-Fly Ye et al. (2024) and ZoomLDM Yellapragada et al. (2024); resolution scaling appears in Diffusion-4K Zhang et al. (2025a). Orthogonal architecture advances such as Switti's scale-wise transformer design Voronov et al. (2024) highlight that hierarchical token aggregation can improve long-range coherence without sacrificing detail, a property complementary to our cross-/self-attention fusion. Domain-specific extensions include DesignDiffusion Wang et al. (2025a), LineArt Wang et al. (2024b), Focus-N-Fix Xing et al. (2025), Type-R Shimoda et al. (2024) and PreciseCam Bernal-Berdun et al. (2025). Alternative representations—triplanes Bilecen et al. (2024), rectified flows Dalva et al. (2024) and dense-aligned guidance Wang et al. (2025b)—expand editing modalities, while Stable Flow Avrahami et al. (2024) and Scene Splatter Zhang et al. (2025b) explore training-free and 3-D generative directions. Style transfer inside diffusion is usually image-referenced or AdaIN-based; text-driven approaches such as diffusion self-distillation Cai et al. (2024) begin to remove exemplar dependence but blur high-frequency details. Our Dual-Prompt Attention Fusion complements these lines by (i) Optimal-Transport cross-attention that reliably blends object identities without retraining and (ii) DSIN-based self-attention that injects text-specified style while explicitly preserving high-frequency texture, outperforming prior art on multi-concept and style-aware edits.

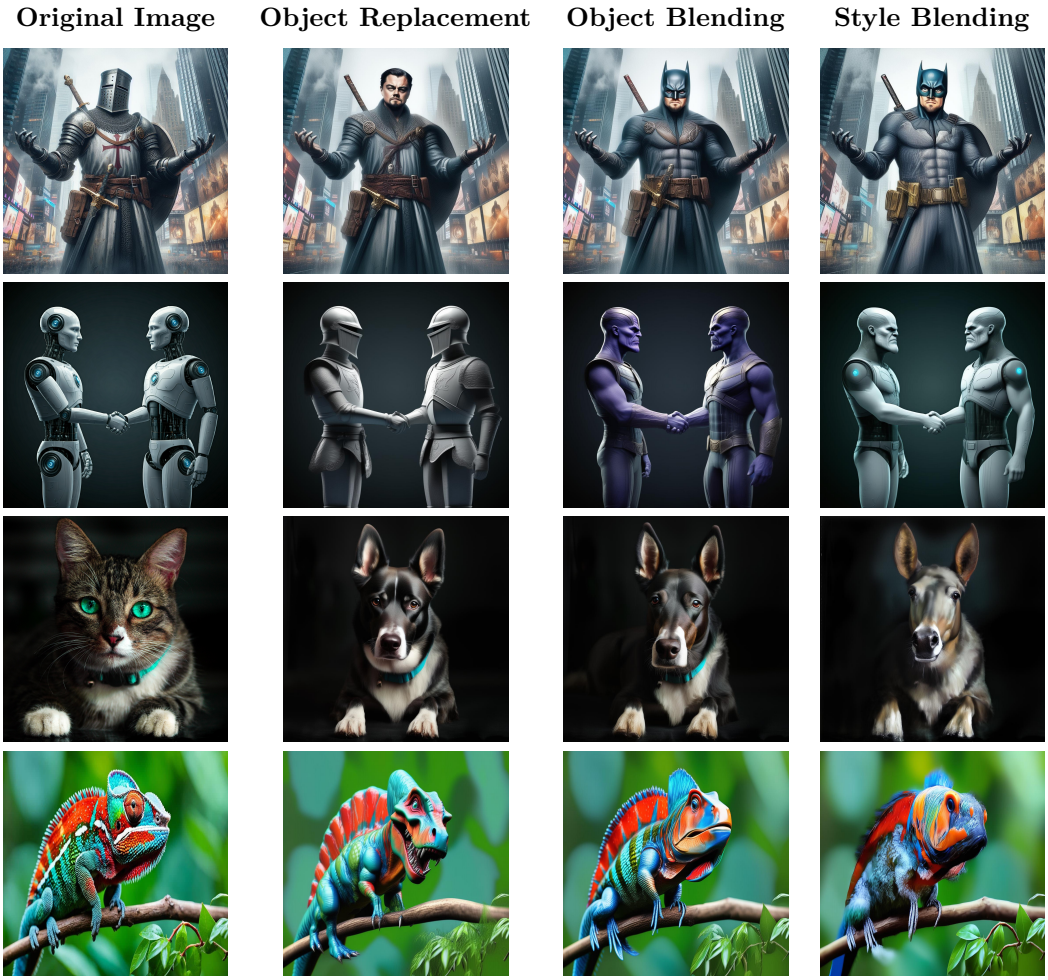

Figure 1: Demonstration of our method's capabilities. Row 1: Original object "**Knight**" is replaced by "**Leonardo DiCaprio**", blended with "**Batman**", and styled with "**Pop Art**". Row 2: Original object "**Robot**" is replaced by "**Knight**", blended with "**Thanos**", and styled with "**Cyberpunk 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**".

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 (e.g., text) 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 scale s_g modifies the predicted noise:

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

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 deterministically recovers $\mathbf{x}_T$ from $\mathbf{x}_0$ without reconstruction

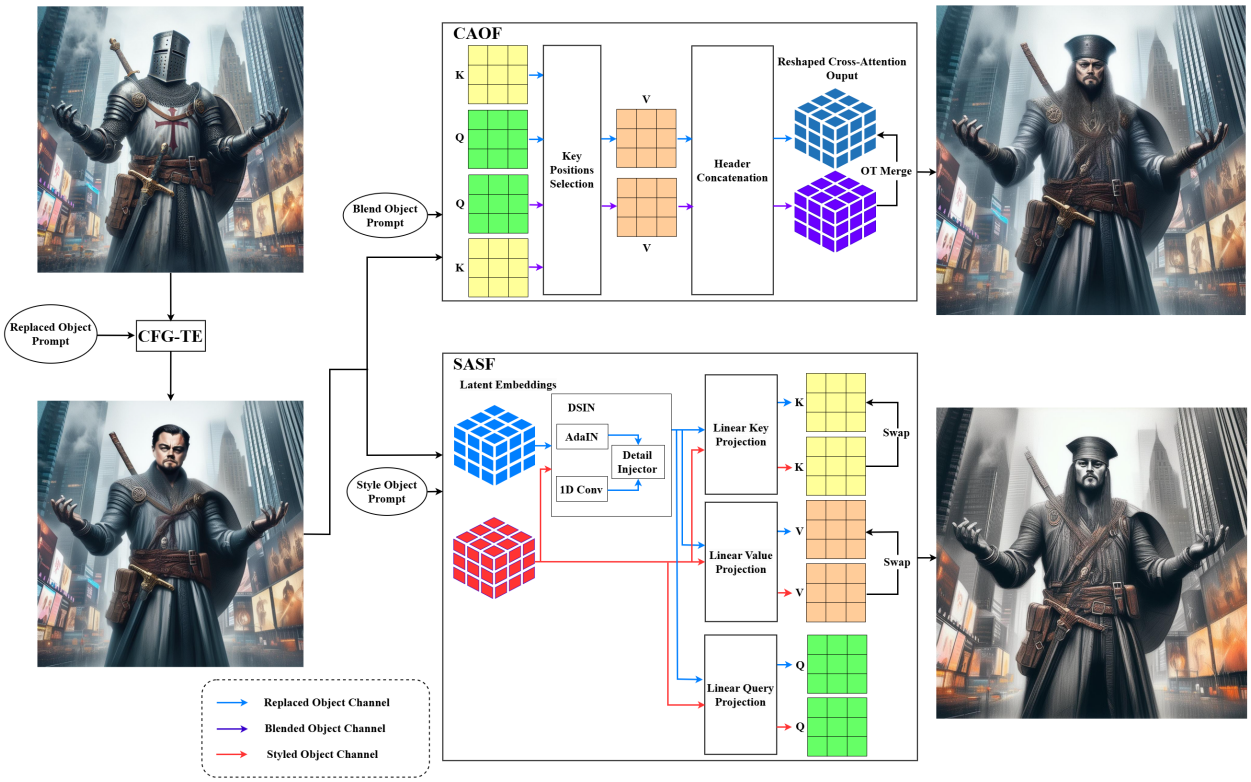

Figure 2: Flowchart of TP-Blend, integrating object replacement, blending, and style transfer within the diffusion process. In this example, the original object "**Knight**" is replaced by "**Leonardo DiCaprio**", blended with "**Captain Jack Sparrow**", and styled with a "**Charcoal Drawing**" effect.

error:

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

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

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

with

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

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).

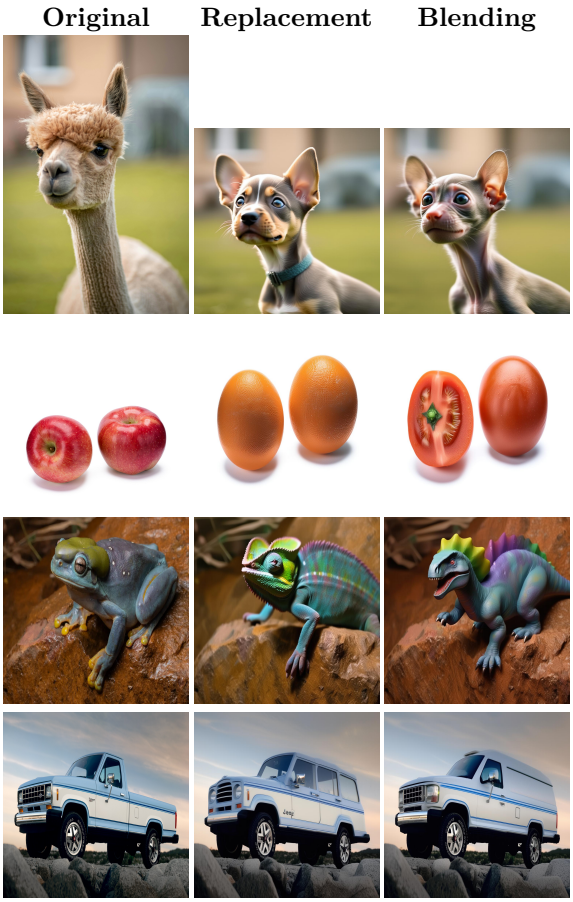

Figure 3: CAOF Object Blending across different sets. Row 1: Original object "**alpaca**" is replaced by "**puppy**" and blended with "**monkey**". Row 2: Original "**apple**" is replaced by "**orange**" and blended with "**tomato**". Row 3: Original "**frog**" is replaced by "**chameleon**" and blended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

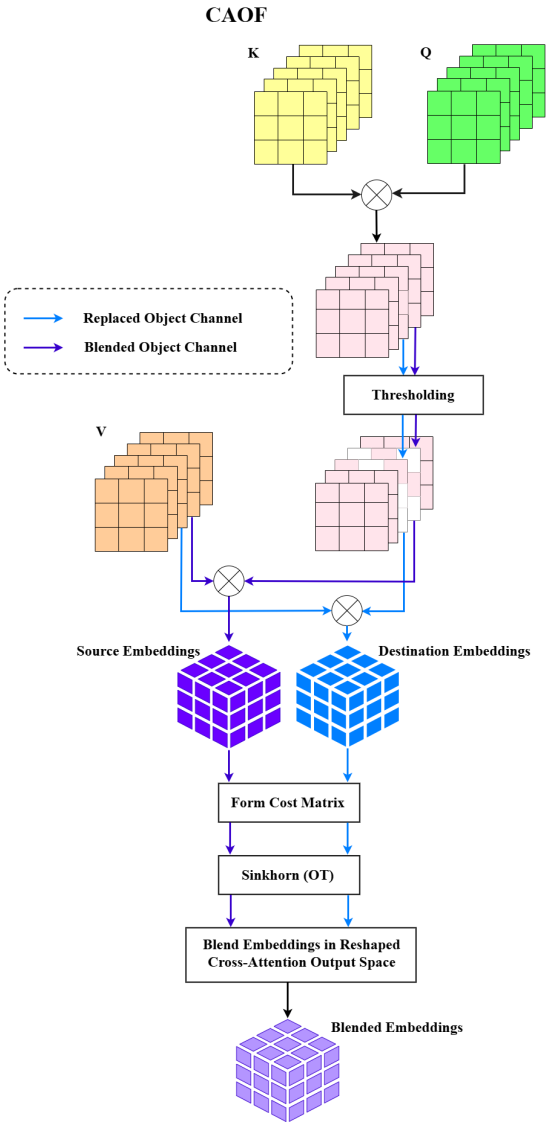

Figure 4: CAOF Flowchart: Cross-Attention Object Fusion merges the blend object's features into the replaced object by identifying key spatial positions in the attention maps and applying an optimal transport framework for coherent morphological transitions.

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 $\mathcal{D}$.

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:

$$\mathbf{O} \;=\; \mathrm{Concat}_{h=1}^{H}\big(\mathbf{A}^{(h)}\mathbf{V}^{(h)}\big) \;\in\; \mathbb{R}^{N\times D}, \tag{7}$$

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

$$\mathbf{f}'_{d_i} \;=\; (1-w_0)\,\mathbf{f}_{d_i} \;+\; w_0 \sum_{s_j\in\mathcal{S}} \frac{T_{ij}}{\sum_{s_k\in\mathcal{S}} T_{ik}}\,\mathbf{f}_{s_j}, \tag{8}$$

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

$$C_{ij} = \lambda_{\text{feature}}\, D_{\text{feature}}(i,j) \;+\; \lambda_{\text{spatial}}\, D_{\text{spatial}}(i,j), \tag{9}$$

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:

$$\min_{\mathbf{T}\geq 0} \quad \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{S}} T_{ij}C_{ij} - \gamma H(\mathbf{T}), \tag{10}$$

$$\text{s.t.} \quad \sum_{j\in\mathcal{S}} T_{ij} = 1, \quad \forall i \in \mathcal{D}, \tag{11}$$

$$\sum_{i\in\mathcal{D}} T_{ij} \geq \frac{1}{|\mathcal{S}|}, \quad \forall j \in \mathcal{S}, \tag{12}$$

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 $\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}$:

$$\mathbf{u}^{(k+1)} \;=\; \frac{\mathbf{1}_{|\mathcal{D}|}}{\mathbf{K}\,\mathbf{v}^{(k)}}, \quad \mathbf{v}^{(k+1)} \;=\; \frac{\frac{1}{|\mathcal{S}|}\,\mathbf{1}_{|\mathcal{S}|}}{\mathbf{K}^{\top}\,\mathbf{u}^{(k+1)}}, \tag{13}$$

until convergence. The transport plan becomes

$$\mathbf{T} \;=\; \mathrm{diag}(\mathbf{u})\,\mathbf{K}\,\mathrm{diag}(\mathbf{v}). \tag{14}$$

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.

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:

$$\mathbf{F}'_{\text{replaced}} \;=\; \left(\frac{\mathbf{F}_{\text{replaced}} \;-\; \mu_{\text{rep}}}{\sigma_{\text{rep}}}\right) \sigma_{\text{style}} \;+\; \mu_{\text{style}}, \tag{15}$$

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:

$$\mathbf{F}^{\text{LF}} \;=\; \mathbf{F} * \mathbf{K}, \quad \mathbf{F}^{\text{HF}} \;=\; \mathbf{F} \;-\; \mathbf{F}^{\text{LF}}, \tag{16}$$

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:

$$\mathbf{F}''_{\text{replaced}} \;=\; \mathbf{F}'_{\text{replaced}} \;+\; \alpha\,\big(\mathbf{F}^{\text{HF}}_{\text{style}} \;-\; \mathbf{F}^{\text{HF}}_{\text{replaced}}\big). \tag{17}$$

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

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

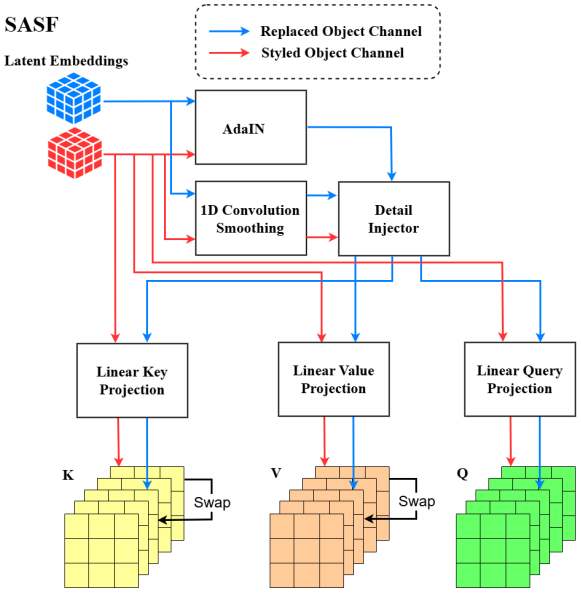

Figure 6: SASF Flowchart: Self-Attention Style Fusion incorporates style prompts by injecting high-frequency details via DSIN and substituting textual Key/Value matrices, ensuring fine-grained style modulation during the diffusion process.

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.

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:

$$\mathbf{K}_{\text{tar}} \leftarrow \mathbf{K}_{\text{sty}}, \qquad \mathbf{V}_{\text{tar}} \leftarrow \mathbf{V}_{\text{sty}}. \tag{18}$$

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.

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. 9).

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. (2025), SeedEdit Shi et al. (2024), LEDITS++ Brack et al. (2024) (CVPR 2024), StyleAligned Hertz et al. (2024) (CVPR 2024), TurboEdit Deutch et al. (2024), and IP2P Brooks et al. (2023) (CVPR 2023). 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. 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:

[1] `https://unsplash.com/`

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)	0.2393	0.2120	0.2866	0.5814
TurboEdit Deutch et al. (2024)	0.3261	0.1984	0.2781	0.6125
LEDITS++ Brack et al. (2024) (CVPR 2024)	0.3980	0.2078	0.2834	0.6145
SeedEdit Shi et al. (2024)	0.5612	0.2096	0.2966	0.6381
Step1X-Edit Liu et al. (2025)	0.7352	0.2120	0.2913	0.7024
CAOF	**0.8388**	0.2014	0.2937	0.8292

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)	0.2762	0.2039	0.2876	0.2236
TurboEdit Deutch et al. (2024)	0.3999	0.1954	0.2820	0.2090
StyleAligned Hertz et al. (2024) (CVPR 2024)	0.4360	0.1888	0.2915	0.1973
Step1X-Edit Liu et al. (2025)	0.4826	0.2145	0.2920	0.2170
SeedEdit Shi et al. (2024)	0.4899	0.1963	0.2915	0.2017
CAOF	**0.7102**	0.2014	0.2937	0.1976
CAOF+SASF	**0.9244**	0.2178	0.3022	0.2161

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

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

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

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 ($1-\text{LPIPS}_O$=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

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

replacement accuracy, blend consistency, and photographic integrity, a trade-off that CAOF is uniquely able to satisfy.

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 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, the blended identity, and the target style are distinctly visible while maintaining the original scene geometry

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

and background texture. In contrast, other methods exhibit clear limitations. In Figure 8, the blending process overemphasizes Batman, leading to a loss of Leonardo DiCaprio's distinct features. Similarly, in Figure 9, the original features are not well preserved: the iconic chocolate box held by Forrest Gump either disappears or is distorted, and the seated pose is unnaturally transformed into a standing position. 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. 8) 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

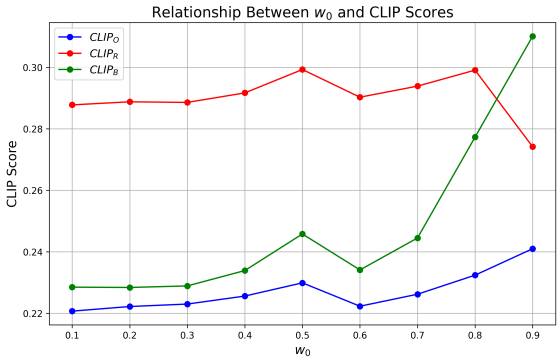

Figure 10: Variation of CLIP scores for original (P_o), replaced (P_r), and blend (P_b) object prompts as the blending coefficient w_0 changes. The curves illustrate CAOF's effectiveness in modulating blending strength, achieving the desired integration of the blend object while replacing the original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

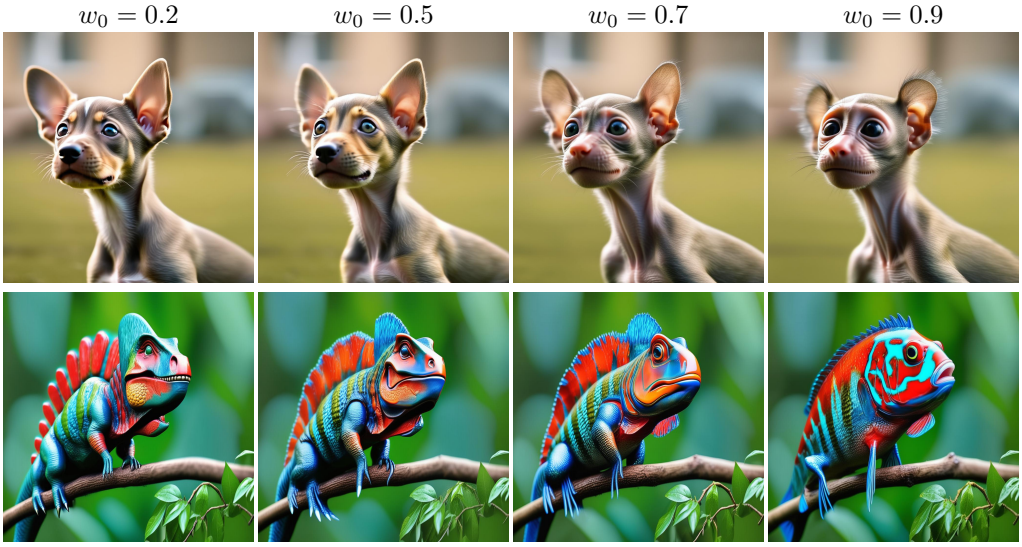

Figure 11: Object blending progression with varying blending coefficients w_0. Row 1: "**monkey**" blended into replaced "**puppy**" (originally "**alpaca**"). **Row 2**: "**fish**" blended into replaced "**dinosaur**" (originally "**chameleon**"). **Row 3**: "**ambulance**" blended into replaced "**jeep**"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

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.

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

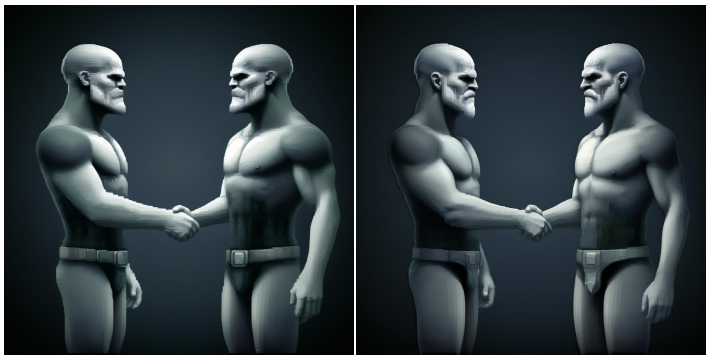

Figure 12: Pixel-art edit of "robot → knight" blended with "Thanos". Left: $\alpha = 0$, right: $\alpha = 0.5,\ \sigma = 2.5$.

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

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 12.

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

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.

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.

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*, 2024.

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.

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.

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

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

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.

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.

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.

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.

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.

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.

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.

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

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, 2024.

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.

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*, 2025.

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.

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.

Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Hancheng Min, Chris Callison-Burch, and René Vidal. Concept lancet: Image editing with compositional representation transplant. *arXiv preprint arXiv:2504.02828*, 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.

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*, 2025.

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.

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.

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.

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.

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.

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

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.

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.

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.

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*, 2024b.

Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Designdiffusion: High-quality text-to-design image generation with diffusion models. *arXiv preprint arXiv:2503.01645*, 2025a.

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

Zixuan Wang, Duo Peng, Feng Chen, Yuwei Yang, and Yinjie Lei. Training-free dense-aligned diffusion guidance for modular conditional image synthesis. *arXiv preprint arXiv:2504.01515*, 2025b.

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, 2024.

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.

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. *arXiv preprint arXiv:2501.06481*, 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.

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.

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.

Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2503.18352*, 2025a.

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*, 2025b.

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.

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. *arXiv preprint arXiv:2503.19839*, 2025.