DECOUPLING SOURCE PRESERVATION FROM TARGET GENERATION IN NEXT-SCALE AUTOREGRESSIVE IMAGE EDITING VIA FIXED-HISTORY CONTRAST

Anonymous authors
Paper under double-blind review

ABSTRACT

Next-scale autoregressive models have emerged as a new paradigm for visual generation, inspiring growing efforts to extend them to training-free image editing. Effective editing requires preserving source content unrelated to the intended change and generating new content aligned with the target prompt. Existing methods often derive preservation signals from source–target prediction contrasts during autoregressive generation. Because these contrasts are evaluated along an evolving edited history, preservation decisions at later scales may depend on earlier changes, potentially causing unintended changes or insufficient edits. To overcome this limitation, we propose DecoEdit, a training-free editing framework that decouples the determination of source preservation from the evolving target generation process. DecoEdit achieves this through fixed-history contrast, which computes source- and target-conditioned prediction contrasts under a fixed source history obtained from the encoded source image. For source preservation, DecoEdit converts these contrasts into preservation evidence to identify source content to retain, and constructs a preservation bias that controls the strength of source retention. For target generation, DecoEdit uses this evidence to modulate the preservation bias and injects the modulated bias into the evolving target-conditioned process to produce the requested edit. Experiments on Infinity-2B and Infinity-8B demonstrate that DecoEdit achieves a better preservation-editability trade-off than existing methods. The anonymous code is attached in the supplementary material.

1 INTRODUCTION

Next-scale autoregressive (AR) generation (Tian et al., 2024; Han et al., 2025) has recently emerged as a new paradigm for visual generation, synthesizing images from scratch through next-scale prediction. Beyond image generation, practical visual content creation also requires the ability to modify existing images according to user intent. Following the successful extension of diffusion models (Ho et al., 2020; Song et al., 2020) from generation to editing (Hertz et al., 2022; Parmar et al., 2023; Cao et al., 2023; Tumanyan et al., 2023), making a similar extension to next-scale AR models is also a natural and important research direction.

Unlike instruction-guided editing (Brooks et al., 2023; Zhang et al., 2024), we focus on training-free editing with a frozen text-to-image generative model, where the desired modification to the source image is specified by the semantic difference between a source prompt and a target prompt. Since the generative model is frozen and lacks task-specific training, it has no explicit knowledge of how to transform the source image into the target image based on this semantic difference. From this perspective, editing involves two key components: source preservation and target generation, which respectively retain relevant source content and produce the desired new content. Target generation can leverage the generative capability of the pretrained model under the target condition, whereas source preservation requires an explicit mechanism.

Existing methods commonly use source-target prediction contrasts to determine source preservation during autoregressive editing. For example, as shown in Figure 1(a), AREdit (Wang et al., 2025) contrasts cached source predictions with target-conditioned predictions based on previously edited scales, while EditMod (Fang et al., 2026) contrasts source- and target-conditioned predictions using

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

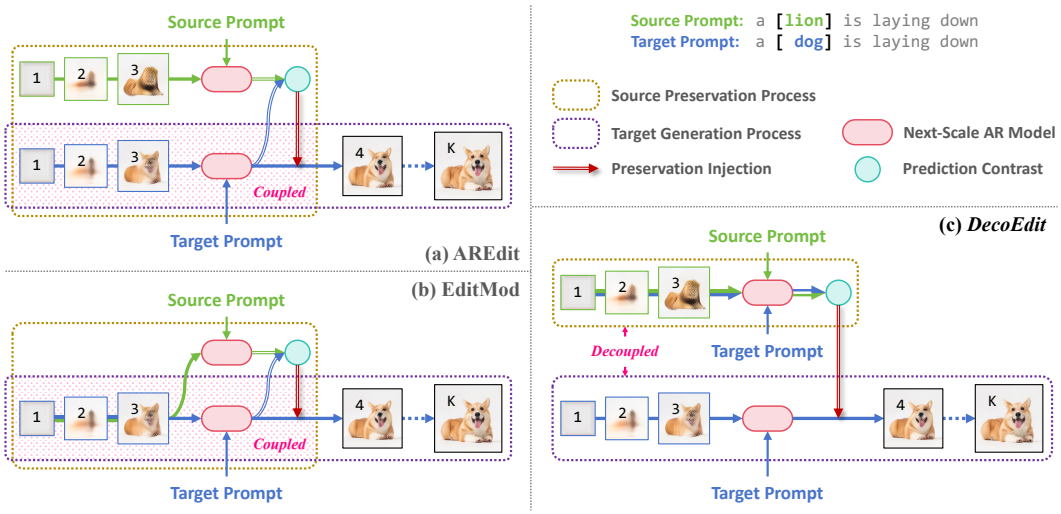


Figure 1: Comparison of source preservation and target generation in next-scale AR image editing. (a) AREdit and (b) EditMod determine source preservation using predictions affected by the evolving edited history, coupling preservation with generation. (c) Our DecoEdit decouples preservation with generation by deriving preservation guidance under a fixed source history.

the same previously edited context, as shown in Figure 1(b). While effective, these methods can still make inaccurate preservation decisions, leading to unintended changes in irrelevant regions or insufficient modification of the intended content. We argue that a key reason is that the determination of source preservation at each scale is conditioned on an evolving target generative history. Specifically, since each scale prediction in next-scale AR models explicitly depends on preceding scales, modifications introduced at earlier scales can propagate into subsequent source- and target-conditioned predictions. Consequently, the contrast may reflect not only the semantic difference between the source and target prompts but also the effects of preceding generation, making the resulting preservation signal less reliable for identifying source content to retain. Overall, existing methods tightly couple source preservation to target generation at each scale. However, the source image, source prompt, and target prompt provides a stable basis for inferring the source preservation requirements. This suggests that the determination of source preservation can be decoupled from the evolving target generation process.

To this end, we propose DecoEdit, a training-free editing framework that decouples the determination of source preservation from the evolving target generation process through fixed-history contrast. Specifically, as shown in Figure 1(c), DecoEdit computes source- and target-conditioned prediction contrasts under a fixed source history obtained from the encoded source image, ensuring that their prediction contrast reflects the requested semantic change without being affected by preceding edits. For source preservation, DecoEdit converts the fixed-history contrasts into preservation evidence that distinguishes source content compatible with the target prompt from content involved in the requested modification. It further constructs a preservation bias from the source-conditioned prediction to compensate for source details at risk of being lost and to control the strength of source retention. For target generation, DecoEdit retains the original autoregressive process under the target prompt and the evolving edited history. At each scale, the preservation evidence modulates the preservation bias before injection, providing stronger source compensation for content intended for retention and weaker compensation for content requiring generation. Experiments on Infinity show that DecoEdit achieves a better preservation-editability trade-off than existing methods.

Our main contributions are summarized as follows:

- We identify the coupling of source-preservation determination with evolving target generation as a key limitation of existing training-free next-scale AR editing methods.
- We propose DecoEdit, which derives preservation evidence via fixed-history contrast and injects an evidence-modulated bias into target generation.
- Experiments demonstrate that DecoEdit achieves a better preservation-editability trade-off.

2 Related Works

Text-Guided Image Generation. Text-guided image generation aims to synthesize visually realistic images that are semantically aligned with textual prompts. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) have become the dominant paradigm, formulating image generation as a conditioned iterative denoising process (Rombach et al., 2022; Esser et al., 2024; Labs et al., 2025). In parallel, autoregressive (AR) image generation (Sun et al., 2024; Li et al., 2024; Han et al., 2025; Yu et al., 2025) has attracted increasing attention by extending the next-token prediction paradigm of language models to visual generation. Among them, next-scale AR models (Tian et al., 2024; Han et al., 2025) reformulate next-token prediction as coarse-to-fine next-scale prediction, better accommodating the spatial structure of images. In this paper, we aim to extend the next-scale AR paradigm from image generation to training-free image editing.

Training-Free Image Editing. Image editing aims to modify image content according to user intent while preserving unrelated source information. Unlike instruction-guided editing (Brooks et al., 2023; Zhang et al., 2024), training-free editing operates with a frozen text-to-image model and represents the desired modification through the semantic difference between a source prompt and a target prompt. A large body of work has explored training-free image editing under the well-established diffusion paradigm (Hertz et al., 2022; Parmar et al., 2023; Cao et al., 2023; Tumanyan et al., 2023; Rout et al., 2024; Avrahami et al., 2025; Wang et al., 2024; Kulikov et al., 2025). More recently, training-free editing has been extended to next-scale AR models. AREdit (Wang et al., 2025) caches source representations and prediction distributions, and contrasts the cached source predictions with target-conditioned predictions based on previously edited scales to guide source preservation and target generation. EditMod (Fang et al., 2026) preserves source content with encoded source representations and generates target changes by applying source-target prediction differences under a shared edited context. In this paper, we focus on training-free image editing within the next-scale AR paradigm, aiming to decouple source preservation from target generation.

3 Preliminary

Next-scale autoregressive (AR) image generation (Tian et al., 2024), also known as Visual Autoregressive Modeling (VAR), is an emerging paradigm for image generation. In this paradigm, images are generated through coarse-to-fine discrete token prediction, where each scale causally conditions all subsequent scales. Specifically, the generation process starts from an initial 1×1 token map $\mathbf{R}_1$ and autoregressively predicts subsequent token maps with larger scales $(\mathbf{R}_2, \cdots, \mathbf{R}_K)$.

As an official extension, Infinity (Han et al., 2025) brings this paradigm to text-guided image generation. It uses Flan-T5 (Chung et al., 2024) to encode the input prompt and inject the resulting textual embeddings into the model. Similar to Latent Diffusion Models (LDMs) (Rombach et al., 2022), Infinity performs generation in a latent space. To represent these latents as discrete tokens and improve generative fidelity, Infinity adopts Binary Spherical Quantization (BSQ) (Zhao et al., 2025) scheme. Under BSQ, each spatial token is represented by D independent bit channels, yielding a token map $\mathbf{R}_k \in \{-1, +1\}^{D \times H_k \times W_k}$ at scale k, where H_k and W_k denote its height and width. For prediction, each bit channel is modeled as a two-way categorical distribution $\mathbf{P}_k \in [0, 1]^{2D \times H_k \times W_k}$.

4 Methodology

4.1 Overview

In this paper, we consider the setting of training-free image editing. Formally, given a source image $\mathcal{I}^{\text{src}}$, a source prompt $\mathcal{C}^{\text{src}}$, and a target prompt $\mathcal{C}^{\text{tgt}}$, the goal is to generate an edited image $\mathcal{I}^{\star}$ that realizes the semantic change from $\mathcal{C}^{\text{src}}$ to $\mathcal{C}^{\text{tgt}}$ while preserving source content unrelated to this change. DecoEdit performs the editing task with a frozen next-scale AR model and requires no additional training or parameter updates.

The key design of DecoEdit is to decouple source preservation from target generation through fixed-history contrast. As shown in Figure 2, at scale k, source preservation is determined from the fixed source history $(\mathbf{R}_1^{\text{src}}, \cdots, \mathbf{R}_{k-1}^{\text{src}})$, whereas target generation is conditioned on the evolving edited history $(\mathbf{R}_1^{\text{edit}}, \cdots, \mathbf{R}_{k-1}^{\text{edit}})$. The former is obtained entirely from the encoded source image and is

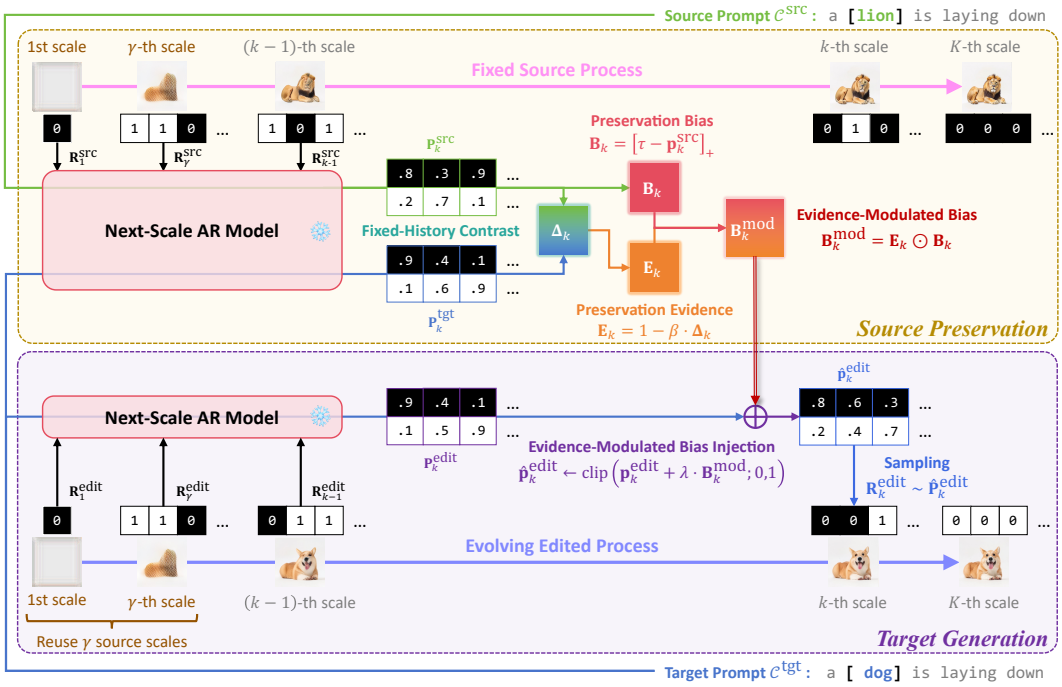


Figure 2: Overview of the proposed DecoEdit framework, which decouples source preservation from target generation for training-free image editing with pretrained next-scale AR models.

therefore unaffected by the generated edit, while the latter is progressively updated with the edited token maps produced at preceding scales.

To compute the fixed-history contrast (Section 4.2), DecoEdit evaluates the source- and target-conditioned predictions under the same fixed source history $(\mathbf{R}_1^{\text{src}}, \cdots, \mathbf{R}_{k-1}^{\text{src}})$, which is obtained by encoding source image. Because their visual histories are identical, the resulting contrast isolates the effect of switching from the source prompt to the target prompt.

For source preservation (Section 4.3), DecoEdit converts the fixed-history contrast into preservation evidence $\mathbf{E}_k$, which identifies the source content that should be retained. To determine the strength of source retention, a preservation bias $\mathbf{B}_k$ is constructed from the source-conditioned prediction.

For target generation (Section 4.4), DecoEdit computes target-conditioned predictions from the evolving edited history $(\mathbf{R}_1^{\text{edit}}, \cdots, \mathbf{R}_{k-1}^{\text{edit}})$. It uses the preservation evidence to modulate the preservation bias and injects the resulting bias $\mathbf{E}_k \odot \mathbf{B}_k$ into target-conditioned prediction. The coarsest source token maps are reused as a structural anchor, while the remaining scales are sampled autoregressively from the bias-adjusted predictions.

4.2 Fixed History Contrast

The goal of fixed-history contrast is to extract the semantic difference between the source and target prompts with respect to the source image, independently of the evolving target generation process. Let $(\mathbf{R}_1^{\text{src}}, \cdots, \mathbf{R}_K^{\text{src}})$ denote the multi-scale token maps obtained by encoding and quantizing the source image $\mathcal{I}^{\text{src}}$. At scale k, we evaluate the frozen model under the source and target prompts using the same fixed source history $(\mathbf{R}_1^{\text{src}}, \cdots, \mathbf{R}_{k-1}^{\text{src}})$, which can be formulated as

$$\mathbf{P}_k^{\text{src}} = p_\theta\left(\cdot \mid \mathbf{R}_1^{\text{src}}, \cdots, \mathbf{R}_{k-1}^{\text{src}}; \mathcal{C}^{\text{src}}\right) , \tag{1}$$

$$\mathbf{P}_k^{\text{tgt}} = p_\theta\left(\cdot \mid \mathbf{R}_1^{\text{src}}, \cdots, \mathbf{R}_{k-1}^{\text{src}}; \mathcal{C}^{\text{tgt}}\right) . \tag{2}$$

Here, p_θ denotes the frozen next-scale AR model with parameters θ. Since the two predictions share exactly the same visual history, their difference can be attributed solely to switching from the source text condition to the target text condition.

To extract the prompt-induced semantic difference with respect to the source image, we first use the current source token map $\mathbf{R}_k^{\text{src}}$ to retrieve the probabilities assigned by the two predictions to the same source bits, which are

$$\mathbf{p}_k^{\text{src}} := \mathbf{P}_k^{\text{src}}\left[\mathbf{R}_k^{\text{src}}\right] \in [0,1]^{D\times H_k\times W_k}\ ,\ \mathbf{p}_k^{\text{tgt}} := \mathbf{P}_k^{\text{tgt}}\left[\mathbf{R}_k^{\text{src}}\right] \in [0,1]^{D\times H_k\times W_k}\ . \tag{3}$$

Each entry of $\mathbf{p}_k^{\text{src}}$ and $\mathbf{p}_k^{\text{tgt}}$ represents the probability assigned to the same source bit under the source and target conditions, respectively. We then define the fixed-history contrast as

$$\mathbf{\Delta}_k = \left[\mathbf{p}_k^{\text{src}} - \mathbf{p}_k^{\text{tgt}}\right]_+\ , \tag{4}$$

where $[\cdot]_+$ denotes the element-wise positive part. A positive value indicates that the target prompt assigns a lower probability to the corresponding source bit than the source prompt, suggesting that the source information represented by this bit may need to be modified. We retain only such probability decreases: when the target-conditioned probability is no lower than the source-conditioned probability, the comparison provides no evidence that retaining the source bit conflicts with the requested edit. Because both probability maps are computed under the same fixed source history and evaluated at the same source bits, $\mathbf{\Delta}_k$ isolates the effect of changing the text condition and remains independent of the evolving target generation process.

4.3 Source Preservation

Source preservation requires reliable identification of source information for retention and appropriate control of its retention strength. DecoEdit addresses these requirements with two complementary quantities. A preservation evidence $\mathbf{E}_k$ is derived from $\mathbf{\Delta}_k$ to indicate the necessity of retaining each source bit. A preservation bias $\mathbf{B}_k$ is further constructed from the source-conditioned prediction to specify the corresponding retention strength. Both quantities are derived under the fixed source history and therefore remain independent of the evolving target generation process.

Preservation Evidence. The fixed-history contrast $\mathbf{\Delta}_k$ quantifies the reduction in probability assigned to each source bit after switching from the source prompt to the target prompt. A larger reduction indicates a lower need to retain the corresponding source bit. We therefore transform this modification-oriented contrast into preservation evidence

$$\mathbf{E}_k = 1 - \beta\cdot\mathbf{\Delta}_k \in [0,1]^{D\times H_k\times W_k}\ , \tag{5}$$

where $\beta\in[0,1]$ controls the effect of the fixed-history contrast on the preservation evidence. When the target prompt causes little or no decrease in the probability of a source bit, the corresponding preservation evidence remains high, indicating that the represented source information should be retained. Conversely, a larger probability decrease produces weaker preservation evidence, indicating a reduced need to preserve the corresponding source bit during editing.

Preservation Bias. While preservation evidence identifies the source information to retain, it does not specify the required retention strength. Even under the source prompt and fixed source history, $\mathbf{P}_k^{\text{src}}$ may assign low probabilities to some source bits, so samples drawn from this distribution may fail to reproduce the current source token map $\mathbf{R}_k^{\text{src}}$. A lower source-conditioned probability increases the risk of losing the corresponding source bit during sampling and therefore requires greater probability compensation. Accordingly, we construct a preservation bias from the probability shortfall relative to a confidence threshold τ,

$$\mathbf{B}_k = \left[\tau - \mathbf{p}_k^{\text{src}}\right]_+ \in [0,\tau]^{D\times H_k\times W_k}\ . \tag{6}$$

If the source-conditioned probability is at least τ, no probability compensation is required and the corresponding preservation bias is zero. Otherwise, $\mathbf{B}_k$ records the probability shortfall $\tau - \mathbf{p}_k^{\text{src}}$ as the base compensation for the corresponding source bit.

4.4 Target Generation

Target generation aims to produce the requested edit under the target prompt while incorporating the source preservation information determined from the fixed source history. DecoEdit performs autoregressive target generation based on the evolving edited history. At scale k, the editing prediction is computed as

$$\mathbf{P}_k^{\text{edit}} = p_\theta\left(\cdot \mid \mathbf{R}_1^{\text{edit}},\cdots,\mathbf{R}_{k-1}^{\text{edit}};\mathcal{C}^{\text{tgt}}\right)\ . \tag{7}$$



Figure 3: Illustration of coarse-to-fine generation in a next-scale autoregressive model.

Unlike $\mathbf{P}_k^{\text{tgt}}$ used for fixed-history contrast, $\mathbf{P}_k^{\text{edit}}$ depends on the edited token maps generated at preceding scales and is used for actual target generation. To incorporate source preservation into this process, DecoEdit first modulates and injects the preservation bias into the editing prediction, and then samples the current token map autoregressively from the adjusted distribution.

Evidence-Modulated Bias Injection. The preservation evidence $\mathbf{E}_k$ and preservation bias $\mathbf{B}_k$ encode complementary source preservation information. $\mathbf{E}_k$ assigns larger values to source bits requiring retention, while $\mathbf{B}_k$ provides the base probability compensation for under-predicted source bits. Combining them yields an adaptive preservation bias that provides stronger probability compensation for content to preserve and weaker compensation for content to generate,

$$\mathbf{B}_k^{\text{mod}} = \mathbf{E}_k \odot \mathbf{B}_k \ . \tag{8}$$

$\mathbf{B}_k^{\text{mod}}$ is then injected into the probabilities assigned to the source bits,

$$\hat{\mathbf{p}}_k^{\text{edit}} \leftarrow \text{clip}\left(\mathbf{p}_k^{\text{edit}} + \lambda \cdot \mathbf{B}_k^{\text{mod}}; 0, 1\right) \ , \tag{9}$$

where $\mathbf{p}_k^{\text{edit}} := \mathbf{P}_k^{\text{edit}}[\mathbf{R}_k^{\text{src}}]$ denotes the probabilities assigned to the current source bits, extracted analogously to Eq. 3; λ controls the overall source retention strength; $\text{clip}(\cdot; 0, 1)$ restricts each adjusted probability to $[0, 1]$.

Autoregressive Sampling. After bias injection, the current edited token map is sampled from the adjusted distribution

$$\mathbf{R}_k^{\text{edit}} \sim \hat{\mathbf{P}}_k^{\text{edit}} \ . \tag{10}$$

Here, $\hat{\mathbf{P}}_k^{\text{edit}}$ denotes the complete two-way categorical distribution reconstructed from $\hat{\mathbf{p}}_k^{\text{edit}}$ in Eq. 9. Each sampled token map is incorporated into the evolving edited history and conditions the prediction at the next scale. As shown in Figure 3, because the coarse scales establish the global layout and low-frequency structure, DecoEdit fixes the first γ scales to the corresponding source token maps, providing a structural anchor for subsequent generation For the first γ scales, DecoEdit bypasses this sampling process and directly reuses the corresponding source token maps:

$$\mathbf{R}_k^{\text{edit}} \leftarrow \mathbf{R}_k^{\text{src}} \ , \ k \leq \gamma \ . \tag{11}$$

After K scales are completed, the edited token maps are decoded to obtain the edited image $\mathcal{I}^\star$.

5 Experiments

5.1 Experimental Settings

Dataset. We conduct our evaluation on **PIE-Bench** (Ju et al., 2024), a widely used benchmark for text-guided image editing. It covers 10 edit types, such as object addition, removal, modification, attribute editing, pose adjustments, style/material transfer, color changes, and background alterations. Each example provides a source image, a source prompt, a target prompt, and a manually annotated editing mask. Note that these masks are used exclusively for evaluation and are not accessible to any method during inference.

Evaluation Metrics. We evaluate editing quality along two dimensions: **background preservation** and **text alignment**. To measure **background preservation**, we compute *PSNR*, *LPIPS* (Zhang et al., 2018), *MSE*, and *SSIM* (Wang et al., 2004) between the source and edited images over the non-edited regions. To measure **text alignment**, we report the CLIP similarity (Radford et al., 2021) between the target prompt and both the whole edited image (i.e., *CLIP-Whole*) and the edited region (i.e., *CLIP-Edited*).

	Base Model	Background Preservation				Text Alignment	
		PSNR ↑	LPIPS ($\times 10^3$) ↓	MSE ($\times 10^4$) ↓	SSIM ($\times 10^2$) ↑	CLIP-Whole ↑	CLIP-Edited ↑
P2P	SDv1.4	17.87	208.80	219.88	71.14	25.01	22.44
Pix2Pix-Zero	SDv1.4	20.44	172.22	144.12	74.67	22.80	20.54
MasaCtrl	SDv1.4	22.17	106.62	86.97	79.67	23.96	21.16
PnP	SDv1.4	22.28	113.40	84.20	79.05	25.41	22.55
RF-Inversion	FLUX.1 [dev]	20.64	193.68	123.33	76.34	25.20	22.38
StableFlow	FLUX.1 [dev]	21.64	92.28	115.21	84.94	24.65	21.70
RF-Edit	FLUX.1 [dev]	23.22	131.18	75.00	81.44	25.22	22.40
FlowEdit	FLUX.1 [dev]	21.74	113.83	98.75	83.46	26.19	22.19
AREdit	Infinity-2B	24.19	87.00	—	83.70	25.42	<u>22.77</u>
EditMod*	Infinity-2B	23.75	97.62	<u>67.30</u>	79.56	25.85	22.55
DecoEdit	Infinity-2B	**24.63**	<u>78.89</u>	79.04	<u>86.18</u>	<u>25.90</u>	22.62
DecoEdit	Infinity-8B	<u>24.49</u>	**73.64**	**65.28**	**88.69**	**26.57**	**22.92**

Table 1: Quantitative comparison of image editing on PIE-Bench. **Bold** and <u>underlined</u> indicate the best and second-best among methods based on Infinity. * denotes our re-implementation.

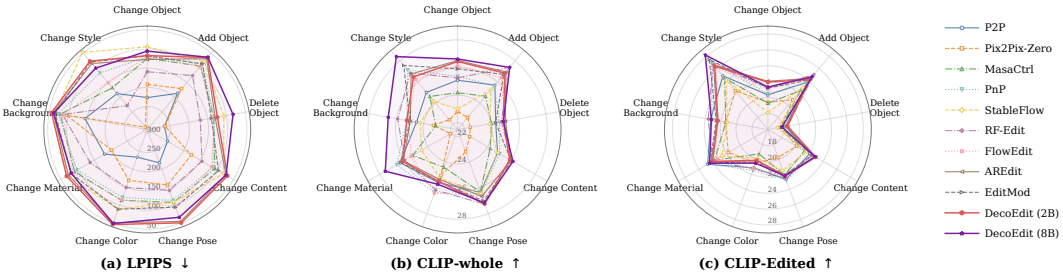


Figure 4: Quantitative comparison on PIE-Bench across editing categories.

Implementation Details. We implement DecoEdit on the pretrained Infinity-2B and Infinity-8B models (Han et al., 2025). We set the preservation evidence strength in Equation 5 to $\beta = 0.5$, the preservation bias threshold in Equation 6 to $\tau = 0.93$, the injection strength in Equation 9 to $\lambda = 0.75$, and the number of reused source scales in Equation 11 to $\gamma = 4$. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

Baselines. We compare our method with ten representative training-free image-editing methods, which are grouped into three categories according to their generative paradigms. **(1) Diffusion-based methods:** P2P (Hertz et al., 2022), Pix2Pix-Zero (Parmar et al., 2023), MasaCtrl (Cao et al., 2023), and PnP (Tumanyan et al., 2023), all of which use SDv1.4 (Rombach et al., 2022) as their backbone. *(2) Flow-based methods:* RF-Inversion (Rout et al., 2024), StableFlow (Avrahami et al., 2025), RF-Edit (Wang et al., 2024), and FlowEdit (Kulikov et al., 2025), all of which use FLUX.1 [dev] Labs et al. (2025) as their backbone. **(3) Next-scale AR methods:** AREdit (Wang et al., 2025) and EditMod (Fang et al., 2026), both of which use Infinity-2B Han et al. (2025) as their backbone.

5.2 Comparison with Existing Methods

Quantitative Comparison. Table 1 shows that, under the same Infinity-2B backbone, DecoEdit outperforms existing next-scale AR baselines on most background-preservation metrics. It also achieves the highest CLIP-Whole score among the Infinity-2B methods while maintaining a competitive CLIP-Edited score, demonstrating that improved preservation does not compromise target alignment. DecoEdit further exhibits a favorable preservation-editability trade-off compared with diffusion- and flow-based baselines. The category-level results in Figure 4 confirm that these improvements are consistent across diverse editing types rather than driven by only a few categories.

Qualitative Comparison. Figure 5 compares DecoEdit with existing methods. DecoEdit effectively realizes the requested modifications, while preserving the source layout, object identity, and unrelated background content. In contrast, AREdit and EditMod may retain residual source semantics or introduce unintended changes to source details. Diffusion- and flow-based methods may miss

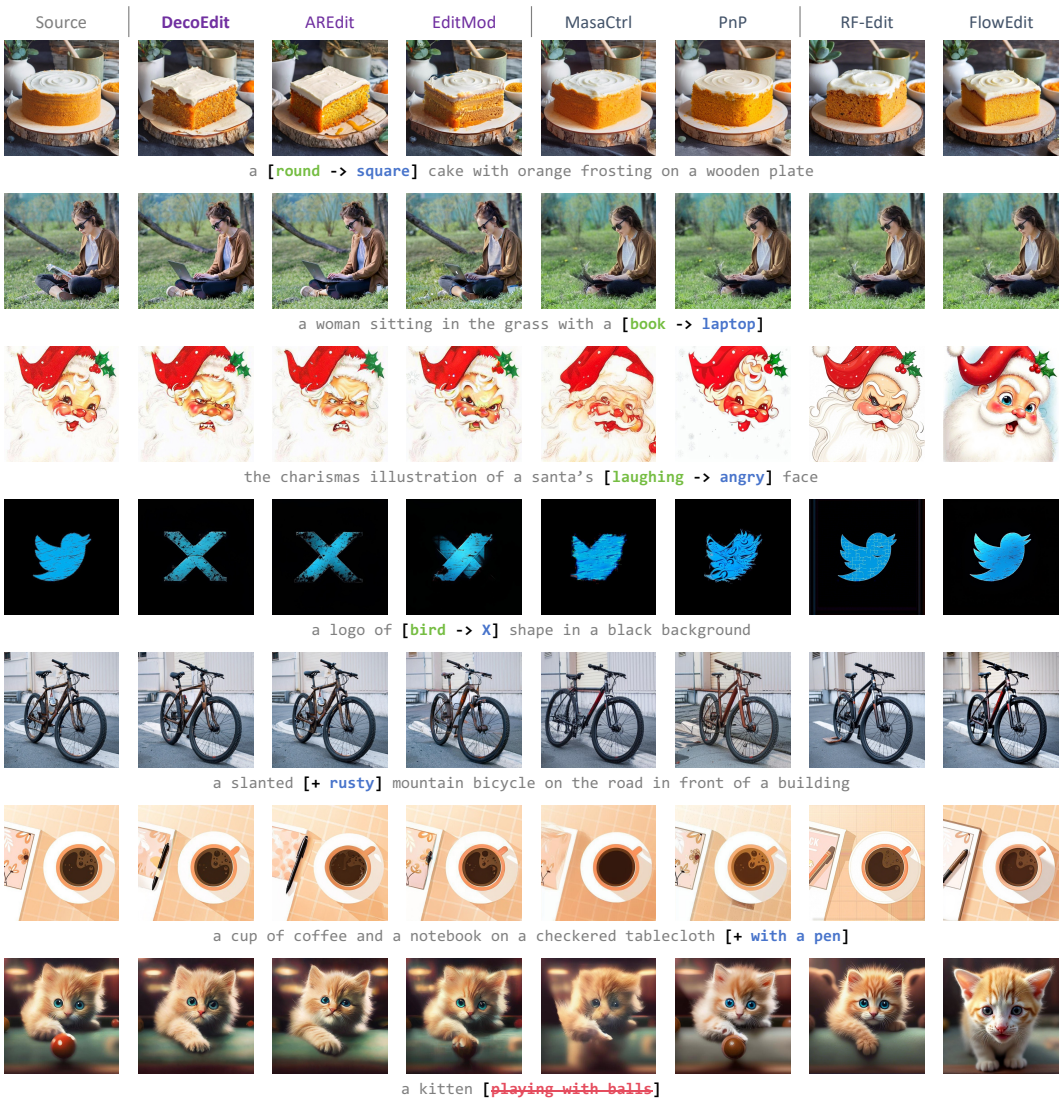


Figure 5: Qualitative comparison of image editing on PIE-Bench.

the requested edits or alter unrelated content. Overall, DecoEdit provides a more favorable visual balance between source preservation and target generation.

5.3 Model Study

Ablation Study. Table 2 and Figure 6(a) demonstrates the contributions of the main components of DecoEdit. **(A)** Removing both *preservation bias* and *preservation evidence* reduces DecoEdit to target-conditioned generation with source token reuse, resulting in weaker source preservation. **(B)** Retaining the *preservation bias* without *preservation evidence* strengthens source preservation but also compensates content requiring generation, thereby limiting editing capability. **(C)** Removing *source token reuse* allows the coarsest scales to be regenerated, leading to larger deviations in global structure. These results demonstrate that all three components contribute to the favorable preservation–editability trade-off achieved by the full model.

Effect of Preservation Evidence Strength β. Figure 6(b) examines the effect of the preservation evidence strength β in Equation 5. A larger β increases the influence of fixed-history contrast on preservation evidence, resulting in weaker evidence for source bits associated with the requested modification. Consequently, less preservation bias is injected for content to generate, improving

	Background Preservation				Text Alignment	
	PSNR ↑	LPIPS ↓	MSE ↓	SSIM ↑	CLIP-Whole ↑	CLIP-Edited ↑
DecoEdit (Full)	24.63	78.89	79.04	86.18	25.90	22.62
(A) – Preservation Bias & Preservation Evidence	16.57	212.32	295.21	71.56	26.57	23.27
(B) – Preservation Evidence	**25.07**	**75.12**	**73.40**	**86.67**	25.70	22.51
(C) – Source Scale Reuse	15.04	282.89	488.39	69.11	**27.87**	**24.55**

Table 2: Ablation study of the main components of DecoEdit on PIE-Bench using Infinity-2B.

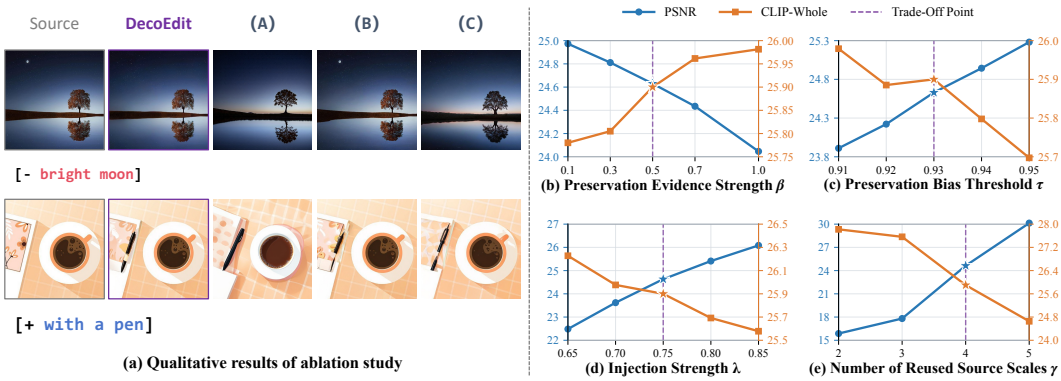


Figure 6: Model study of DecoEdit. (a) Qualitative ablation results ablation study. (b)–(e) Effects of β, τ, λ, and γ on the preservation-editability trade-off.

CLIP-Whole at the cost of lower PSNR. A moderate value of $\beta = 0.5$ provides a favorable balance between source preservation and target alignment.

Effect of Preservation Bias Threshold τ. Figure 6(c) studies the effect of the preservation bias threshold τ in Equation 6. A larger τ increases the probability compensation for under-predicted source bits, leading to higher PSNR and stronger source preservation. However, excessive compensation can restrict the generation of target content, resulting in lower CLIP-Whole. A moderate value of $\tau = 0.93$ achieves a favorable balance between source preservation and target alignment.

Effect of Injection Strength λ. Figure 6(d) studies the effect of the injection strength λ in Equation 9. A larger λ strengthens the injected probability compensation, leading to higher PSNR and improved source preservation. However, excessive compensation can interfere with the generation of target content, resulting in lower CLIP-Whole. A moderate value of $\lambda = 0.75$ achieves a favorable preservation-editability trade-off.

Effect of Number of Reused Source Scales γ. Figure 6(e) analyzes the effect of the number of reused source scales γ in Equation 11. Increasing γ anchors more coarse-scale information to the source image, leading to higher PSNR and stronger preservation of global structure. However, reusing too many source scales restricts large-scale semantic modification and reduces CLIP-Whole. A moderate value of $\gamma = 4$ provides an effective structural anchor while retaining sufficient flexibility for target generation.

6 Conclusion

In this paper, we propose DecoEdit, a training-free image editing framework that decouples source preservation from target generation in next-scale autoregressive models. DecoEdit determines source preservation through fixed-history contrast, which compares source- and target-conditioned predictions under the same source history without being affected by preceding edits. To balance source retention and target generation, the contrast is converted into preservation evidence, which modulates a preservation bias before its injection into the evolving target-conditioned process. Extensive experiments on Infinity-2B and Infinity-8B demonstrate that DecoEdit achieves a better preservation-editability trade-off than existing methods.

AI USE STATEMENT

In this work, we used generative AI tools for helping develop conceptual frameworks, designing or providing feedback on research methodology or experiments, assisting with translation, and interpreting results. We have not used generative AI tools for generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, implementing methods, or cleaning and reformatting datasets, and supporting qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for brainstorming, drafting parts of the research paper, editing the research paper to improve readability, suggesting a structure for the research paper, proposing titles, summarizing or analyzing existing literature, identifying relevant literature, and sourcing or searching for information. We have reviewed all AI-assisted work. AI-assisted methodological suggestions were independently evaluated by the authors, manuscript content and technical claims were checked against the actual method and experimental results, and literature-related suggestions were manually verified against the original sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work focuses on improving training-free image editing with next-scale autoregressive models and does not involve human subjects or the collection of private data. As with general-purpose image editing techniques, the proposed method could potentially be misused to manipulate images without authorization or to create misleading visual content. We encourage responsible use of the method in accordance with applicable data licenses, copyright requirements, and privacy regulations. The authors take responsibility for conducting this research in accordance with standard research integrity and ethical practices.

REPRODUCIBILITY STATEMENT

We make several efforts to facilitate the reproducibility of our work. The main paper provides a complete description of the proposed DecoEdit framework and its inference procedure. Additional implementation details, hyperparameter settings, model configurations, and experimental protocols are provided in the appendix. We also describe the datasets, evaluation metrics, and baseline configurations used in our experiments to enable consistent comparison and reproduction of the reported results. Further qualitative results and ablation studies are included in the supplementary material to support a more comprehensive evaluation of the method. Anonymous source code and evaluation scripts are provided in the supplementary materials.

REFERENCES

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Hongyi Fang, Chuwen Xie, Benjia Zhou, Yu-Xuan Qiu, Chenggong Hu, Zhibin Wang, Chao Chen, Jianbin Qin, and Rui Mao. Model the edit, not the image: Visual autoregressive editing from a source-centric perspective. *arXiv preprint arXiv:2608.09057*, 2026.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744. IEEE, 2025.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *International Conference on Learning Representations*, volume 2024, pp. 23395–23422, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730. IEEE, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

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.

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. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

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

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

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

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

Yufei Wang, Lanqing Guo, Zhihao Li, Jiaxing Huang, Pichao Wang, Bihan Wen, and Jian Wang. Training-free text-guided image editing with visual autoregressive model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17577–17586. IEEE, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18431–18441, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9026–9036, 2024.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *International Conference on Learning Representations*, pp. 90844–90868, 2025.