FLASHEDIT: DECOUPLING SPEED, STRUCTURE, AND SEMANTICS FOR PRECISE IMAGE EDITING

Anonymous authors
Paper under double-blind review

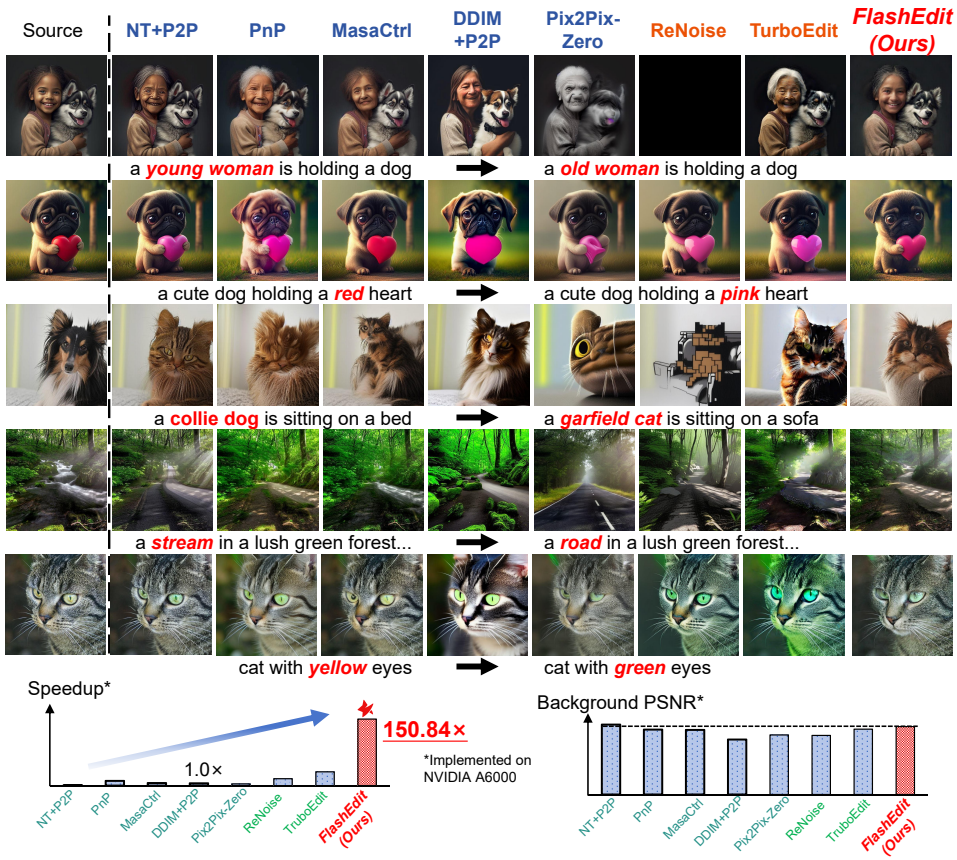

Figure 1: **FlashEdit** produces superior visual results for text-guided image editing, addressing background instability and semantic entanglement with an over $150\times$ speedup against DDIM (Song et al. (2020b)) + P2P (Hertz et al. (2022)).

ABSTRACT

Text-guided image editing with diffusion models has achieved remarkable quality but suffers from prohibitive latency, hindering real-world applications. We introduce **FlashEdit**, a novel framework designed to enable high-fidelity, real-time image editing. Its efficiency stems from three key innovations: (1) a **One-Step Inversion-and-Editing (OSIE)** pipeline that bypasses costly iterative processes; (2) a **Background Shield (BG-Shield)** technique that guarantees background preservation by selectively modifying features only within the edit region; and (3) a **Sparsified Spatial Cross-Attention (SSCA)** mechanism that ensures precise, localized edits by suppressing semantic leakage to the background. Extensive experiments demonstrate that FlashEdit maintains superior background consistency and structural integrity, while performing edits in under 0.2 seconds, which is an over $150\times$ speedup compared to prior multi-step methods. Our code will be made publicly available.

1 INTRODUCTION

Text-guided image editing with diffusion models (Brooks et al. (2023),Dong et al. (2023)) has demonstrated remarkable capabilities, allowing users to perform complex semantic modifications with high fidelity. The standard methodology is built upon a two-stage inversion-denoising pipeline: an initial inversion process maps a source image to its corresponding noise latent, which is then progressively denoised to generate the edited output according to a target prompt (Ju et al. (2023),Cao et al. (2023)). The objective is to achieve high fidelity in both content preservation and target prompt alignment, which often necessitates a computationally intensive, multi-step process.

Recent research has pursued several distinct strategies to improve accuracy and speed. To tackle the latency of the multi-step denoising process, methods based on model distillation have been proposed to enable editing in a faster way (Deutch et al. (2024)). These approaches must carefully address challenges such as mismatched noise statistics and insufficient editing strength that arise when adapting multi-step frameworks to fast samplers (Mokady et al. (2023b),Miyake et al. (2025)). To improve edit precision and prevent semantic leakage into the background, another category of work modifies the model's internal mechanisms, primarily by re-weighting or replacing attention maps to ensure the edit is spatially constrained (Fang et al. (2024); Xu et al. (2024)). Recognizing that the final edit quality is highly dependent on the starting point, other approaches focus on refining the inversion technique itself (Ju et al. (2023)). These methods aim to find a more accurate initial latent vector, with recent insights revealing that separating the objectives of content preservation and edit fidelity can yield significant performance gains and speedups (Wang et al. (2025b)).

However, these existing methods approach speed and quality as a trade-off rather than as interconnected components of a singular, complex control problem. They offer partial solutions like accelerating the sampler at the cost of inversion fidelity, or preserving the background without addressing the precision of the foreground edit. This results in a fragmented landscape of techniques that fail to deliver a solution that is simultaneously fast, robust, and precise. A truly practical editing framework requires a more holistic methodology that addresses control at every level of editing.

To address this multifaceted challenge, we introduce a novel editing methodology that establishes control at three progressively finer levels of granularity. At the foundational level, we tackle the macro-problem of **temporal control**. We propose a **One-Step Inversion-and-Editing (OSIE)** pipeline, built upon an "Anchor-and-Refine" training strategy, which conquers the prohibitive latency of prior work and makes real-time interaction possible. With this temporal control established, we address the meso-level problem of **spatial control**. Our **Background Shield (BG-Shield)** mechanism provides structural integrity by performing a surgical intervention in the self-attention layers. It uses a background memory and foreground-core querying to create a hard separation between edited and unedited regions, guaranteeing background stability. Finally, with speed and structure secured, we target the micro-level problem of **semantic control**. We develop **Sparsified Spatial Cross-Attention (SSCA)**, a refinement of the cross-attention mechanism that prunes irrelevant text tokens pre-softmax, ensuring the edit is guided by a clean, unambiguous semantic signal. Each component logically builds upon the last, forming a cohesive solution (Figure 1). Our main contributions can be summarized as follows:

- We propose a novel, multi-level methodology for image editing that cohesively integrates control over three distinct levels: the temporal latency of the pipeline, the spatial structure of the image, and the semantic content of the edit with an over $150\times$ speedup compared to prior multi-step methods.
- At the temporal level, we introduce the **One-Step Inversion-and-Editing (OSIE)** pipeline and its "Anchor-and-Refine" training strategy, which for the first time enables high-fidelity inversion for one-step diffusion models.
- At the spatial level, we propose **Background Shield (BG-Shield)**, a structural intervention in self-attention that uses memory caching and selective core querying to enforce pixel-perfect background preservation, ensuring the structural integrity of the edit.
- At the semantic level, we develop **Sparsified Spatial Cross-Attention (SSCA)**, a cross-attention mechanism that performs pre-softmax token pruning. This provides the final layer of fine-grained control, eliminating attribute bleeding and enabling precise edits with complex text prompts.

2 Related Works

2.1 Diffusion Models

Recent advances in image synthesis have been largely driven by diffusion models (Peebles & Xie (2023),Kulikov et al. (2024)), which have become a leading paradigm for generating high-fidelity images from text. The core mechanism involves an iterative denoising process that progressively refines a random noise vector into a coherent image conditioned on a text prompt. A landmark contribution in this area is Stable Diffusion (Rombach et al. (2021)), a Latent Diffusion Model (LDM) (Rombach et al. (2022)) that performs the computationally intensive denoising process in a lower-dimensional latent space, making the technology widely accessible. Parallel to this, alternative frameworks have emerged, such as Flow Matching models like Flux (Labs (2024)). Instead of an iterative refinement process, these models learn to map noise to an image via a more direct, straight-line trajectory, representing a different theoretical foundation for high-quality generative modeling.

To mitigate the high computational cost of these iterative models, various acceleration techniques have been proposed. Model quantization (Li et al. (2024a;b;c); Yan et al. (2025b)), cache mechanism (Xu et al. (2025); Pan et al. (2025)), sparse attention (Li et al. (2025a)), pruning (Wang et al. (2025a),Yan et al. (2025a)), and distillation (Hinton et al. (2015)) are general acceleration techniques for deep learning model. In diffusion models, specifically, one primary category is *model quantization* (Li et al. (2025b)), which reduces memory footprint and computational load by converting full-precision model weights and activations into lower-bit representations. Another category involves *cache mechanisms* (Liu et al. (2025); Xu et al. (2018)), which enhance inference efficiency by exploiting temporal redundancy. These methods reuse intermediate features computed at earlier denoising steps to avoid redundant calculations in later steps. While effective in isolation, recent work like QuantCache (Wu et al. (2025)) demonstrates a unified framework can yield greater gains.

2.2 Editing Models

The task of editing real images with pre-trained generative models introduces the fundamental challenge of *inversion*: finding a latent representation that can faithfully reconstruct a given source image. This problem was first extensively studied in the context of Generative Adversarial Networks (GAN) Inversion (Wang et al. (2022),Zhu et al. (2020),Zhu et al. (2016)). In comparison, **DDIM Inversion** (Song et al. (2020b)) provides a deterministic method to find a corresponding noise latent for a source image. Once this latent is obtained, various editing mechanisms are employed during the denoising process to apply the desired changes. A prominent family of methods focuses on *attention control*, where the cross-attention maps between text and image are manipulated. For example, to change a "photo of a red car" to a "blue car," Prompt-to-Prompt (Hertz et al. (2022)) identifies the attention weights corresponding to the word "red" and replaces them with those for "blue," preserving the attention for "car" and the background. Another powerful technique is *feature injection*, exemplified by Plug-and-Play (PnP) (Zhang et al. (2021)). To preserve the identity of a subject, PnP injects the self-attention features—which encode structure and appearance—from the source image's generation process into the edited one. A third approach is *mask-based editing*, where methods like DiffEdit (Couairon et al. (2022)) generate a mask indicating the region to be altered and then apply the denoising process only within that area. Despite these advances, a core challenge persists in perfectly disentangling the edited foreground from the unedited background.

3 Method

3.1 One-Step Inversion-and-Editing

Challenge: A Dual-Constraint Optimization Problem. The task of learning an effective inversion mapping is fundamentally a dual-constraint optimization problem. The predicted noise latent, ε_{inv}, must simultaneously satisfy two competing objectives. The first is a *fidelity constraint*, requiring ε_{inv} to encode sufficient information to perfectly reconstruct the source image. The second is a *distributional constraint*, requiring ε_{inv} to adhere to the generator's prior distribution, $\mathcal{N}(0, I)$, to ensure editability. While both constraints can be explicitly supervised when using synthetic data, the distributional constraint becomes non-trivial and unsupervised for real-world images where the ground-truth noise is unknown. Naively optimizing for fidelity alone causes a severe violation of the distributional constraint, leading to uneditable latents.

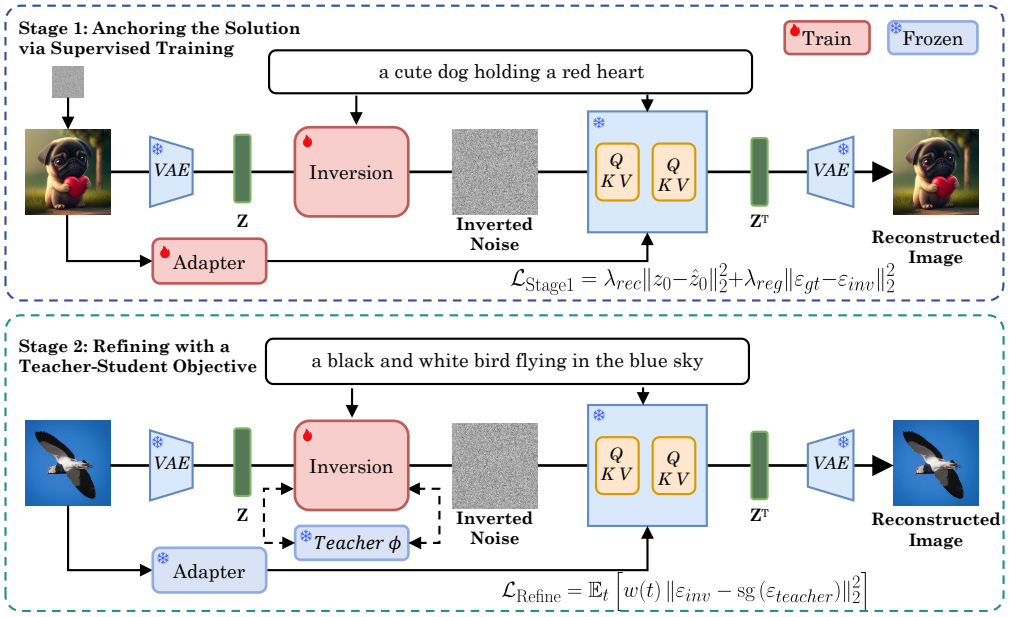

Figure 2: Overview of our **One-Step Inversion-and-Editing framework**, which introduces a direct image conditioning branch, trained via a two-stage "Anchor-and-Refine" strategy that uses direct supervision for synthetic data (Stage 1) and a teacher-student objective for real images (Stage 2).

Motivation. Our motivation is to design a training strategy that explicitly decouples and progressively solves these two constraints. We posit that the network must first learn to jointly satisfy both objectives in a fully-supervised setting before it can be adapted to handle the unsupervised nature of real-image inversion. This leads to our "Anchor-and-Refine" approach. The "Anchor" stage uses synthetic data to ground the network in a parameter space that respects both constraints. The "Refine" stage then adapts this mapping to real images, where we introduce a powerful generative prior from a teacher model to act as a proxy for the now-unsupervised distributional constraint. This ensures that even for real images, fidelity is pursued without sacrificing editability.

Proposed Method. Shown in Figure 2, our primary architectural modification is designed to resolve a fundamental tension in the inversion process. The inverted noise vector is typically burdened with two conflicting tasks: perfectly preserving the source image's identity and remaining generic enough for subsequent editing. To decouple these roles, we introduce a dedicated visual adapter which provides the decoder D with a direct visual information from the source image.

This way, the decoder's output—the reconstructed latent z'—becomes a function of three distinct inputs: the inverted noise n, the text condition c_t, and the explicit image features c_i. By directly supplying the visual identity via c_i, we liberate the noise vector n from its strict reconstruction duty. It can now remain closer to a pure Gaussian distribution, drastically improving its malleability for downstream editing tasks.

Stage 1: Anchoring the Solution via Supervised Training. The first stage aims to find a robust initialization, or "anchor," for the inversion network I_θ. We use a synthetic dataset of (ε_{gt}, z_0) tuples from the base generator G, which allows for direct and strong supervision. The training objective is twofold:

$$\mathcal{L}_{\text{Stage1}} = \lambda_{rec}\|z_0 - \hat{z}_0\|_2^2 + \lambda_{reg}\|\varepsilon_{gt} - \varepsilon_{inv}\|_2^2. \tag{1}$$

The regression term $\mathcal{L}_{reg}$ is critical in this stage. It constrains the network to a region of the loss landscape where its outputs naturally conform to the target distribution $\mathcal{N}(0, I)$. During this stage, we train both the inversion network I_θ and the newly introduced image adapter. This teaches the adapter how to effectively provide visual priors that aid in reconstruction. This anchoring step prevents the network from converging to trivial solutions in the next stage.

Stage 2: Refining with a Teacher-Student Objective. With the network anchored, the second stage refines its mapping for the complexities of real-world images where the ground-truth noise ε_{gt} is unknown. To prevent the distribution of ε_{inv} from drifting, we introduce a regularization scheme

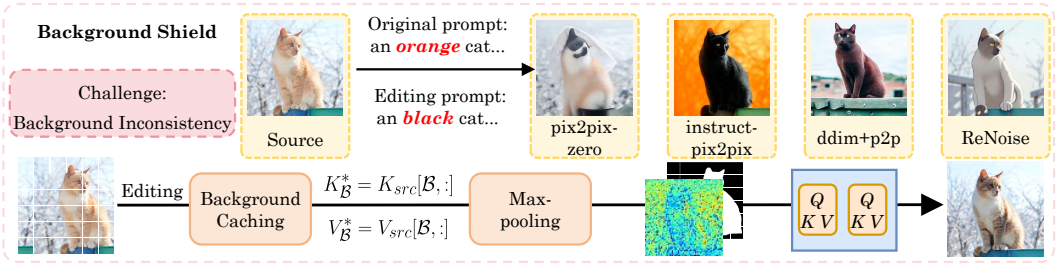

Figure 3: Illustration of our **Background Shield (BG-Shield) mechanism.** The top of the figure illustrates the problem of *background inconsistency* in standard editing, while the bottom details the pipeline of our method designed to solve it.

framed as a **teacher-student distillation** process. We leverage a pre-trained "teacher" model, ϕ, to provide a dynamic, supervisory signal for our "student" inversion network, I_θ.

For each real image latent z_0, we first create a noisy version $z_t = \alpha_t z_0 + \sigma_t \varepsilon_{inv}$ at a random timestep t. The teacher model ϕ then predicts the noise from this input, yielding a "pseudo-ground-truth" target, $\varepsilon_{teacher}$:

$$\varepsilon_{teacher} = \phi(\alpha_t z_0 + \sigma_t \varepsilon_{inv}, t, c). \tag{2}$$

We then define a refinement loss, $\mathcal{L}_{\text{Refine}}$, that minimizes the L2 distance between our network's output ε_{inv} and the teacher's prediction. Crucially, we treat the teacher's output as a fixed target by applying a stop-gradient operator.

$$\mathcal{L}_{\text{Refine}} = \mathbb{E}_t \left[w(t) \left\| \varepsilon_{inv} - \text{sg}\left(\varepsilon_{teacher}\right) \right\|_2^2 \right], \tag{3}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. This formulation turns the problem into a simple regression task where the student (I_θ) is trained to produce a noise latent that the teacher (ϕ) would have predicted. This distillation-style loss effectively regularizes the training, ensuring that for any given real image, the predicted noise ε_{inv} is a solution that is not only perceptually accurate (as enforced by a parallel perceptual loss) but also highly plausible under the teacher's learned world model.

3.2 Background Shield

Challenge: Background Inconsistency. A critical challenge in localized image editing is maintaining strict background consistency. We observe that even with precise masks, many methods fail at this task. For instance, in Figure 3 when performing a seemingly simple edit such as changing "an orange cat" to "a black cat", the background suffers from unintended alterations, leading to shifts in color, lighting, or style. We identify the root cause of this instability as the inherent nature of the self-attention mechanism. As a global operator that computes all-to-all relationships between image tokens, it allows the strong semantic signal from the foreground edit to propagate and contaminate the background features, undermining the goal of a truly localized edit.

Motivation. Having identified the global nature of self-attention as the cause of this background inconsistency, our motivation is to move beyond merely scaling influences and propose a direct structural intervention. To achieve background stability, a hard constraint that structurally isolates the background from the editing process is required. We introduce **Background Shield (BG-Shield)**, a method designed to enforce this consistency by replacing the background's feature computation with a direct recall from a "background memory".

Proposed Method. Shown in Figure 3, BG-Shield operates as a two-pass mechanism within self-attention layers. Let $X \in \mathbb{R}^{S \times D}$ be the input feature sequence, and let a binary mask $M \in \{0, 1\}^S$ define the foreground indices $\mathcal{F}$ and background indices $\mathcal{B}$.

Background Memory Caching. During a forward pass with the source prompt c_{src}, we compute the Key and Value matrices, K_{src}, V_{src}. We then extract and cache the background-specific key-value pairs:

$$K^*_{\mathcal{B}} = K_{src}[\mathcal{B}, :], \quad V^*_{\mathcal{B}} = V_{src}[\mathcal{B}, :]. \tag{4}$$

This cached memory, $(K^*_{\mathcal{B}}, V^*_{\mathcal{B}})$, serves as a high-fidelity record of the original background state.

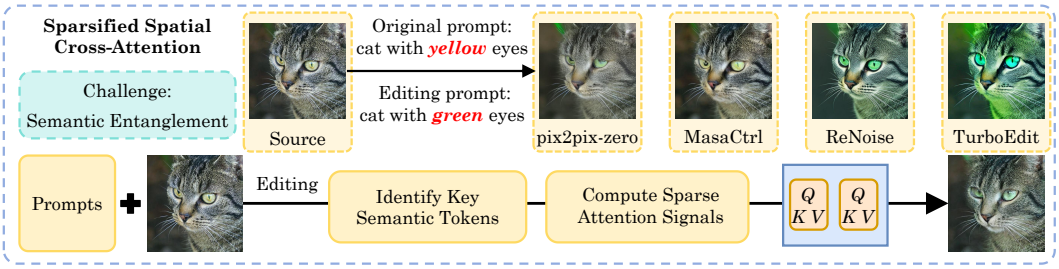

Figure 4: Illustration of our **Sparsified Spatial Cross-Attention (SSCA)** method resolving *semantic entanglement*. The top row demonstrates how standard attention fails on precise edits, resulting in edit attenuation and attribute leakage. The bottom row details our SSCA mechanism, which prevents this by computing attention only over a subset of relevant text tokens to ensure a clean edit.

Mask-Driven Recomposition and Selective Querying. During the editing pass with the target prompt c_{tgt}, we compute new queries, keys, and values ($Q_{tgt}, K_{tgt}, V_{tgt} \in \mathbb{R}^{S \times d_k}$). We then construct a spatially-aware, full key-value set, K_{full}, V_{full}, by combining the background memory with the current foreground features:

$$K_{full}[j,:] = \begin{cases} K^*_{\mathcal{B}}[\text{rank}_{\mathcal{B}}(j),:] & \text{if } j \in \mathcal{B} \\ K_{tgt}[j,:] & \text{if } j \in \mathcal{F} \end{cases}, \quad V_{full}[j,:] = \begin{cases} V^*_{\mathcal{B}}[\text{rank}_{\mathcal{B}}(j),:] & \text{if } j \in \mathcal{B} \\ V_{tgt}[j,:] & \text{if } j \in \mathcal{F} \end{cases}, \tag{5}$$

where $\text{rank}_{\mathcal{B}}(j)$ ensures correct positional alignment. To mitigate boundary artifacts, we introduce a *foreground core* by applying a morphological erosion to the mask M. This is implemented using a 2D max-pooling operation (with kernel size k, stride s, and padding p) on the inverted mask. The resulting core mask M_{core} is binarized with a threshold τ to yield the core index set $\mathcal{F}_c \subset \mathcal{F}$:

$$M_{core} = \mathbf{1} - \text{MaxPool2d}(\mathbf{1} - M, kernel\_size, stride, padding), \tag{6}$$

$$\mathcal{F}_c = \{i \mid (M_{core})_i > \tau\}. \tag{7}$$

The attention computation is then performed *only* for queries within this core region. Let $Q_{tgt,c} = Q_{tgt}[\mathcal{F}_c,:]$ be the subset of queries corresponding to the core indices. The attention output for this region, $H_c \in \mathbb{R}^{|\mathcal{F}_c| \times d_k}$, is computed as:

$$H_c = \text{softmax}\left(\frac{Q_{tgt,c}K^T_{full}}{\sqrt{d_k}}\right)V_{full}. \tag{8}$$

The full output matrix $H \in \mathbb{R}^{S \times d_k}$ is then constructed by scattering the computed values H_c back to their original positions, while all other positions corresponding to the background and boundary are set to zero.

Residual Fusion. The sparse output matrix H is projected and added back to the input features: $Y = \text{Proj}(H) + X$. Since $H_i = 0$ for all $i \notin \mathcal{F}_c$, this step functions as an identity map for the background and boundary regions, ensuring they are perfectly preserved.

3.3 Sparsified Spatial Cross-Attention

Challenge: Semantic Entanglement in Image Editing. A key challenge in precise editing is *semantic entanglement*, where textual attributes are not cleanly bound to their intended objects. This is clearly demonstrated in Figure 4, where the task is to change "a cat with yellow eyes" to "a cat with green eyes." Standard models often fail, resulting in either *edit attenuation*, where the eyes are incompletely colored, or significant *attribute leakage*, causing an unnatural green tint to bleed onto the cat's face. This failure stems from the competitive nature of the softmax function in **cross-attention**. It forces all text tokens to compete for influence over each pixel, allowing the powerful "green" signal to suppress the essential structural tokens like "cat," which leads to the incorrect generalization.

Motivation. Based on this diagnosis, we contend that semantic concepts must be disentangled *before* the attention softmax allows them to interfere. Our motivation is to implement a **pre-emptive disentanglement** strategy. Instead of allowing all text tokens to participate in the attention calculation for the foreground, we introduce Sparsified Spatial Cross-Attention (SSCA), a method that

Figure 5: Qualitative comparison of editing results. Each row corresponds to a unique editing task, with the source image displayed in the first column and the source/target prompts listed below.

forces the softmax to operate only on a clean, disentangled subset, thus preventing attribute leakage at its source.

Proposed Method. Our Sparsified Spatial Cross-Attention (SSCA) mechanism fundamentally redefines the text attention computation by breaking it down into three sequential steps: identifying key semantic tokens, computing a focused sparse attention signal, and integrating this signal into the final feature map, shown in Figure 4.

Identifying Key Semantic Tokens. Before computing attention, we first identify the most relevant tokens from the text prompt y for the given edit region M. We compute the similarity between the set of image queries within the mask, $Q_{l,M}$, and all text keys K_y. The top-k text key-value pairs that exhibit the highest aggregate similarity are selected. This pre-selection step acts as a filter, creating a task-relevant subset of textual information, denoted as (K_y^k, V_y^k).

Computing Sparse Attention Signals. With the pruned set of text tokens, we then compute a sparse attention result, A_{sparse}, *only* for the image queries within the edit region, $Q_{l,M}$. This ensures that the computationally expensive attention operation is focused where it is needed most.

$$A_{\text{sparse}} = \text{softmax}\left(\frac{Q_{l,M}(K_y^k)^T}{\sqrt{d}}\right)V_y^k. \tag{9}$$

The resulting matrix $A_{\text{sparse}} \in \mathbb{R}^{|\mathcal{F}|\times d}$ contains a highly precise and disentangled guidance signal, where $|\mathcal{F}|$ is the number of foreground pixels. **Constructing and Integrating the Full Attention Matrix.** The sparse signal A_{sparse} must be placed into a full-size matrix to be used in the model. We construct the final text attention matrix, $A_{\text{SSCA}} \in \mathbb{R}^{S\times d}$, by scattering the values from A_{sparse} into a zero matrix according to the mask indices $\mathcal{F}$. This structurally enforces that the text prompt has zero influence on the background.

$$A_{\text{SSCA}}[i,:] = \begin{cases} A_{\text{sparse}}[\text{rank}_{\mathcal{F}}(i),:] & \text{if } i \in \mathcal{F} \\ \mathbf{0} & \text{if } i \notin \mathcal{F} \end{cases}, \tag{10}$$

Table 1: Comprehensive comparison of editing quality, evaluating background preservation and CLIP similarity across various methods.

Method		Background Preservation				CLIP Similarity	
Inverse	**Editing**	**PSNR** ↑	**LPIPS**$_{\times 10^3}$ ↓	**MSE**$_{\times 10^4}$ ↓	**SSIM**$_{\times 10^2}$ ↑	**Whole** ↑	**Edited** ↑
DDIM	P2P	17.87	208.80	219.88	71.14	25.01	22.44
NT-Inv	P2P	27.03	60.67	35.86	84.11	24.75	21.86
DDIM	MasaCtrl	22.17	106.62	86.97	79.67	23.96	21.16
Direct Inversion	MasaCtrl	22.64	87.94	81.09	81.33	24.38	21.35
DDIM	P2P-Zero	20.44	172.22	144.12	74.67	22.80	20.54
Direct Inversion	P2P-Zero	21.53	138.98	127.32	77.05	23.31	21.05
DDIM	PnP	22.28	113.46	83.64	79.05	25.41	22.55
Direct Inversion	PnP	22.46	106.06	80.45	79.68	25.41	22.62
ReNoise(SDXL)		20.85	176.84	51.78	72.44	24.41	21.88
TurboEdit		22.51	107.27	9.32	80.09	25.49	21.82
FlashEdit		25.29	62.55	4.36	83.21	25.43	22.13
FlashEdit(w/ GT masks)		25.26	62.78	4.39	83.08	25.53	22.25

where $\text{rank}_{\mathcal{F}}(i)$ maps the global index to its local index within the foreground. Finally, this purified text guidance is integrated with the source image condition, A_{img}, to compute the updated hidden state h_l:

$$h_l = s_y \cdot A_{\text{SSCA}} + s_{edit} \cdot M \odot A_{img} + s_{non-edit} \cdot (1 - M) \odot A_{img}. \tag{11}$$

This multi-step process provides a maximally disentangled and precise guidance signal for the edit.

4 Experiment

4.1 Experimental Setup

Implementation Details. Our inversion network, I_θ, is initialized from SwiftBrush (Nguyen & Tran (2024),Dao et al. (2024)). Inspired by (Song et al. (2024),Ye et al. (2023),Zhang et al.), the image conditioning branch is based on an adapter, utilizing a pre-trained CLIP image encoder. We train the model using the Adam optimizer (Kingma & Ba (2014)) with a learning rate of 2e-5, weight decay of 2e-4, and an Exponential Moving Average (EMA). Anchoring the Solution via Supervised Training runs for 150k iterationson synthetic data from SwiftBrush. Refining with a Teacher-Student Objective continues for 200k iterations using real images from CommonCanvas (Gokaslan et al. (2024)). All experiments were conducted on a single NVIDIA A6000 GPU.

Metrics. We evaluate our method on the PieBench benchmark (Zhang et al. (2021)), which features 700 samples across 10 editing types. We report metrics along two primary axes. As for **Background Preservation**, We compute PSNR (Huynh-Thu & Ghanbari (2008)), LPIPS (Zhang et al. (2018)), MSE and SSIM (Wang et al. (2004)) on the unedited regions to measure fidelity to the source image. As for **Semantic Alignment**, We report CLIP-Whole (Radford et al. (2021)) for prompt-image alignment and CLIP-Edited (Radford et al. (2021)) for alignment within the masked edit region.

Baselines. We compare our method against state-of-the-art **multi-step** and **few-step** baselines. For multi-step methods, we evaluate Prompt-to-Prompt (P2P) (Hertz et al. (2022)), MasaCtrl (Cao et al. (2023)), Pix2Pix-Zero (Parmar et al. (2023)), and Plug-and-Play (PnP) (Zhang et al. (2021)), paired with powerful inversion techniques like DDIM (Song et al. (2020a)), Null-text Inversion (NT-Inv) (Mokady et al. (2023a)), and Direct Inversion (Ju et al. (2023)). For few-step methods, we compare against Renoise (Garibi et al. (2024)) and TurboEdit (Deutch et al. (2024)).

4.2 Quantitative Analysis

As shown in Table 1, our method establishes a new state-of-the-art for accelerated editing. FlashEdit significantly outperforms recent **few-step methods** like ReNoise (Garibi et al. (2024)) and Tur-

Table 2: **Ablation Study on Core Model Components.** We evaluate the contribution of each module by measuring the impact on background preservation and semantic similarity (CLIP Score). The final row represents our full method.

Components			Background Preservation				CLIP Similarity	
OSIE	**BG-Shield**	**SSCA**	**PSNR↑**	**LPIPS$_{\times 10^3}$↓**	**MSE$_{\times 10^4}$↓**	**SSIM$_{\times 10^2}$↑**	**Whole↑**	**Edited↑**
✓	-	-	23.33	92.37	6.60	79.97	24.14	21.23
✓	✓	-	24.63	75.36	5.01	81.65	24.77	21.22
✓	✓	✓	25.29	62.55	4.36	83.21	25.43	22.13

boEdit (Deutch et al. (2024)) across all reported metrics. Crucially, it also achieves quality on par with, and in several metrics superior to, top-performing but prohibitively slow **multi-step methods**. This high fidelity is delivered with an extraordinary efficiency gain of over **150**× (Table 3). Furthermore, an experiment using ground-truth (GT) masks reveals a negligible performance difference, confirming the high accuracy of our self-guided masking mechanism.

4.3 Qualitative Analysis

Visual comparisons in Figure 5 reinforce our quantitative findings. The outputs from FlashEdit consistently exhibit high semantic fidelity to the target prompt while maintaining pristine background integrity, avoiding the "bleeding" artifacts common in other methods. In contrast, other baselines often display noticeable quality degradation or fail to preserve background details. FlashEdit is unique in providing both state-of-the-art visual quality and the real-time performance that multi-step methods lack.

Table 3: Efficiency comparison of individual editing methods, with the denoising steps and speedup factor for each specific combination.

Method		Denoising Steps	Speedup
Inverse	**Editing**		
DDIM	P2P	Multi-steps	**1.00**×
NT-Inv	P2P		0.19×
DDIM	MasaCtrl		1.12×
Direct Inversion	MasaCtrl		0.88×
DDIM	P2P-Zero		0.73×
Direct Inversion	P2P-Zero		0.73×
DDIM	PnP		2.06×
Direct Inversion	PnP		2.03×
ReNoise(SDXL)		Few-steps	5.08×
TurboEdit			19.68×
FlashEdit(Ours)		**One-step**	**150.84**×

4.4 Ablation Studies

To validate the contribution of each component in our framework, we conduct a comprehensive ablation study, with the results presented in Table 2. Our baseline, consisting of the **OSIE** pipeline alone, establishes a strong performance foundation. Integrating **BG-Shield** brings a marked improvement across background preservation metrics, confirming its effectiveness in isolating background features. The final addition of **SSCA** further boosts metrics. It substantially enhances semantic alignment, evidenced by a large increase in the CLIP-Edited score, which validates our pre-softmax token pruning strategy. SSCA also improves reconstruction quality, suggesting a synergistic effect where cleaner textual guidance benefits the entire process. This demonstrates that all three components are critical and work in concert to achieve the final state-of-the-art performance of **FlashEdit**.

5 Conclusion

This paper introduces **FlashEdit**, a new paradigm for text-guided image editing that redefines the performance standard for real-time generative applications. We demonstrate that the long-standing trade-off between speed and quality is not fundamental but can be overcome with a holistic, multi-level control strategy. Our approach begins by establishing temporal control with a foundational **OSIE** pipeline for one-step inversion and editing. It then enforces spatial control with **BG-Shield** and fine-grained semantic control with **SSCA**. Together, these components transform diffusion-based editing from a slow, offline process into an interactive and expressive creative tool.

Ethics Statement

The research conducted in the paper conforms, in every respect, with the ICLR Code of Ethics.

Reproducibility Statement

We have provided implementation details in Sec. 4. We will also release all the code and models.

LLM Usage Statement

Large Language Models (LLMs) were used solely for polishing writing. They did not contribute to the research content or scientific findings of this work.

References

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.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22560–22570, October 2023.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. In *European Conference on Computer Vision*, pp. 176–192. Springer, 2024.

Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024.

Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7430–7440, 2023.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*, 2024.

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*, pp. 395–413. Springer, 2024.

Aaron Gokaslan, A Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. Commoncanvas: Open diffusion models trained on creative-commons images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8250–8260, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801, 2008. doi: 10.1049/el:20080522. URL https://digital-library.theiet.org/doi/abs/10.1049/el%3A20080522.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Jinhao Li, Jiaming Xu, Shan Huang, Yonghua Chen, Wen Li, Jun Liu, Yaoxiu Lian, Jiayi Pan, Li Ding, Hao Zhou, et al. Large language model inference acceleration: A comprehensive hardware perspective. *arXiv preprint arXiv:2410.04466*, 2024a.

Jinhao Li, Jiaming Xu, Shiyao Li, Shan Huang, Jun Liu, Yaoxiu Lian, and Guohao Dai. Fast and efficient 2-bit llm inference on gpu: 2/4/16-bit in a weight matrix with asynchronous dequantization. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pp. 1–9, 2024b.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, et al. Radial attention: O (nlog n) sparse attention with energy decay for long video generation. *arXiv preprint arXiv:2506.19852*, 2025a.

Zhiteng Li, Xianglong Yan, Tianao Zhang, Haotong Qin, Dong Xie, Jiang Tian, zhongchao shi, Linghe Kong, Yulun Zhang, and Xiaokang Yang. Arb-llm: Alternating refined binarizations for large language models, 2024c. URL https://arxiv.org/abs/2410.03129.

Zhiteng Li, Hanxuan Li, Junyi Wu, Kai Liu, Linghe Kong, Guihai Chen, Yulun Zhang, and Xiaokang Yang. Dvd-quant: Data-free video diffusion transformers quantization, 2025b. URL https://arxiv.org/abs/2505.18663.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with taylorseers. *arXiv preprint arXiv:2503.06923*, 2025.

Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2063–2072. IEEE, 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, 2023a.

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, 2023b.

Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7807–7816, 2024.

Jiayi Pan, Jiaming Xu, Yongkang Zhou, and Guohao Dai. Specdiff: Accelerating diffusion model inference with self-speculation, 2025. URL https://arxiv.org/abs/2509.13848.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020a. URL https://arxiv.org/abs/2010.02502.

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

Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation. 2024.

Hanzhen Wang, Jiaming Xu, Jiayi Pan, Yongkang Zhou, and Guohao Dai. Specprune-vla: Accelerating vision-language-action models via action-aware self-speculative pruning, 2025a. URL https://arxiv.org/abs/2509.05614.

Jia Wang, Jie Hu, Xiaoqi Ma, Hanghang Ma, Xiaoming Wei, and Enhua Wu. Image editing with diffusion models: A survey. *arXiv preprint arXiv:2504.13226*, 2025b.

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Junyi Wu, Zhiteng Li, Zheng Hui, Yulun Zhang, Linghe Kong, and Xiaokang Yang. Quantcache: Adaptive importance-guided quantization with hierarchical latent and layer caching for video generation, 2025. URL https://arxiv.org/abs/2503.06545.

Jiaming Xu, Jiayi Pan, Yongkang Zhou, Siming Chen, Jinhao Li, Yaoxiu Lian, Junyi Wu, and Guohao Dai. Specee: Accelerating large language model inference with speculative early exiting. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture*, pp. 467–481, 2025.

Mengwei Xu, Mengze Zhu, Yunxin Liu, Felix Xiaozhu Lin, and Xuanzhe Liu. Deepcache: Principled cache for mobile deep vision. In *Proceedings of the 24th annual international conference on mobile computing and networking*, pp. 129–144, 2018.

Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Xiaoyu Kong, Jintao Li, Oliver Deussen, and Tong-Yee Lee. Headrouter: A training-free image editing framework for mm-dits by adaptively routing attention heads. *arXiv preprint arXiv:2411.15034*, 2024.

Xianglong Yan, Zhiteng Li, Tianao Zhang, Linghe Kong, Yulun Zhang, and Xiaokang Yang. Recalkv: Low-rank kv cache compression via head reordering and offline calibration. *arXiv preprint arXiv:2505.24357*, 2025a.

Xianglong Yan, Tianao Zhang, Zhiteng Li, and Yulun Zhang. Progressive binarization with semi-structured pruning for llms, 2025b. URL https://arxiv.org/abs/2502.01705.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pp. 592–608. Springer, 2020.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pp. 597–613. Springer, 2016.