PANDORA: Pixel-wise Attention Dissolution and Latent Guidance for Zero-Shot Object Removal

immediate

***Abstract*—Removing objects from natural images remains a formidable challenge, often hindered by the inability to synthesize semantically appropriate content in the foreground while preserving background integrity. Existing methods often rely on fine-tuning, prompt engineering, or inference-time optimization, yet still struggle to maintain texture consistency, produce rigid or unnatural results, lack precise foreground-background disentanglement, and fail to flexibly handle multiple objects—ultimately limiting their scalability and practical applicability. In this paper, we propose a zero-shot object removal framework that operates directly on pre-trained text-to-image diffusion models—requiring no fine-tuning, no prompts, and no optimization. At the core is our Pixel-wise Attention Dissolution, which performs fine-grained, pixel-wise dissolution of object information by nullifying the most correlated keys for each masked pixel. This operation causes the object to vanish from the self-attention flow, allowing the coherent background context to seamlessly dominate the reconstruction. To complement this, we introduce Localized Attentional Disentanglement Guidance, which steers the denoising process toward latent manifolds that favor clean object removal. Together, Pixel-wise Attention Dissolution and Localized Attentional Disentanglement Guidance enable precise, non-rigid, scalable, and prompt-free multi-object erasure in a single pass. Experiments show our method outperforms state-of-the-art methods even with fine-tuned and prompt-guided baselines in both visual fidelity and semantic plausibility. The project page is available at here.**

***Index Terms*—Stable Diffusion, Multi-Object Removal, Zero-Shot Algorithm, Attention**

I. Introduction

Recent advances in diffusion models (DMs) [1]–[12] have revolutionized generative image modeling, enabling high-quality image synthesis guided by textual prompts or other conditions. These models have opened new opportunities for semantic image editing [13]–[19], particularly in object removal—a longstanding and challenging task [20]–[23] that involves selectively erasing undesired regions from an image and reconstructing the void with content that blends naturally into the scene. This task goes beyond typical inpainting by requiring both precise object elimination and faithful restoration that aligns with the scene's semantics and visual flow.

Traditional object removal techniques have mainly utilized patch-based approaches [24]–[26] or Generative Adversarial Networks (GANs) [21], [27]–[30]. Patch-based methods often result in inconsistencies by filling in the masked areas with patches from other parts of the image, which can lead to unnatural blending with the surrounding regions. GANs, while enhancing realism, still face challenges with artifact generation and lack versatility in handling complex scenes.

More recent methods based on Latent Diffusion Models (LDMs) [11] have significantly improved the realism of image synthesis. However, when adapted to object removal, these models often struggle to reliably eliminate target objects. One such adaptation, Stable Diffusion Inpainting [11], extends the base model by conditioning on a binary mask to enable end-to-end inpainting. Despite its tailored design, the model frequently produces incomplete removals or introduces unexpected artifacts, even with considerable fine-tuning effort. Moreover, they require extensive prompt engineering [31] or fully fine-tuning such large-scale models [11] is often impractical in low-resource settings, limiting their accessibility and scalability in broader research applications. Some methods [22], [32] have been proposed zero-shot methods for object removal, but they often struggle to handle multi-object removal and lack precise control over the interaction between foreground and background elements during the generation process, ultimately leading to incomplete removals or noticeable artifacts.

To address these challenges, we propose PANDORA, a zero-shot object removal framework that operates directly on pre-trained diffusion models in a single pass, without any fine-tuning, prompt engineering, or inference-time optimization, thus fully leveraging their latent generative capacity for inpainting. Central to our approach are two synergistic modules: Pixel-wise Attention Dissolution (PAD) and Localized Attentional Disentanglement Guidance (LADG). Specifically, we first adopt the preservation adaptation module from CPAM [32] to invert the input image into the latent noise space and ensure that the background regions remain unaffected during the object removal process, allowing targeted and artifact-free editing. Then, we introduce Pixel-wise Attention Dissolution to achieve fine-grained control within the self-attention mechanism. PAD computes similarity scores between each query and all key positions, and dissolves the top-k most correlated keys through a percentile-based thresholding strategy. This dissolution disconnects each masked query pixel from its most correlated regions, effectively eliminating foreground dominance in the attention computation. As a result, object information is dissolved, allowing the masked areas to vanish and be naturally reconstructed with coherent background content in a precise, query-specific manner. Finally, we incorporate Localized Attentional Disentanglement Guidance to steer latent noise away from the object regions while preserving other areas, thereby refining the denoising trajectory in the latent space to effectively suppress residual artifacts and produce smoother, cleaner results. While PAD dissolves object

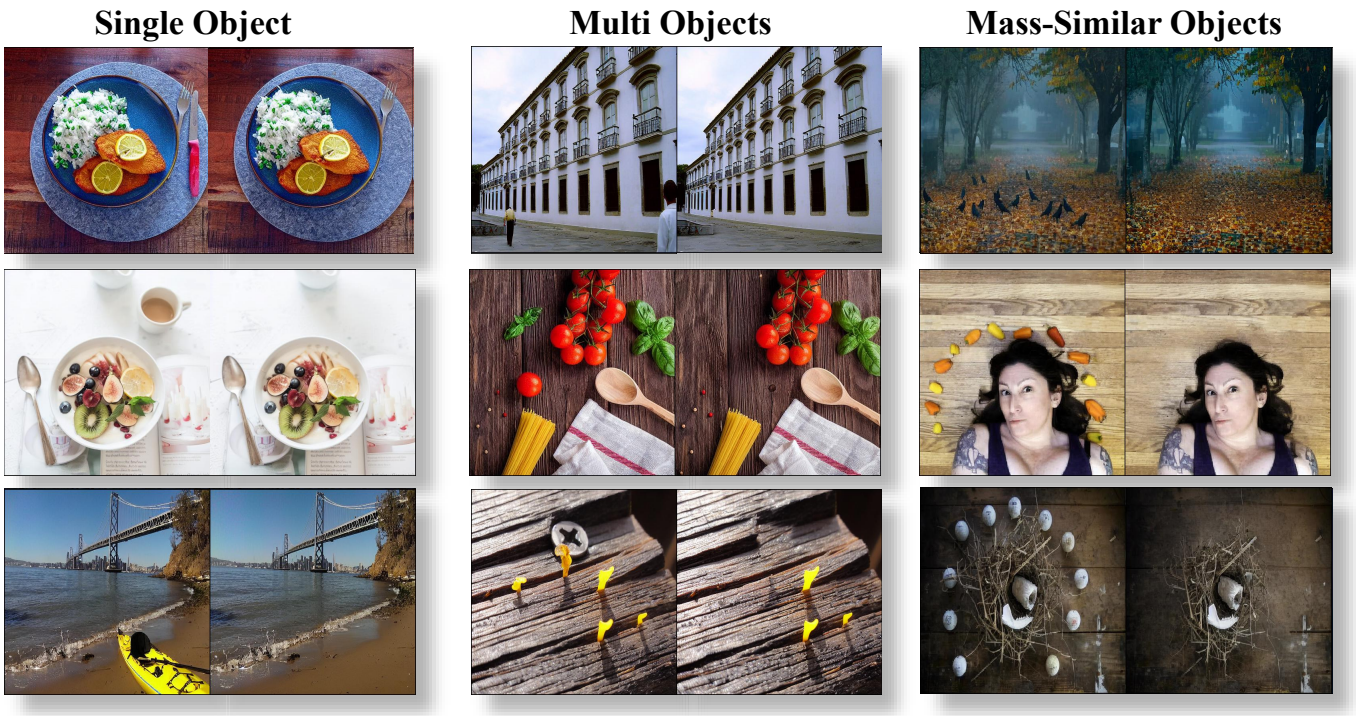

Fig. 1: **Our method enables prompt-free, fine-tuning-free object removal across various scenarios in a single forward.** Without requiring training or textual prompts, our approach handles diverse and challenging removal settings—from a single object to multiple similar or distinct targets and even densely packed similar objects—while preserving background fidelity and structural consistency.

information at the attention level, LADG complements it by reshaping the denoising trajectory in latent space, ensuring that masked regions are reconstructed coherently with their surrounding context.

PANDORA's full pipeline implemented on top of pre-trained Stable Diffusion [11], achieves architecture-agnostic, end-to-end object removal in a single pass. It excels in complex and cluttered scenes, supports multi-object erasure, and requires no prompts or training-time supervision. Our contributions are summarized as follows:

- We propose PANDORA, a novel zero-shot object removal framework that directly leverages pre-trained diffusion models to remove objects in a single pass without fine-tuning, prompt engineering, or inference-time optimization.
- We propose Pixel-wise Attention Dissolution (PAD), which dissolves object information by disconnecting masked query pixels from their most correlated keys in self-attention, enabling fine-grained removal and reconstruction with coherent background content.
- We present Localized Attentional Disentanglement Guidance (LADG), which steers latent noise away from object regions while preserving unaffected areas, refining the denoising trajectory to remove residual artifacts and produce cleaner outputs.
- Extensive experiments on a challenging benchmark demonstrate that our method achieves state-of-the-art zero-shot object removal performance, excelling in both single- and multi-object scenarios.

II. Related work

A. Attention-Guided Zero-Shot Image Editing

To reduce the need for expensive training on large datasets, tuning-free methods were proposed that leveraged frozen text-to-image models such as Stable Diffusion [11], enabling *zero-shot image editing* [33]–[35]. These approaches typically exploited attention mechanisms to preserve image content, such as structure or identity [15], [16], [36]–[38]. Some methods modified self-attention [16], [38], while others replaced cross-attention maps to guide the edit [15]. However, many approaches assumed similarity between the source and target prompts or objects [16], [38], [39]. MasaCtrl [37] enhanced flexibility by preserving query features and replacing key/value components using masks, but its simultaneous editing of both foreground and background limited its controllability in tasks that required preserving one region unchanged. Moreover, most methods affected the whole image due to global cross-attention conditioning, making local editing difficult. Some approaches [15], [40]–[42] attempted localized blending in the latent space but often ignored foreground–background interaction, leading to rigid edits. Despite their success in general editing, these methods remained underexplored for object removal, especially in selectively erasing specific regions

without affecting the surrounding content, and often required carefully crafted prompts to achieve satisfactory results.

B. Zero-Shot Object Removal

Zero-shot object removal methods generally operated by suppressing information from masked foreground regions during the diffusion process while compensating with background content to ensure visual plausibility. CPAM [32] provided flexible control over semantic preservation and foreground–background interaction, enabling object removal by suppressing masked regions and filling them with background content. However, it still attended to object-correlated areas, often leaving residual traces or failing to completely erase the target object. To address this issue, Attentive Eraser [22] directly scaled down attention weights within the masked region, thereby suppressing its semantic contribution during image synthesis and improving removal quality. Nonetheless, this approach applied a uniform downscaling to all query pixels in the object region, even though each pixel should have been treated differently based on its correlations. It also required manual tuning of the suppression parameter, which was highly sensitive—small values often failed to fully erase the object, while large values produced unnatural or degraded results.

To overcome these limitations, we propose Pixel-wise Attention Dissolution (PAD), a mask-guided modulation strategy that preserves the internal consistency of self-attention. Instead of globally scaling attention maps, PAD dissolves object information by disconnecting each masked query pixel from its most correlated keys. This pixel-level operation enables independent and precise multi-object removal in a single pass, while maintaining contextual harmony with the surrounding content and producing more coherent results.

C. Fine-Tuning and Optimization-Based Inpainting Approaches

SuppressEOT [43] suppressed undesired concepts via text embedding optimization but lacked spatial control, which limited its applicability to region-specific tasks such as object removal. Several works explored concept erasure through fine-tuning [44], [45], Inst-Inpaint [46] supported localized inpainting with text, but relied on paired training data and model retraining, reducing its practicality. MagicRemover [47] proposed a tuning-free inpainting method that optimizes text embeddings at inference time to remove unwanted objects given a textual prompt specifying the removal target. MAT [30] introduced a mask-aware transformer to handle large holes by learning spatial structure priors, but it required training from scratch and lacked semantic controllability.

These approaches highlighted the potential of fine-tuning and optimization for object removal but remained limited in flexibility and efficiency. In contrast, PANDORA enabled zero-shot, mask-aware, end-to-end multi-object removal in a single pass, without any model fine-tuning, prompt engineering, or inference-time optimization.

Algorithm 1: PANDORA's algorithm: Zero-Shot Object Removal

Inputs: A mask M, the intermediate latent noises x^i, the target initial latent noise map x_T.
Output: Erased Image I_t.
1) **For** $t = T, T-1, \ldots, 1$ **do:**
   a) $\{\_, K_i, V_i\} \xleftarrow{\text{get}} \epsilon_\theta(x_t^i, t)$
   b) $\{Q, \_, \_\} \xleftarrow{\text{get}} \epsilon_\theta(x_t, t)$
   c) self-attention $\xleftarrow{\text{inject}}$ BPA(Q, K_i, V_i, M)
   d) self-attention $\xleftarrow{\text{inject}}$ PAD(Q, K_i, V_i, M, top_k)
   e) $\epsilon \leftarrow \epsilon_\theta(x_t, t,$ self-attention)
   f) $\epsilon \leftarrow$ LADG$(\epsilon, \epsilon_\theta(x_t^i, t), M, \alpha_t)$
   g) $x_{t-1} \leftarrow$ Sample(x_t, ϵ)
2) **End For.**
Return: VAE(x_0).

III. Preliminaries of Diffusion Models

A. Latent Inversion via DDIM

Given a pre-trained diffusion model, an input image x_0 can be projected into its corresponding latent state x_T, which can then be used for guided sampling. This process, known as DDIM inversion [48], reconstructs the noise trajectory deterministically.

Assume a forward diffusion process:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

with β_t denoting the noise schedule. The marginal distribution becomes:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

In the DDIM sampling process, the reverse step is:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \sigma_t \cdot z, \tag{3}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$ and σ_t controls the stochasticity.

To invert an image, the process was run *backwards* from $t = 0$ to T, estimating noise at each step using the trained denoiser ϵ_θ. This yielded a latent code x_T from which the original image x_0 could be reconstructed deterministically.

B. Guidance Mechanisms

To control the image generation process from x_T, several guidance mechanisms were introduced to steer the denoising trajectory toward desired outcomes.

a) Classifier Guidance: Proposed by Dhariwal and Nichol [3], classifier guidance modified the reverse process using the gradient of a classifier $p(y \mid x_t)$:

$$\hat{\epsilon}_\theta(x_t; y) = \epsilon_\theta(x_t) - s \cdot \nabla_{x_t} \log p(y \mid x_t), \tag{4}$$

where s was the guidance scale. This method required a separately trained classifier.

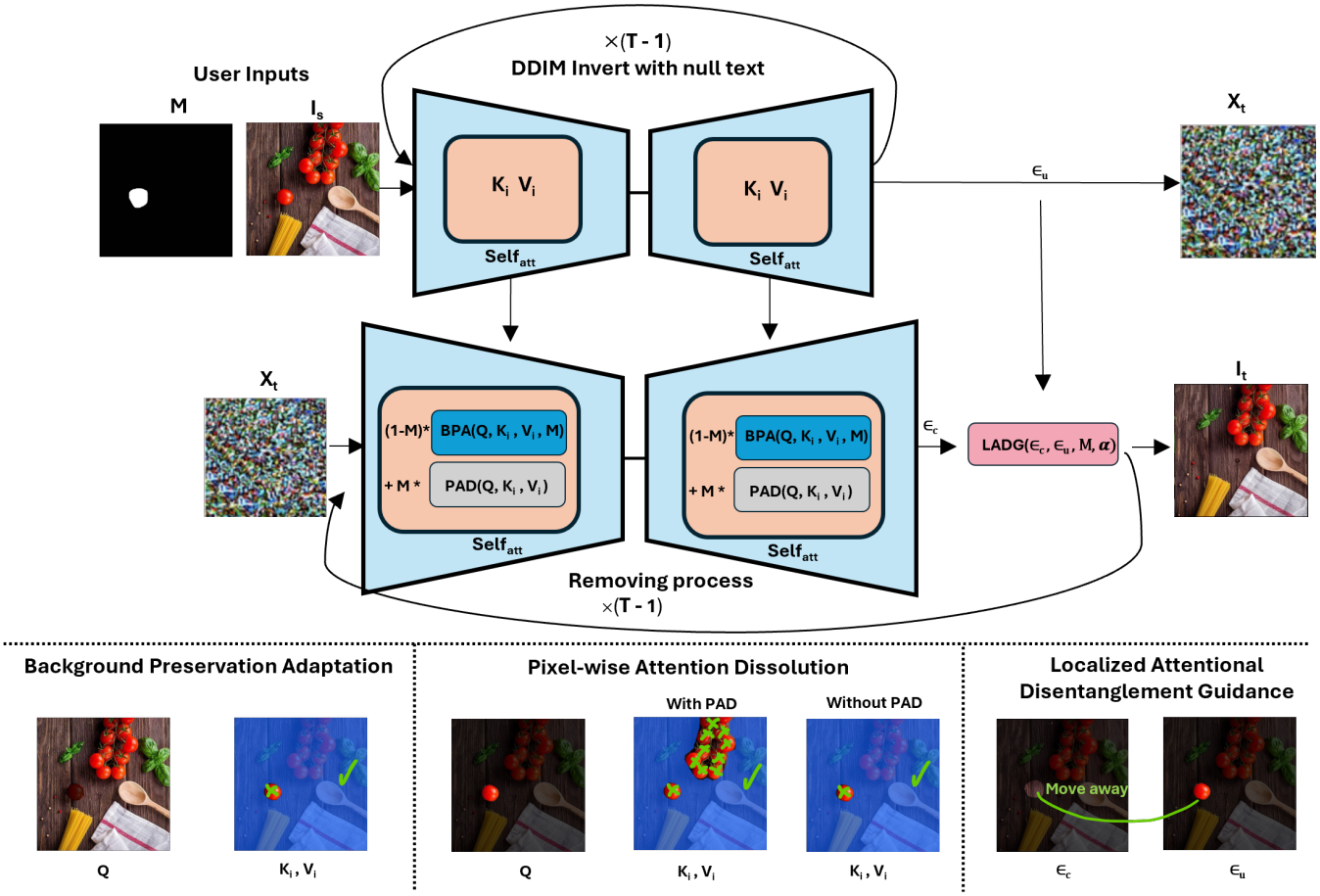

Fig. 2: Overview of our proposed pipeline with an intuitive illustration of each module. The image is inverted into noise with intermediate latents stored and injected into BPA and PAD to preserve background and dissolve objects, respectively. Specifically, BPA restricts background queries to background regions, while PAD operates at the pixel level to constrain object queries to unrelated regions. Finally, LADG steers denoising away from masked object regions for seamless synthesis.

b) Classifier-Free Guidance: Introduced by Ho and Salimans [49], this technique avoided external classifiers by interpolating between unconditional and conditional predictions:

$$\hat{\epsilon}_\theta(x_t; y) = (1 + s) \cdot \epsilon_\theta(x_t; y) - s \cdot \epsilon_\theta(x_t), \tag{5}$$

where $\epsilon_\theta(x_t; y)$ and $\epsilon_\theta(x_t)$ were predictions conditioned and unconditioned on prompt y, respectively.

Various guidance methods [50]–[54] steered the sampling process by subtracting the score of undesired distributions and/or adding the score of desired ones, thereby shifting the generation toward the targeted distribution. In the context of object removal, guidance could also be applied to mask-aware attention structures to eliminate target regions while preserving the rest of the scene.

C. Self-Attention

Within Stable Diffusion (SD) [11], the attention mechanism [55] of the denoising U-Net, which includes both self-attention and cross-attention, is mathematically expressed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V,$$

where Q represented the query features projected from spatial features, while K and V were the key and value features projected from spatial features (in self-attention layers) or textual embeddings (in cross-attention layers).

Prior studies [16], [37], [38] showed that incorporating self-attention features into U-Net layers supported semantic consistency during image translation. CPAM [32] further demonstrated that self-attention inherently allowed each region to establish flexible interactions, including self-referencing and selective connections with others. This property enabled smooth visual transitions and coherent global structures, even when suppressing specific regions.

IV. Proposed Method

A. Overview

PANDORA performs zero-shot object removal directly on a pre-trained diffusion model. Given an input image I_s and a binary mask M specifying the target objects, the model produces an edited image I_t where the masked regions are erased and seamlessly reconstructed with contextually consistent background. The process begins with latent inversion to

map the input image into the noise space while preserving unaffected regions in the denoising process. We then apply Pixel-wise Attention Dissolution (PAD) to disconnect masked query pixels from their most correlated keys, effectively dissolving object information at the attention level. Next, Localized Attentional Disentanglement Guidance (LADG) steers the denoising trajectory in latent space away from the object regions, refining the reconstruction to suppress residual artifacts. Together, PAD and LADG enable precise, pixel-level control for single- and multi-object removal in a single forward pass, without any fine-tuning, prompt engineering, or inference-time optimization. Figure 2 illustrates the overall pipeline and algorithm is outlined in Algorithm 1.

B. Background Preservation Adaptation (BPA)

To ensure that non-masked regions remain unaffected during the object removal process, we adopt the preservation adaptation module from CPAM [32] for background consistency. Specifically, we first invert the input image I_s into the latent noise space using DDIM inversion [48], which reconstructs the noise trajectory deterministically. During this process, the intermediate latent states x_i are stored at each timestep t to preserve semantic information.

At a denoising step t, let (Q, K, V) denote the query, key, and value features of the current noise in Unet's self-attention, to retain the background unchanged we get (K_i, V_i) the key and value features extracted from the stored latent noise at step t while retaining the query feature Q. The background semantic content SC_{bg} is obtained by applying mask-guided attention:

$$SC_{bg} = \text{Att}(Q, K_i, V_i; 1 - M), \tag{6}$$

where $\text{Att}(\cdot)$ denotes the attention mechanism and $(1 - M)$ ensures that only non-masked (background) areas contribute to the attention computation.

This operation effectively transfers semantic content from the original latent noise to the background region of the edited image, thereby preserving structural and visual fidelity in areas outside the object mask. As a result, the subsequent object removal modules can focus exclusively on masked regions without degrading the integrity of the surrounding scene.

C. Pixel-wise Attention Dissolution (PAD)

In contrast to background preservation, which reuses (K_i, V_i) to retain non-masked content, **Pixel-wise Attention Dissolution (PAD)** aims to erase object regions specified by the mask M by dissolving their strongest semantic connections in self-attention. Given the query features Q and stored (K_i, V_i), the attention logits are first computed as:

$$\mathbf{S}_{t,l} = \frac{QK_i^\top}{\sqrt{d}}, \tag{7}$$

where d is the feature dimension, and t, l denote the denoising step and self-attention layer in the U-Net. To suppress dominant associations, we apply percentile-based thresholding over each query's attention distribution and set the top-k strongest connections to $-\infty$:

$$\mathbf{S}^{\text{diss}}_{t,l}[i,j] = \begin{cases} -\infty & \text{if } j \in \text{Top-}k(\mathbf{A}_{t,l}[i,:]) \ \vee \ j \in M, \\ \mathbf{S}_{t,l}[i,j] & \text{otherwise,} \end{cases} \tag{8}$$

where $\mathbf{A}_{t,l} = \text{softmax}(\mathbf{S}_{t,l})$ is the normalized attention map. The dissolved self-attention output is then computed as:

$$\mathbf{B}_{t,l} = \text{softmax}(\mathbf{S}^{\text{diss}}_{t,l}) \cdot V_i. \tag{9}$$

Finally, to integrate PAD with background preservation, the final output representation is obtained by blending object and background content:

$$\mathbf{O}_{t,l} = \mathbf{B}_{t,l} \odot M + \mathbf{SC}^{\text{bg}}_{t,l} \odot (1 - M), \tag{10}$$

where $\mathbf{SC}^{\text{bg}}_{t,l}$ denotes the background-preserved semantic content.

Through this pixel-wise dissolution and selective recombination, each masked query abandons its strongest semantic ties, allowing the object region to vanish and be naturally reconstructed from surrounding background context, while ensuring global scene consistency.

D. Localized Attentional Disentanglement Guidance (LADG)

Diffusion models are score-based generators: their denoising trajectory is guided by noise predictions that approximate the gradient of the data distribution. Existing guidance strategies such as classifier-based [3] (requiring an auxiliary classifier) or classifier-free [49] (requiring training with conditional vs. unconditional embedding) operate globally. In contrast, our approach requires no auxiliary classifier or re-training with conditional and unconditional embedding; instead, we introduce **Localized Attentional Disentanglement Guidance (LADG)**, a spatially gated mechanism that *pushes the latent distribution away from the original (unconditioned) trajectory only inside the object mask*. Denote the unconditional and conditional noise predictions at step t by:

$$\epsilon_u(x_t^i, t) := \epsilon_\theta(x_t^i, t; \varnothing), \qquad \epsilon_c(x_t, t) := \epsilon_\theta(PAD(x_t), t; \varnothing)$$

Rather than applying a global guidance scale, LADG forms a *mask-aware* noise prediction:

$$\begin{aligned} \hat{\epsilon}(x_t, t) = (1 - M) \odot \epsilon_c(x_t, t) \\ + M \odot \Big[\alpha_t \, \epsilon_c(x_t, t) + (1 - \alpha_t) \, \epsilon_u(x_t^i, t)\Big], \end{aligned} \tag{11}$$

where α_t is a real number controlling the degree to which masked latents are steered away from the unconditional trajectory and toward the conditional prediction.

Intuitively, outside the mask we preserve the conditional trajectory to maintain scene fidelity, while inside the mask we explicitly drive the latents away from the original object-unconditioned score, thereby accelerating object removal and mitigating residual artifacts to produce cleaner and more coherent outputs.

Method	Text	FID↓	LPIPS↓	MSE↓	CLIP score↑
Fine-tuning-based methods					
PowerPaint [56]	✓	22.81	0.1322	0.0104	24.15
LaMa [21]	✗	**0.71**	**0.0012**	**0.0001**	24.50
SD2-Inpaint [11]	✗	17.93	0.1106	0.0073	24.06
SD2-Inpaint-wprompt [11]	✓	18.01	0.1098	0.0072	24.32
Zero-shot methods (no training required)					
CPAM [32]	✗	29.54	0.1564	0.0138	24.32
Attentive Eraser [22]	✗	118.09	0.2567	0.0270	24.42
PANDORA w/o LADG (Ours)	✗	42.17	0.1844	0.0171	24.55
PANDORA w/o PAD (Ours)	✗	35.59	0.1702	0.0156	24.4
PANDORA (Ours)	✗	44.98	0.1895	0.0184	**24.57**

TABLE I: Quantitative comparison between fine-tuning and zero-shot object removal methods, averaged across all dataset types. PANDORA achieves competitive background realism among zero-shot approaches, while outperforming even fine-tuned methods in object removal quality.

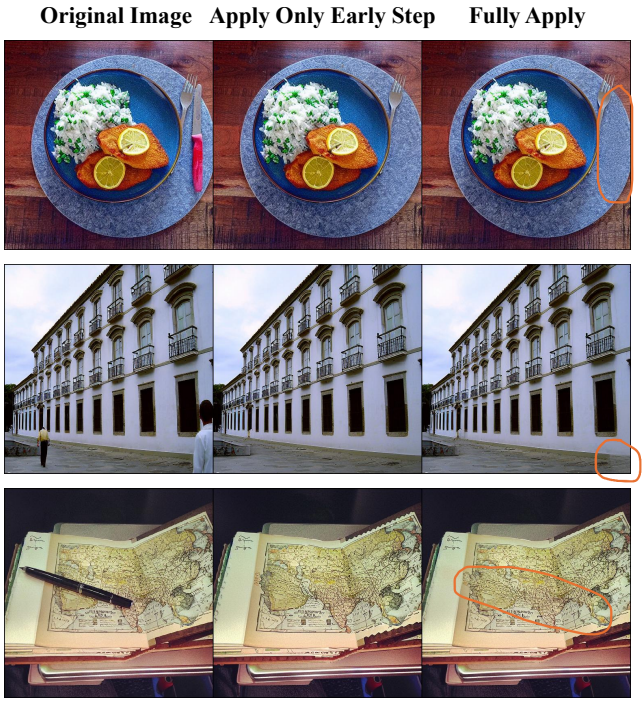

Fig. 3: Qualitative comparison of object removal with and without the adaptive application schedule. Regions of difference are highlighted with orange notations. Our method enforces object removal in the early denoising phase while restoring self-attention in later steps, yielding cleaner erasure and more harmonious results synthesis. Please zoom in for a clearer view.

V. Experiments

A. Ablation Study

a) Adaptive Application Schedule: While both Background Preservation Adaptation (BPA) and Pixel-wise Attention Dissolution (PAD) are effective in isolating background and object regions for retention and erasure, their strict separation can cause attention maps to become disconnected. In particular, the removed object regions may lose fine details or appear less harmonized with the surrounding content, as they only maintain connections to the original latent's background rather than dynamically adapting to the current denoising context.

To mitigate this issue, we apply BPA and PAD only during the early denoising phase, typically from step 1 to 40–45, suppose the maximum step is 50. Within this interval, object removal is enforced by constraining the self-attention mechanism. Once the undesired content has been sufficiently suppressed and the sampling trajectory aligns with the target distribution, the subsequent steps proceed without intervention. Inside the masked region, the conditional signal associated with the original object is progressively attenuated. After the mask is released, the self-attention module resumes its canonical form, allowing unrestricted pixel-to-pixel interactions, thereby restoring coherence and visual harmony across the image (as shown in Fig. 3).

b) Effect of LADG: To evaluate the contribution of the proposed Localized Attentional Disentanglement Guidance (LADG), we conduct experiments with and without its integration into the diffusion process. Without LADG, the model often fails to fully suppress the conditional signal associated with the target object, resulting in residual structures or ghost-like artifacts inside the masked region. This limitation is

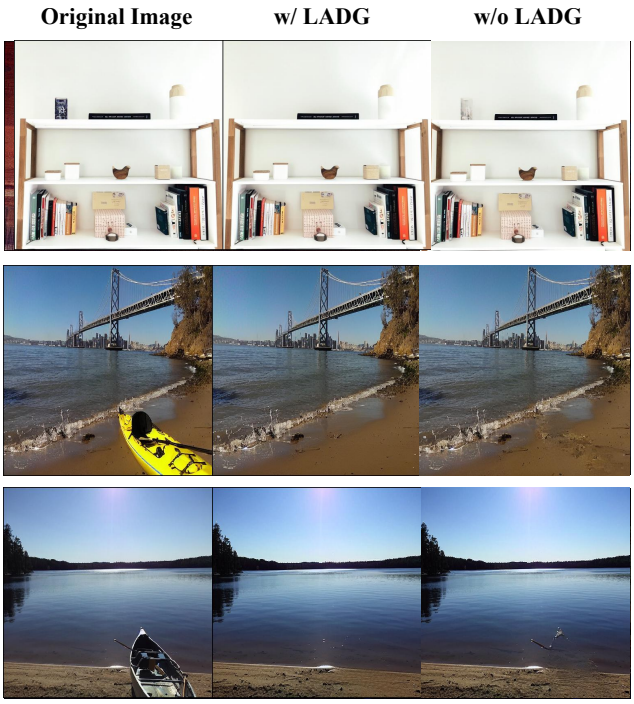

Fig. 4: Qualitative comparison with and without LADG. Without LADG, residual object traces and ghost-like artifacts remain, whereas LADG achieves cleaner removal and smoother background blending.

especially evident when the object strongly dominates local attention maps, leading to incomplete removal.

In contrast, incorporating LADG enforces spatially localized guidance away from original objects within the mask while preserving the global denoising trajectory outside. Qualitatively, this yields cleaner object removal and smoother results, where the inpainted regions seamlessly blend with their surroundings. Quantitatively, the improvement is reflected in lower FID and LPIPS scores, as well as higher background consistency metrics (as illustrated in Fig. 4). These results confirm that LADG plays a critical role in disentangling object-specific signals from the diffusion trajectory, enabling fine-grained removal and enhancing overall scene fidelity.

B. Implementation Details

PANDORA's framework is built upon Stable Diffusion v1.5 with official weights from Hugging Face diffusers. We employ DDIM inversion with 50 denoising steps. The proposed PAD and LADG modules are integrated into the U-Net's self-attention layers without retraining. Specifically, PAD suppresses the top 2-5% of correlated activations within masked regions to detach object dependencies, while LADG applies a decaying latent guidance weight (α_t from 1.0-1.6) to stabilize background refinement. To prevent disconnection artifacts, BPA and PAD are only activated during the early denoising phase (steps 1–40-45), after which full self-attention is restored for final refinement.

C. Baseline

We compare PANDORA against a diverse set of mask-guided state-of-the-art object removal approaches, covering both fine-tuned and zero-shot paradigms. For fine-tuning-based methods, we include LaMa [21], PowerPaint [56], and SD2-Inpaint [11] (evaluated with and without text prompts). Specifically, we use PowerPaint v2-1 with the prompt "empty scene blur" and SD2-Inpaint with the prompt "a clean, natural background, seamless and realistic lighting." For zero-shot methods, we evaluate CPAM [32], Attentive Eraser [22], and our proposed PANDORA, all built upon the Stable Diffusion v1.5 checkpoint with 50 denoising steps. All methods are tested under identical masking conditions for fair comparison, using their official implementations and default configurations.

D. Benchmark Dataset

To evaluate our method, we construct a benchmark dataset for multi-object removal by collecting images from multiple sources. Specifically, single- and multi-object samples are taken from PIE-Bench [57] and the Open Images Dataset [58], both of which provide high-quality paired images and corresponding object masks. For the mass-similar object removal scenario, we include samples from Ranjan et al. [59], which feature clusters of visually similar objects such as fruits, flowers, and birds. We remove low-quality or overly simple samples and retain diverse scenes with varied object shapes, textures, and contextual relationships. In total, the dataset comprises 75 single-object samples, 17 multi-object samples, and 94 mass-similar object samples, with masks obtained through manual annotation or automatic extraction.

E. Evaluation Metrics

To comprehensively evaluate object removal performance, we assess both background fidelity and object removal quality. For background regions, we employ Mean Squared Error (MSE) and LPIPS [60] to quantify the pixel-level and perceptual similarity between the generated background and the ground truth background. To measure how effectively the target object is removed, we adopt the CLIP score [61], which evaluates the semantic alignment between the edited region and a background-related text prompt ("A background without any objects."). In addition, we compute the Fréchet Inception Distance (FID) [62] to assess the overall realism and distributional consistency of the generated images compared to the ground truth background regions (excluding masked areas). To ensure a fair comparison across all methods, all images are resized to 512×512 resolution, and the masked regions are uniformly processed for each dataset.

F. Qualitative and Quantitative Results

Table. I presents the averaged results across all dataset types. Among zero-shot methods, PANDORA achieves the best balance between background realism and object removal quality. It attains the highest CLIP score, showing strong semantic consistency and effective object elimination, while maintaining competitive LPIPS and MSE scores.CPAM [32]

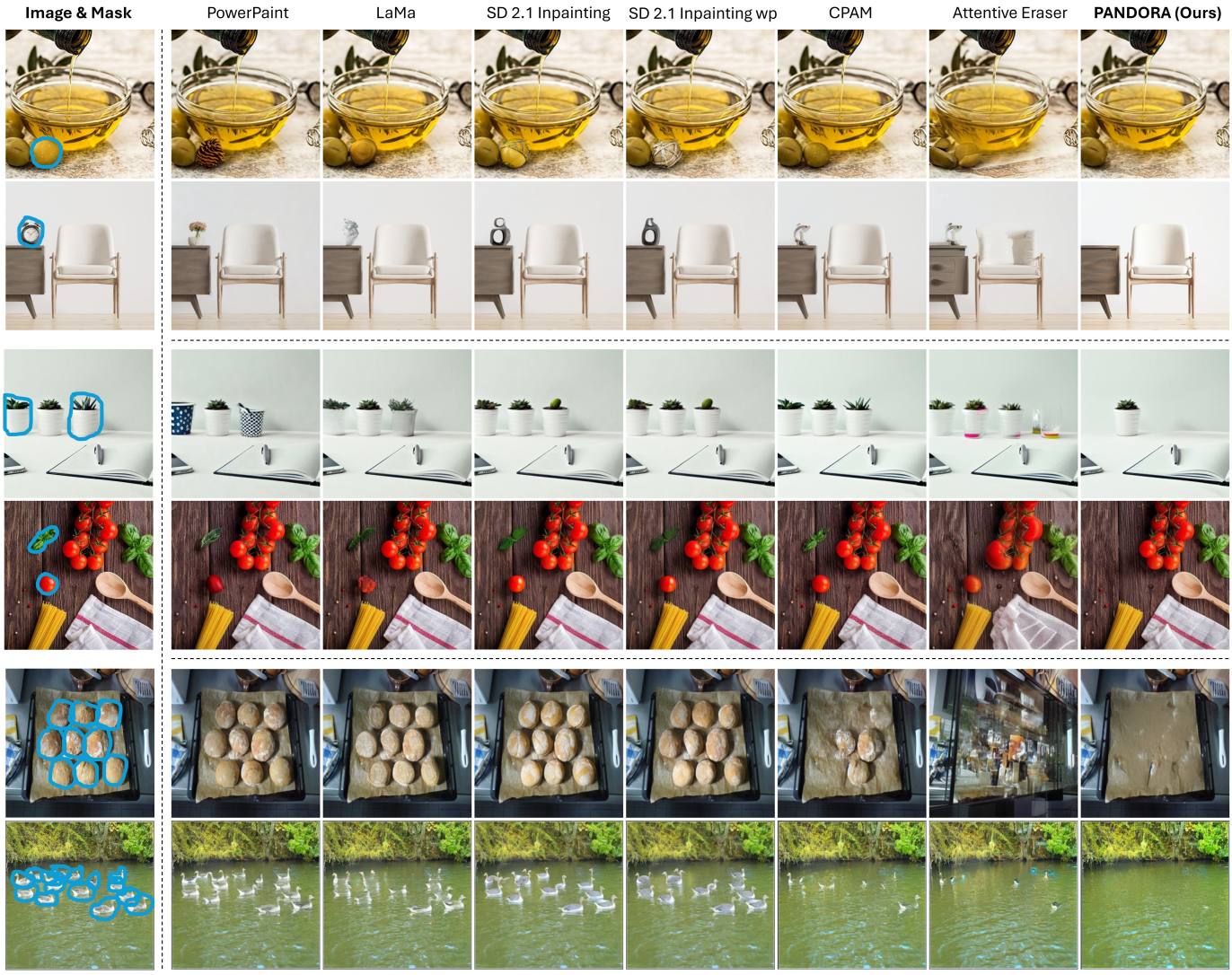

Fig. 5: Qualitative comparison on various object removal scenarios. From left to right: original image with a mask, and results from different methods. The top two rows show single-object removal, the middle two rows show multi-object cases, and the bottom two rows show mass-similar object removal. The last three columns show zero-shot methods.

produces balanced reconstructions but struggles to remove objects effectively, whereas Attentive Eraser [22] often distorts scene structure, resulting in high error metrics. For fine-tuned models, LaMa [21] records the lowest FID, LPIPS, and MSE due to its rigid blending strategy, which merges original background patches but limits realistic restoration. PowerPaint [56] and SD2-Inpaint [11] perform moderately yet depend on training or prompts. Overall, PANDORA achieves the most coherent background and effective removal, confirming the strength of its attention-guided erasure mechanism.

Figure. 5 presents qualitative comparisons across various scenarios. PANDORA effectively removes both distinct and clustered objects while preserving fine background details and maintaining global coherence. In contrast, fine-tuning-based methods often modify or replace target objects instead of removing them. Among zero-shot baselines, CPAM [32] preserves background structure relatively well, whereas Attentive Eraser [22] tends to lose fine details or produce overly altered textures; however, both struggle to remove objects when similar instances appear elsewhere in the scene. All methods except PANDORA frequently introduce residual artifacts or distortions, while PANDORA produces clean, contextually consistent reconstructions with minimal artifacts.

G. User Study

To evaluate the effectiveness of PANDORA, we conducted a user study with 20 participants from diverse backgrounds. Each participant was asked to select the best image from sets of outputs, where the original image and the results of six different methods were presented side-by-side. To ensure objectivity, the methods were shuffled and blinded so participants did not know which image corresponded to which method, including PANDORA. The evaluation was organized into 20

Method	Chosen (%)
Fine-tuning-based methods	
PowerPaint [56]	4.5
LaMa [21]	6.75
SD2-Inpaint [11]	6.25
SD2-Inpaint-wprompt [11]	7
Zero-shot methods (no training required)	
CPAM [32]	6
Attentive Eraser [22]	22
PANDORA (Ours)	47.5

TABLE II: User study results showing the percentage of times each method was chosen as the best. PANDORA is consistently favored by users.

batches, each containing 20 randomly selected samples, yielding a total of 400 responses and 2,800 image considerations across all methods. Table II summarizes the results, showing that PANDORA was consistently favored by users. Overall, the user study reinforces our qualitative and quantitative findings, highlighting PANDORA's effectiveness in the object removal task.

VI. Limitations and Future Work

Although effective, our approach has several limitations. First, suppression based on a fixed percentile threshold may occasionally over-filter or under-filter attention responses, leading to incomplete or excessive object removal. Second, the framework depends on accurate binary masks; imprecise segmentation can reduce disentanglement quality and introduce artifacts. In future work, we plan to explore adaptive mechanisms that adjust percentile thresholds according to attention statistics or mask confidence, as well as automatic region selection to identify removable objects without manual input.

VII. Conclusion

In this paper, we introduced PANDORA, a novel zero-shot framework for object removal that operates directly on pre-trained diffusion models without requiring any fine-tuning, prompts, or inference-time optimization. Our approach addresses the key challenges of maintaining background integrity while achieving clean, semantically coherent object erasure. The core of our method lies in two synergistic components: Pixel-wise Attention Dissolution (PAD), which precisely dissolves object information at a granular level by nullifying the most correlated keys in self-attention, and Localized Attentional Disentanglement Guidance (LADG), which steers the denoising process away from object-related latent manifolds. Together, these modules enable flexible and effective removal of single, multiple, and even densely packed objects in a single forward pass. Extensive experiments demonstrate that PANDORA significantly outperforms existing state-of-the-art methods, including those that rely on fine-tuning and prompt guidance, setting a new benchmark for zero-shot object removal in terms of both visual quality and semantic plausibility.

References

[1] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[3] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[4] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *International conference on machine learning*, Pmlr, 2021, pp. 8821–8831.

[5] A. Nichol *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *International conference on machine learning*, 2022.

[6] J. Yu *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

[7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents. arxiv 2022," *arXiv preprint arXiv:2204.06125*, 2022.

[8] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 36 479–36 494. [Online]. Available: https : / / proceedings . neurips . cc / paper_files / paper / 2022 / file / ec795aeadae0b7d230fa35cbaf04c041 - Paper-Conference.pdf.

[9] Black Forest Labs, *Flux*, https : / / github . com / black - forest-labs/flux, Accessed: 2024, 2024.

[10] P. Esser *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis, 2024," *URL https://arxiv. org/abs/2403.03206*, vol. 2,

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[12] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*, PMLR, 2021, pp. 8162–8171.

[13] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435.

[14] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. 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*, 2023, pp. 6038–6047.

[15] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *International Conference on Learning Representations*, 2023.

[16] N. Tumanyan, M. Geyer, S. Bagon, and T. 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*, 2023, pp. 1921–1930.

[17] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[18] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

[19] R. Gal *et al.*, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *International Conference on Learning Representations*, 2023.

[20] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[21] R. Suvorov *et al.*, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2149–2159.

[22] W. Sun, X.-M. Dong, B. Cui, and J. Tang, "Attentive eraser: Unleashing diffusion model's object removal potential via self-attention redirection guidance," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 20 734–20 742.

[23] Y. Ekin, A. B. Yildirim, E. E. Çağlar, A. Erdem, E. Erdem, and A. Dundar, "Clipaway: Harmonizing focused embeddings for removing objects via diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 17 572–17 601, 2024.

[24] Zomet, "Learning how to inpaint from global image statistics," in *Proceedings Ninth IEEE international conference on computer vision*, IEEE, 2003, pp. 305–312.

[25] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Transactions on graphics (TOG)*, vol. 26, no. 3, 4–es, 2007.

[26] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[27] S. Zhao *et al.*, "Large scale image completion via co-modulated generative adversarial networks," *arXiv preprint arXiv:2103.10428*, 2021.

[28] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7508–7517.

[29] Q. Dong, C. Cao, and Y. Fu, "Incremental transformer structure enhanced image inpainting with masking positional encoding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 358–11 368.

[30] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 758–10 768.

[31] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023.

[32] D.-K. Vo, T.-T. Do, T. V. Nguyen, M.-T. Tran, and T.-N. Le, "Cpam: Context-preserving adaptive manipulation for zero-shot real image editing," *arXiv preprint arXiv:2506.18438*, 2025.

[33] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 208–18 218.

[34] C. Meng *et al.*, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations*, 2022.

[35] M. Brack *et al.*, "Ledits++: Limitless image editing using text-to-image models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8861–8870.

[36] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–10, 2023.

[37] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. 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)*, Oct. 2023, pp. 22 560–22 570.

[38] B. Liu, C. Wang, T. Cao, K. Jia, and J. Huang, "Towards understanding cross and self-attention in stable diffusion for text-guided image editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7817–7826.

[39] V. Titov, M. Khalmatova, A. Ivanova, D. Vetrov, and A. Alanov, "Guide-and-rescale: Self-guidance mechanism for effective tuning-free real image editing," in

European Conference on Computer Vision, Springer, 2024, pp. 235–251.

[40] O. Avrahami, O. Fried, and D. Lischinski, "Blended latent diffusion," *ACM transactions on graphics (TOG)*, vol. 42, no. 4, pp. 1–11, 2023.

[41] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," *International Conference in Learning Representations*, 2023.

[42] S. Li *et al.*, "Zone: Zero-shot instruction-guided local editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6254–6263.

[43] S. Li, J. van de Weijer, F. Khan, Q. Hou, Y. Wang, *et al.*, "Get what you want, not what you don't: Image content suppression for text-to-image diffusion models," in *The Twelfth International Conference on Learning Representations*, 2023.

[44] R. Gandikota, J. Materzyńska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023.

[45] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu, "Ablating concepts in text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 22 691–22 702.

[46] A. B. Yildirim, V. Baday, E. Erdem, A. Erdem, and A. Dundar, *Inst-inpaint: Instructing to remove objects with diffusion models*, 2023. arXiv: 2304.03246 [cs.CV].

[47] S. Yang, L. Zhang, L. Ma, Y. Liu, J. Fu, and Y. He, "Magicremover: Tuning-free text-guided image inpainting with diffusion models," *arXiv preprint arXiv:2310.02848*, 2023.

[48] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *International Conference on Learning Representations*, 2021.

[49] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. [Online]. Available: https://openreview.net/forum?id=qw8AKxfYbI.

[50] S. Hong, G. Lee, W. Jang, and S. Kim, "Improving sample quality of diffusion models using self-attention guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7462–7471.

[51] D. Ahn *et al.*, "Self-rectifying diffusion sampling with perturbed-attention guidance," in *European Conference on Computer Vision*, Springer, 2024, pp. 1–17.

[52] A. Bansal *et al.*, "Universal guidance for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 843–852.

[53] D. Epstein, A. Jabri, B. Poole, A. Efros, and A. Holynski, "Diffusion self-guidance for controllable image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 16 222–16 239, 2023.

[54] S. Mo *et al.*, "Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7465–7475.

[55] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[56] J. Zhuang, Y. Zeng, W. Liu, C. Yuan, and K. Chen, "A task is worth one word: Learning with task prompts for high-quality versatile image inpainting," in *European Conference on Computer Vision*, Springer, 2024, pp. 195–211.

[57] X. Ju, A. Zeng, Y. Bian, S. Liu, and Q. Xu, "Direct inversion: Boosting diffusion-based editing with 3 lines of code," *arXiv preprint arXiv:2304.04269*, 2023.

[58] A. Kuznetsova *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International journal of computer vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[59] V. Ranjan, U. Sharma, T. Nguyen, and M. Hoai, "Learning to count everything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3394–3403.

[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[61] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[62] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.