ConsistEdit: Highly Consistent and Precise Training-free Visual Editing

ZIXIN YIN, Hong Kong University of Science and Technology

LING-HAO CHEN, Tsinghua University and International Digital Economy Academy

LIONEL NI, Hong Kong University of Science and Technology, Guangzhou and Hong Kong University of Science and Technology

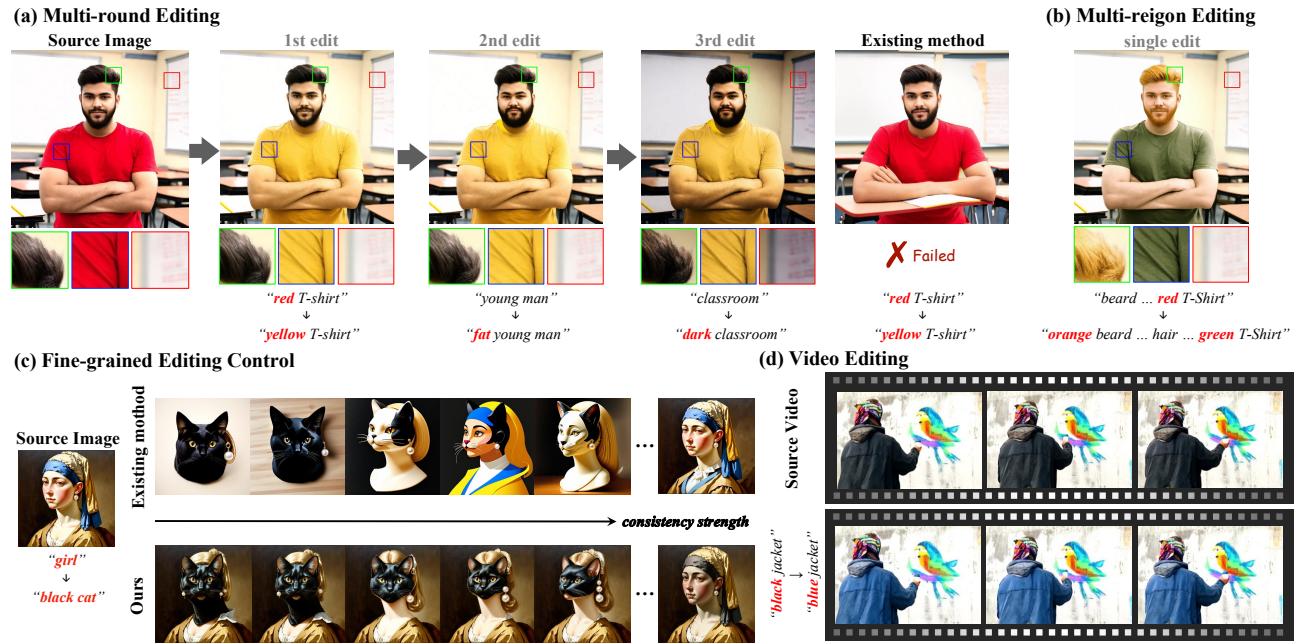XILI DAI, Hong Kong University of Science and Technology, Guangzhou

Fig. 1. **(a) ConsistEdit** enables multi-round editing by allowing users to specify both the target region and the nature of the editing through prompts. Unlike existing methods, it can perform structure-preserving (hair, clothing folds) and shape-changing with identity-preserving edits in edited regions while keeping non-edited regions intact. **(b)** ConsistEdit handles multi-region edits in one pass and preserves both the edited structure and unedited content. **(c)** Our method enables smooth control over consistency strength in the edited region. In contrast, existing approaches lack smooth transitions and often alter non-edited areas. **(d)** Beyond image editing and rectified flow models, ConsistEdit generalizes well to all MM-DiT variants, including diffusion and video models.

Recent advances in training-free attention control methods have enabled flexible and efficient text-guided editing capabilities for existing image and video generation models. However, current approaches struggle to simultaneously deliver strong editing strength while preserving consistency with the source. For instance, in color-editing tasks, they struggle to maintain structural consistency in edited regions while preserving the rest intact. This limitation becomes particularly critical in multi-round and video editing, where visual errors can accumulate over time. Moreover, most existing methods enforce global consistency, which limits their ability to modify individual attributes such as texture while preserving others, thereby hindering fine-grained editing. Recently, the architectural shift from U-Net to Multi-Modal Diffusion Transformers (MM-DiT) has brought significant improvements in generative performance and introduced a novel mechanism for integrating text and vision modalities. These advancements pave the way for overcoming challenges that previous methods failed to resolve. Through an in-depth analysis of MM-DiT, we identify three key insights into its attention mechanisms. Building on these, we propose ConsistEdit, a novel attention control method specifically tailored for MM-DiT. ConsistEdit incorporates vision-only attention control, mask-guided pre-attention fusion, and differentiated manipulation of the query, key, and value tokens to produce consistent, prompt-aligned edits. Extensive experiments demonstrate that ConsistEdit achieves state-of-the-art performance across a wide range of image and video editing tasks, including both structure-consistent and structure-inconsistent scenarios. Unlike prior methods, it is the first approach to perform editing across all inference steps and attention layers without handcraft, significantly enhancing reliability and consistency, which enables robust multi-round and multi-region editing. Furthermore, it supports progressive adjustment of structural consistency, enabling finer control. ConsistEdit represents a significant advancement in generative model editing and unlocks the full editing potential of MM-DiT architectures. Here is the project website.

1 Introduction

Attention control techniques, which manipulate the query (Q), key (K), and value (V) tokens in the attention mechanism, have been widely adopted because their training-free nature enables flexible

and efficient extensions of generative models to image and video editing tasks. For example, in image editing, Prompt-to-Prompt (P2P) [Hertz et al. 2023] introduced a method to control cross-attention layers, enabling text-guided editing without the need for additional data or retraining. This inspired follow-up work in video editing, such as Video-P2P [Liu et al. 2024b], while Masactrl [Cao et al. 2023] extended P2P from rigid to non-rigid editing.

Despite these advancements, there are still two fundamental challenges in text-guided editing: **(1)** the method must not only modify content in according to the input text but also preserve consistency in both editing and non-editing regions. In editing regions, for example, structure should remain stable when changing color, and the character identity must stay recognizable when adjusting shape. In non-editing regions, all visual elements should remain identical to the original. However, these two requirements are often not simultaneously satisfied in existing training-free methods [Cao et al. 2023; Hertz et al. 2023; Jiao et al. 2025], leading to unacceptable results in tasks such as color and material modifications. As shown in part Fig. 1 (a), existing methods tend to introduce noticeable changes but compromise consistency in both edited and non-edited regions. Therefore, maintaining editing strength and consistency is essential for multi-round and video editing, where both iterative accumulation and additional temporal dimension can exacerbate visual errors. **(2)** Beyond the inability to satisfy both requirements, existing methods typically enforce overall (*e.g.*, texture and structure) consistency, which severely limits performance in fine-grained editing. When a task requires preserving texture while altering structure, or vice versa, these methods often fail, see Fig. 1 (c). Allowing more targeted control, such as focusing consistency on structure alone, would enable more flexible and effective editing.

Amid these unresolved issues, the field of image and video generation has undergone astonishing advancements due to the transition in architectures from U-Net [Rombach et al. 2022] to Multi-Modal Diffusion Transformer (MM-DiT) [Esser et al. 2024], shedding a new light on solving these problems mentioned above. Since training-free attention control methods heavily depend on the underlying architecture of the generative model, it is crucial to study how to tailor it to MM-DiT. To date, only one work, DiTCtrl [Cai et al. 2025], has investigated attention control in MM-DiT, and even that does not target editing tasks. Instead, it targeted multi-prompt long video generation. As a result, the lack of investigation into the attention mechanisms of MM-DiT in editing tasks significantly limits existing approaches [Deng et al. 2025; Wang et al. 2025].

To address this gap, we conduct a detailed study of MM-DiT's attention architecture, starting by contrasting it with the more commonly studied U-Net. In U-Net, cross-attention governs text guidance, while self-attention drives visual generation, resulting in a two-stage separation of modalities. In contrast, MM-DiT merges textual and visual information, applying self-attention to jointly process modalities. Through in-depth analysis and experimental exploration, we derive three key insights of MM-DiT models:

- Vision-only is crucial: Editing effectiveness relies on modifying only the vision parts, since interfering with text tokens often leads to generation instability (Fig. 4).

- Homogeneous for all layers: Visualizations of the vision parts of Q, K, and V across attention layers (Fig. 2) show that, unlike U-Net, each layer in MM-DiT retains rich semantic content. Thus, attention control must be applied to all layers.
- Strong structure controllability from Q and K: Applying attention control solely on the vision parts of Q, K results in strong controllable structural preservation (Fig. 9).

By grounding these insights, we introduce ConsistEdit, a novel attention control method specifically adapted to MM-DiT to address the challenges through three core operations: (1) Vision-only attention control: attention control is applied solely to the vision parts across all layers; (2) Pre-attention mask fusion: editing and non-editing regions are fused before the attention calculation; (3) Differentiated control for Q, K, and V: we apply distinct control mechanisms to Q and K for structure, and V for content.

Through extensive experiments, we show that ConsistEdit enables structurally consistent at finer levels such as lighting and shading in edited regions, while keeping non-edited regions unchanged. As a result, ConsistEdit can address the two fundamental challenges in text-guided editing mentioned before: **(1)** ConsistEdit achieves **state-of-the-art (SOTA)** performance across diverse editing tasks including structure-consistency and -inconsistency tasks, enabling iterative multi-round editing, as well as single-pass multi-region editing, see Fig. 1 (a) (b). Additionally, it demonstrates strong generalization across diverse generation models and editing tasks, including video editing, showcasing its versatility and practical potential, as shown in Fig. 1 (d) and 13. **(2)** Instead of enforcing overall consistency, ConsistEdit supports progressive adjustment of structural consistency, allowing fine-grained control in various downstream tasks, as shown in Fig. 1 (c).

To our best knowledge, ConsistEdit is the **first** approach that enables editing across all steps and layers without manual parameter adjustment, significantly improving reliability and consistency. Overall, we list our contributions as follows.

- We identify three key insights from MM-DiT foundation generation models that enable effective training-free attention control for editing tasks.
- We propose ConsistEdit, a novel attention control approach designed to extend the editing capabilities of MM-DiT-based models.
- Our method supports both structure-consistent and -inconsistent edits while maintaining fidelity in non-edited regions. Extensive experiments demonstrate that ConsistEdit sets new SOTA results in both image and video editing tasks, and enables reliable multi-round and multi-region editing.

2 Related Work

2.1 Text-to-image/video Generation

Early visual generation methods were primarily based on GANs [Reed et al. 2016; Tao et al. 2022; Wang et al. 2023; Yu et al. 2023] due to their ability to synthesize high-fidelity content. However, diffusion models [Guo et al. 2024; Ho et al. 2020; Reed et al. 2016; Rombach et al. 2022] have demonstrated superior generative performance, with U-Net-based architectures [Rombach et al. 2022] becoming the dominant paradigm. As U-Net designs encounter scalability

limitations in data and model parameter size, the field has progressively shifted toward transformer-based architectures, particularly diffusion transformers (DiT) [Peebles and Xie 2023]. Among these, MM-DiT [Esser et al. 2024] has emerged as a widely adopted backbone for text-conditional generation. Many recent state-of-the-art models [AI 2024; Esser et al. 2024; Kong et al. 2024; Labs 2024; Liu et al. 2025; Yang et al. 2024] build upon MM-DiT, achieving remarkable performance, such as SD3 [Esser et al. 2024] and FLUX [Labs 2024] for image generation, as well as CogVideoX [Yang et al. 2024] for video generation. In this work, we tailor a new attention control method for MM-DiT-based models.

2.2 Text-guided Editing

Early work focused on training-based approaches that leveraged generative models to achieve controllable image editing [Karras et al. 2019]. With the rapid advancement of generative models, attention has shifted toward training-free editing methods, which offer greater flexibility and efficiency. These training-free approaches generally fall into two categories: sampling-based and attention-based methods. Sampling-based approaches introduce controlled randomness into the generation process to enable flexible editing [Huberman-Spiegelglas et al. 2024; Jiao et al. 2025; Kulikov et al. 2024], while attention-based methods achieve editing ability by directly altering attention tokens. Prompt-to-Prompt [Hertz et al. 2023] was the first to introduce attention manipulation on the cross-attention layers of U-Net, adopted in many subsequent editing methods [Chen et al. 2024; Yang et al. 2023]. Video-P2P [Liu et al. 2024b] extends this cross-attention control to video editing. FateZero [Qi et al. 2023] combines self-attention with a blending mask derived from cross-attention features of the source prompt. Methods such as MasaCtrl [Cao et al. 2023] and DiTCtrl [Cai et al. 2025] employ similar attention control strategies, applied to U-Net and MM-DiT architectures respectively. Despite their differences, all existing attention control methods can be understood as multi-branch frameworks [Cai et al. 2025; Cao et al. 2023; Ju et al. 2024; Rout et al. 2025; Wang et al. 2025], and can be uniformly expressed as special cases of Eq. 3. Notably, all above methods rely on selectively manipulating specific inference steps or attention layers, which limits their robustness and consistency with respect to the source. In contrast, our approach is the first one requires no manual selection of steps or layers.

3 Method

3.1 Preliminary

3.1.1 Generation procedure. The current visual generation procedure is a systematical method which includes generation algorithm and foundation network architecture. $z^T \rightarrow z^{T-1} \rightarrow \ldots \rightarrow z^t \rightarrow \ldots \rightarrow z^0$ shows the procedure for generating the final image or video from random noise z^T in T steps. The generation algorithm could be latent diffusion, flow matching, or rectified flow.

Beyond the generation algorithm, the foundation network architecture plays a crucial role in affecting the final generation results. In each step, the network $f(\cdot)$ integrates the text prompt tokens **P** and the result of previous step z^t to generate the result of the next step z^{t-1}: $z^t \xrightarrow{f(z^t|\mathbf{P})} z^{t-1}$. The network $f(\cdot)$ can be U-Net or MM-DiT. It takes the pair of $(z^t, \mathbf{P})$ as input, which present the

vision z^t and text **P** tokens respectively. It goes through each layer of the network, and Eq. 1 shows how each attention layer works.

$$\{z^t(l), \mathbf{P}(l)\} \xrightarrow{g(\cdot)} Q^l, K^l, V^l,$$

$$z^t(l+1) = \text{Attention}(Q^l, K^l, V^l) = V^l \cdot \text{Softmax}\left(\frac{Q^l (K^l)^\top}{\sqrt{d}}\right). \quad (1)$$

We unify the formulation of cross-attention and self-attention in Eq. 1. The function $g(\cdot)$ denotes a pre-attention operation that plays different roles in cross-attention and self-attention layers. Specifically, in the l-th layer of a U-Net, if it is a cross-attention layer, $g(\cdot)$ maps the text tokens $\mathbf{P}(l)$ to K^l and V^l, and maps the vision tokens $z^t(l)$ to Q^l. In contrast, for self-attention layers, $g(\cdot)$ ignores the text tokens and maps $z^t(l)$ to all of Q^l, K^l, and V^l.

In contrast, MM-DiT is a self-attention-only architecture, without cross-attention. Before computing attention, each MM-DiT block applies a pre-attention transformation $g(\cdot)$, which includes operations such as MLP modulation, residual connections, normalization, and other components. In each block, the pre-attention $g(\cdot)$ maps the vision $z^t(l)$ and text $\mathbf{P}(l)$ tokens respectively and concatenate every vision and text pair to get Q^l, K^l, V^l. In other words, Q^l, K^l, V^l all contain text and vision parts.

3.1.2 Inversion. The inversion procedure aims to accurately reverse the generation process to recover the initial noise z^T that can reconstruct the real image or video tokens z^0.

3.1.3 Editing. The original editing method can trace back to the image processing era [Jähne 2005] and the task was formulated as follows:

$$I_{tg} = (1 - M) \odot I_s + M \odot I_e, \quad (2)$$

where the user offers source image I_s and editing regions (mask M). The goal of the editing task was to generate the edited content I_e and then blend it back to the source image while preserving the non-edited regions of the source image. $\odot$ denotes the element-wise Hadamard product of two matrices.

3.1.4 Attention control approach for training-free editing. The current visual editing methods in the background of generation models [Cao et al. 2023; Hertz et al. 2023], leverage the attention control approach to extend the editing capability of the foundation generation model in a training-free manner. In concrete, they employ a dual-network architecture: one network is dedicated to reconstructing the original source given the prompt tokens $\mathbf{P}_s$ and random noise z^T, while the other is focused on editing. The dual-network shares the same network parameters.

We formulate the procedure of the editing in a way of dual-network architecture. By applying the generation process to the source I_s, we obtain the full generation trajectory of the source tokens: $z^T \rightarrow z_s^{T-1} \rightarrow \cdots \rightarrow z_s^t \rightarrow \cdots \rightarrow z_s^0$. At each step, the update follows $z_s^t \xrightarrow{f(z_s^t|\mathbf{P}_s)} z_s^{t-1}$, and each attention layer is computed as $z_s^t(l+1) = \text{Attention}(Q_s^l, K_s^l, V_s^l)$.

The generation procedure for the target I_{tg} starting from the same noise, $z^T \rightarrow z_{tg}^{T-1} \rightarrow \cdots \rightarrow z_{tg}^t \rightarrow \cdots \rightarrow z_{tg}^0$, and each step $z_{tg}^t \xrightarrow{f(z_{tg}^t|\mathbf{P}_t)} z_{tg}^{t-1}$, are very similar to that of the source, but with different attention operation which we call it attention control.

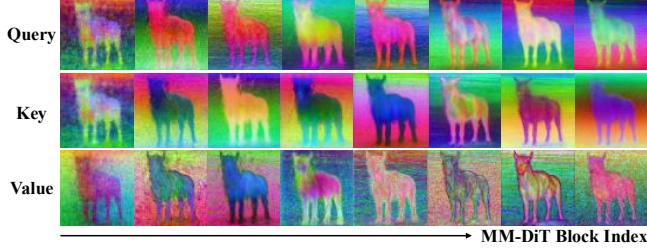Fig. 2. Visualization of projected Q, K, V vision tokens in attention layers of the MM-DiT blocks at 15th sampling step of prompt "A standing horse."

When we get the $Q_{tg}^l, K_{tg}^l, V_{tg}^l$ from the l-th attention layer of t-th step in the target generation procedure, the attention control no longer apply them directly to attention operation, but replace some of them from the generation procedure of the source.

$$\{z_{tg}^t(l), \mathbf{P}_{tg}\} \xrightarrow{g(\cdot)} Q_{tg}^l, K_{tg}^l, V_{tg}^l,$$
$$f_o^t = \text{Attention}(Q_{tg}^l, K_s^l, V_s^l | \mathbf{M}_c),$$
$$f_b^t = \text{Attention}(Q_{tg}^l, K_s^l, V_s^l | \mathbf{1} - \mathbf{M}_c), \quad (3)$$
$$z_{tg}^t(l+1) = (\mathbf{1} - \mathbf{M}) \odot f_b^t + \mathbf{M} \odot f_o^t.$$

Eq. 3 is a example of attention control formulation of MasaCtrl [Cao et al. 2023] and DiTCtrl [Cai et al. 2025], where they replace the K, V from the source in self-attention layers. Here, $\mathbf{M}_c$ and $\mathbf{M}$ donate masks extracted from attention maps[1] of the source and target generation procedure, respectively. $\mathbf{M}$ is used as the attention mask.

3.2 ConsistEdit: A New Attention Control for MM-DiT

3.2.1 The analysis of the attention mechanism in MM-DiT. MM-DiT [Esser et al. 2024] fundamentally differs from U-Net [Rombach et al. 2022] in its attention mechanism. In U-Net, cross-attention handles text guidance, while self-attention focuses on visual content generation, creating a two-stage process. In contrast, MM-DiT merges text and visual information and employs self-attention to process both modalities simultaneously. This architecture shift renders existing U-Net-based attention control methods ineffective.

For instance, DiTCtrl [Cai et al. 2025] adopts the strategy of MasaCtrl [Cao et al. 2023] by applying attention control to the latter blocks of the model. This design originates from the downsampling–upsampling structure of the U-Net encoder–decoder architecture, where MasaCtrl performs edits in the decoder stages. However, MM-DiT does not exhibit such stage separation, as it lacks a distinct decoder stage on which editing can be focused, as illustrated by the PCA decomposition visualization in Fig. 2. Consequently, directly transferring this strategy leads to structural artifacts, as shown in Fig. 10, Fig. 7, and Fig. 8. Further experiments (Fig. 11) confirm that editing across all blocks yields superior results.

Moreover, Fig. 4 compares FireFlow [Deng et al. 2025] and RF-Solver [Wang et al. 2025], each using their original attention control method on either all parts (original) or vision-only parts of tokens. Under higher consistency strength, the original approach often fails,

[1]DiTCtrl adopts all-one masks for both $\mathbf{M}_c$ and $\mathbf{M}$ in editing tasks, despite using extracted masks for long video generation.

while vision-only edits better preserve the source content. At lower consistency strength, both approaches perform similarly. These results clearly *highlight* that restricting attention control to the vision parts is critical for robust editing. The detailed implementations are provided in Appendix A.

Hence, we would execute the attention control on the vision parts across all blocks. $\hat{Q}_s^l$, $\hat{K}_s^l$ and $\hat{V}_s^l$ have the same vision parts as Q_s^l, K_s^l and V_s^l but the text parts are from Q_{tg}^l, K_{tg}^l and V_{tg}^l.

3.2.2 Structural consistency in edited region. Besides the visual parts only replaced from source components, we also move the blending procedure before the attention operation. Then, after extensive exploration of all potential combinations of Q, K, and V from the source and target generation procedure, we find the combination shown in Eq. 4 best preserves structural consistency. The mask $\mathbf{M}$, representing the editing region, is extracted from the source attention maps similar to Cai et al. [2025] and is applied only to the vision parts. We refer to this method as *Structure Fusion*. Furthermore, the spatially-resolved visualizations in Fig. 2 enable us to perform mask blending directly based on spatial regions. To enable controllable editing strength, we define the **consistency strength** α as a ratio of steps for applying attention control, which determines the level of structural preservation during editing. Technically, in structure consist editing, the attention calculation can be formulated as:

$$\tilde{Q}_{tg}^l = \begin{cases} \mathbf{M} \odot \hat{Q}_s^l + (1 - \mathbf{M}) \odot Q_{tg}^l, & \text{if } t > (1-\alpha)T \\ \mathbf{M} \odot Q_{tg}^l + (1 - \mathbf{M}) \odot Q_{tg}^l, & \text{otherwise} \end{cases},$$

$$\tilde{K}_{tg}^l = \begin{cases} \mathbf{M} \odot \hat{K}_s^l + (1 - \mathbf{M}) \odot K_{tg}^l, & \text{if } t > (1-\alpha)T \\ \mathbf{M} \odot K_{tg}^l + (1 - \mathbf{M}) \odot K_{tg}^l, & \text{otherwise} \end{cases}, \quad (4)$$

$$z_{tg}^t(l+1) = \text{Attention}(\tilde{Q}_{tg}^l, \tilde{K}_{tg}^l, V_{tg}^l).$$

This operation enforces structural consistency while enabling precise text control to adjust appearance and texture.

3.2.3 Content preservation in non-edited region. We find that using $\hat{Q}_s^l$ and $\hat{K}_s^l$ in the non-editing regions can maintain structural consistency, but often leads to color shifts. To achieve high-fidelity content preservation, we further use $\hat{V}_s^l$ in non-editing regions, which yields the best results. We refer to following strategy as *Content Fusion*.

As a result, Eq. 5 defines the final formulation of ConsistEdit:

$$\tilde{Q}_{tg}^l = \begin{cases} \mathbf{M} \odot \hat{Q}_s^l + (1 - \mathbf{M}) \odot \hat{Q}_s^l, & \text{if } t > (1-\alpha)T \\ \mathbf{M} \odot Q_{tg}^l + (1 - \mathbf{M}) \odot \hat{Q}_s^l, & \text{otherwise} \end{cases},$$

$$\tilde{K}_{tg}^l = \begin{cases} \mathbf{M} \odot \hat{K}_s^l + (1 - \mathbf{M}) \odot \hat{K}_s^l, & \text{if } t > (1-\alpha)T \\ \mathbf{M} \odot K_{tg}^l + (1 - \mathbf{M}) \odot \hat{K}_s^l, & \text{otherwise} \end{cases}, \quad (5)$$

$$\tilde{V}_{tg}^l = \mathbf{M} \odot V_{tg}^l + (1 - \mathbf{M}) \odot \hat{V}_s^l,$$

$$z_{tg}^t(l+1) = \text{Attention}(\tilde{Q}_{tg}^l, \tilde{K}_{tg}^l, \tilde{V}_{tg}^l).$$

4 Experiments

4.1 Setup

4.1.1 Baselines. We compare our method against several recent state-of-the-art approaches built upon MM-DiT, including UniEdit-Flow [Jiao et al. 2025], DiTCtrl [Cai et al. 2025], FireFlow [Deng et al. 2025], RF-Solver [Wang et al. 2025], and SDEdit [Meng et al.
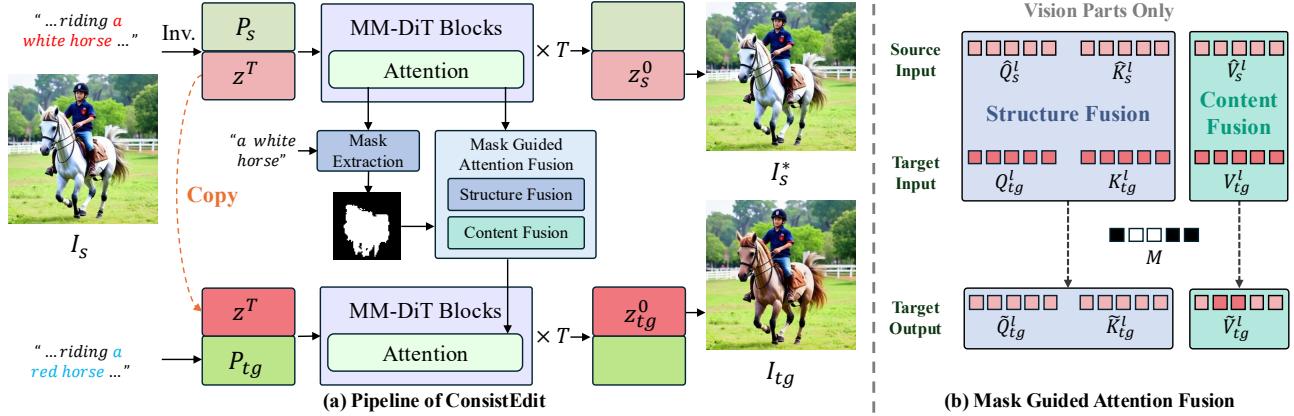
(a) Pipeline of ConsistEdit

(b) Mask Guided Attention Fusion

Fig. 3. **(a)** shows the **ConsistEdit** pipeline. Given a real image or video I_s and source text tokens P_s, we first invert the source to obtain the vision tokens z^T, which is concatenated with the target prompt tokens P_{tg} and passed into the generation process to produce the edited image or video I_{tg}. During inference, a mask M generated by our extraction method delineates editing and non-editing regions. We apply structure and content fusion to enable prompt-aligned edits while preserving structural consistency within edited regions and maintaining content integrity elsewhere. **(b)** illustrates the mask-guided attention fusion for timesteps where $t > (1 - \alpha)T$, which is applied exclusively to the vision parts, while the text parts remain unchanged.
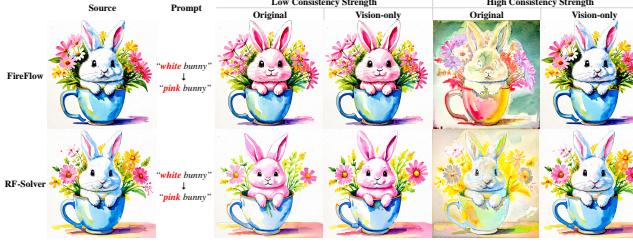

Fig. 4. Comparison of V token swapping strategies for content consistency. Swapping vision-only V tokens leads to superior content consistency under high consistency strength settings, while maintaining comparable editing capability to original methods when the consistency strength is low.

2022]. We focus exclusively on MM-DiT-based baselines, as previous works [Deng et al. 2025; Jiao et al. 2025] and our preliminary experiments (Fig. 6) show that U-Net-based methods perfom significantly worse. Methods (SDEdit) that can be adapted to MM-DiT are included in comparisons, while those cannot be transferred are excluded.

4.1.2 Implementation. We primarily conduct experiments using Stable Diffusion 3 Medium (*a.k.a.* SD3) [Esser et al. 2024] for image generation and CogVideoX-2B [Yang et al. 2024] for video generation, both of which employ a pure MM-DiT architecture. Unless otherwise specified, we use the Euler sampler and adopt UniEdit-Flow [Jiao et al. 2025] for inversion. For all baseline methods, we carefully tune the hyperparameters to ensure a fair comparison. Implementation details are provided in Appendix A.

4.1.3 Benchmark. We adopt prompts from PIE-Bench [Ju et al. 2024] which comprises 700 editing pairs across 10 types of edits. Although our method is fully compatible with inversion methods, we adopt a noise-to-image setting to better isolate and highlight the editing capabilities, minimizing the influence of reconstruction and inversion quality. To ensure fair comparison across baselines,

Table 1. Quantitative results of structural consistency comparison with RF-Solver and FireFlow using Canny SSIM ↑.

Edit Method	Sampler	
	RF-Solver	FireFlow
Fix seed	0.5507	0.5557
RF-Solver [Wang et al. 2025]	0.6225	—
FireFlow [Deng et al. 2025]	—	0.5136
Ours	**0.8714**	**0.8776**

we use a fixed sampler and identical random seeds for each method within a comparison group, so that source images are consistent across all methods. For structure-consistent image editing, we adopt prompts on two tasks that require preserving the original structure: *change color* and *change material*, covering 80 image pairs in total. For structure-inconsistency image editing, we use the remaining cateogries including *Change Object*, *Add Object*, *Delete Object*, *Change Content*, *Change Style*, etc.

4.1.4 Metrics and settings. Unlike the original PIE-Bench, which uses structural distance [Tumanyan et al. 2022] to evaluate structural similarity, we employ the Structural Similarity Index (SSIM) [Wang et al. 2004] computed on Canny edge maps [Canny 1986] borrowed from Zhao et al. [2023] for a more accurate assessment. To evaluate the preservation of non-edited regions (*a.k.a.* BG preservation), we compute PSNR and SSIM exclusively on those regions, which are manually annotated by human annotators. The semantic alignment of the edits is assessed using CLIP similarity [Radford et al. 2021], applied to both the entire image and the edited regions.

4.2 Quantitative Evaluation

While prior editing methods [Cai et al. 2025; Cao et al. 2023; Hertz et al. 2023] typically lack quantitative evaluation, we incorporate evaluation metrics inspired by related tasks (*i.e.* PIE-Bench) to more effectively showcase the capabilities of our method.
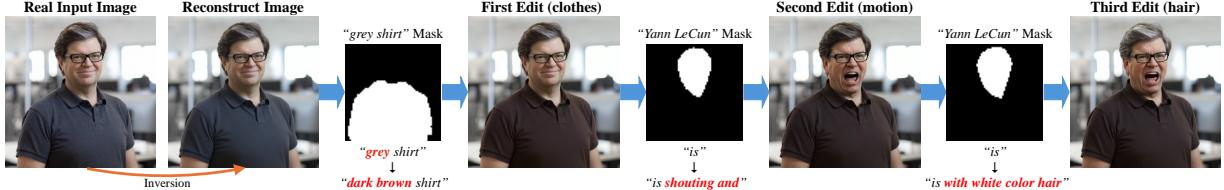
Fig. 5. Real image multi-round editing results. Starting from a real image, we first perform inversion to project it into the latent space. We then sequentially edit the clothing color, motion, and hair.

Fig. 6. Qualitative comparison of methods on real image editing tasks.

Table 2. Quantitative results of *Change Color* and *Change Material* tasks.

Method	Canny	BG Preservation		Clip Similarity ↑	
	SSIM ↑	PSNR ↑	SSIM ↑	Whole	Edited
SDEdit [2022]	0.6795	23.99	0.8697	26.59	22.80
UniEdit-Flow [2025]	0.8029	30.56	0.9554	26.55	22.59
DiTCtrl [2025]	0.8235	29.54	0.9632	26.63	22.97
Ours	**0.8811**	**36.76**	**0.9869**	**27.19**	**23.73**

4.2.1 Evaluation results. Tab. 1 reports a structural consistency comparison with RF-Solver [Wang et al. 2025] and FireFlow [Deng et al. 2025]. We evaluate each baseline using its native sampler and editing strategy, while applying our method under the same sampler but with our own attention control strategy. This setup ensures a fair comparison and demonstrates our method's robustness across varying samplers. As shown in the table, the evaluation metrics of RF-Solver and FireFlow closely match those from fixed-seed generation, suggesting an inability to preserve structural consistency. Therefore, we exclude these two methods from subsequent comparisons on structure-consistent editing tasks. In contrast, our method consistently produces the best structure-preserving results.

Tab. 2 presents the whole benchmark with other baselines. Our method delivers superior results in both preserving source content and executing accurate edits, achieving **state-of-the-art** performance across the board.

4.3 Qualitative Evaluation

In this section, the evaluation begins with structure-consistent editing, highlighting the method's ability to preserve structural consistency. This is followed by demonstrations on real images to validate practical effectiveness. Performance on structure-inconsistent editing is then presented, showcasing adaptability across varied scenarios. Finally, multi-round editing examples are provided, combining both structure-consistent and -inconsistent editing to demonstrate the method's robustness and flexibility.

4.3.1 Structure-consistent editing. Fig. 7 presents a qualitative comparison across all methods on structurally consistent editing tasks. Our approach accurately changes the color or material according to the target prompt while preserving the structure of the edited region same to that of the source image. Notably, beyond merely replacing colors, the edited outputs are also well adapted to the surrounding lighting conditions. In contrast, other methods often produce incorrect or insufficient edits and fail to maintain structural consistency. Additionally, our method faithfully preserves the non-edited regions, whereas others introduce undesirable changes. More results are shown in Appendix B.1.

4.3.2 Structure-consistent editing on real images. We compare our real image editing results with several existing methods, including U-Net-based approaches such as FreePromptEditing [Liu et al. 2024a], EditFriendly [Huberman-Spiegelglas et al. 2024], InfEdit [Xu et al. 2023], PnP-Inv. [Ju et al. 2024], PnP [Tumanyan et al. 2023], and P2P [Hertz et al. 2023]. As shown in Fig. 6, conventional U-Net-based and MM-DiT-based methods all struggle to preserve the non-edited regions and often fail to accurately modify the hat color. In contrast, our method achieves the best performance, delivering precise edits in the target region while preserving the consistency of non-edited areas. Please refer to Appendix B.1 for further examples.

4.3.3 Structure-inconsistent editing. We compare various methods on structure-inconsistent editing tasks in Fig. 8. In these experiments, the consistency strength (α) is set to 0.3, allowing the model to moderately edit structures for improved prompt alignment, while still preserving the overall layout. As shown, our method achieves better results in the edited regions, producing more precise editing with fewer artifacts. Moreover, it more effectively preserves the non-edited areas compared to other approaches, maintaining high content fidelity with respect to the source image. More results are shown in Appendix B.1.

4.3.4 Multi-round interactive editing on real images. Fig. 5 presents an example of multi-round editing on a real input image. The image
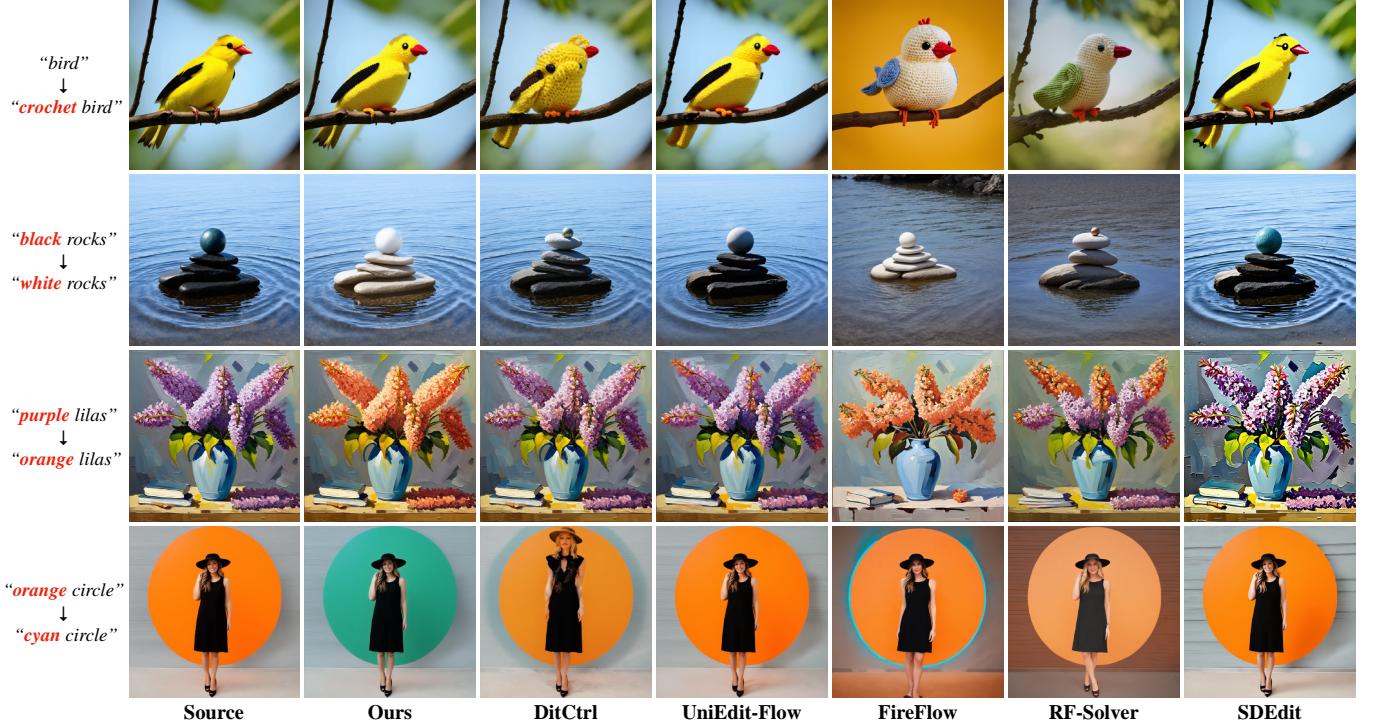
Fig. 7. Qualitative comparison of methods on structure-consistent editing tasks.

is first inverted into the latent space. Then, we perform a series of edits on it, including modifying the clothing color, motion, and hair. These flexible operations and the reliability of the results open up new possibilities for interactive or iterative editing tasks.

4.4 Fine-grained Editing Control

Fig. 9 shows the controllability of structural consistency during editing by varying the consistency strength (α). A high value enforces strict structure preservation, even when the prompt includes shape-altering instructions. At the same time, it still enables accurate texture editing, *e.g.*, color changes, as specified by the prompt. In contrast, a low consistency strength permits structural editing with the prompt. Additionally, the similar color appearance under varying consistency strengths demonstrates the effectiveness of our disentangled structure-preserving control mechanism, enabling precise and independent editing of structure and texture.

Furthermore, thanks to this disentanglement, our method enables smooth and controllable adjustment of consistency strength. In contrast, other methods struggle to maintain stable editing performance across varying strength levels, often relying on specific parameter values, as illustrated in Fig. 10. This property further highlights the potential for integrating a controllable consistency strength slider into interactive editing interfaces.

4.5 Ablation

We conduct two ablation studies to validate the effectiveness of our approach.

Table 3. Evaluation of content preservation in non-edited regions.

	DiffEdit [2023]	V	Q & K	Q & K & V (Ours)
PSNR ↑	51.49	37.98	24.32	**38.85**
SSIM ↑	0.9972	0.9905	0.9286	**0.9917**

4.5.1 Structural consistency in edited regions. As shown in Fig. 11, we conduct an ablation study to investigate the effects of different Q and K tokens swapping strategies. Starting from the same seed ensures a well-initialized structural layout for subsequent editing. Swapping all (text and vision) Q and K tokens preserves structural consistency to a certain extent but significantly impairs text-driven editability, as it discarding the text tokens of target. In contrast, selectively swapping only the vision part of Q and K tokens across all blocks maintains the structural layout of the source image while preserving strong editing capabilities. To verify the necessary of swapping in all layers, we find that only swap the latter half of the model's blocks will substantially weaken structural control and can lead to corrupted generation results. Finally, by incorporating our content fusion method on top of the full-block vision-only Q and K swapping, we further enforce preservation in non-edited regions, achieving the best quality. These findings emphasize the importance of applying editing across all blocks while restricting editing to the vision parts of the attention mechanism.

4.5.2 Content preservation in non-edited regions. Tab. 3 reports PSNR and SSIM scores on the non-edited regions of 80 image pairs from the *Change Color* and *Change Material* tasks in PIE-Bench [Ju

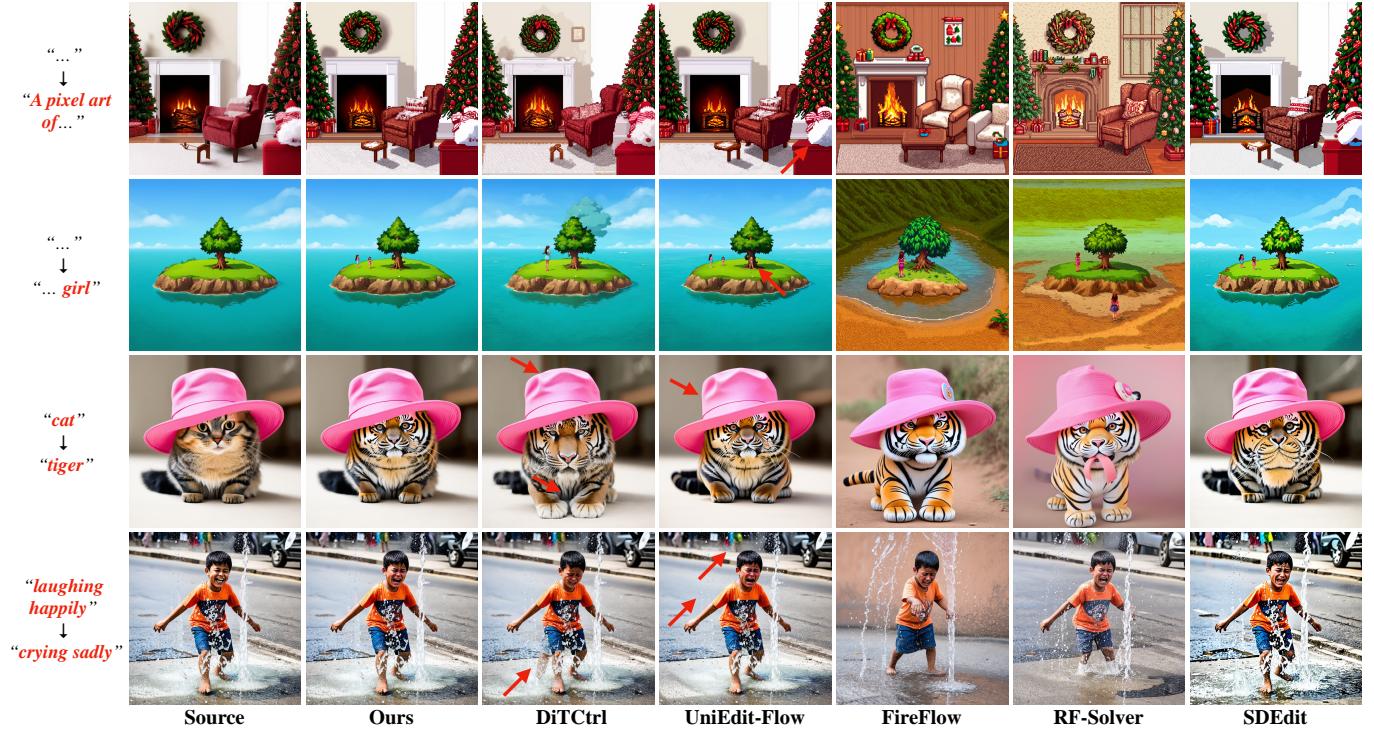| Source | Ours | DiTCtrl | UniEdit-Flow | FireFlow | RF-Solver | SDEdit |

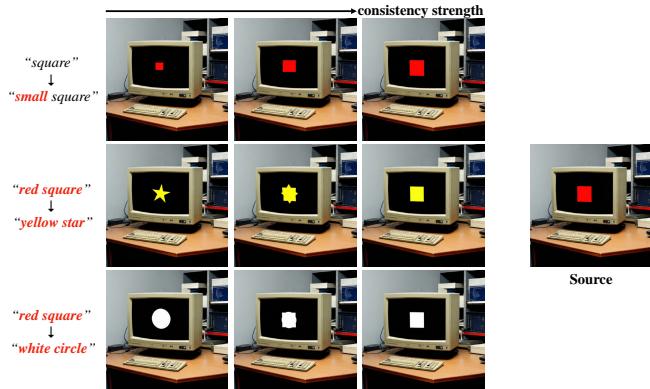Fig. 8. Qualitative comparison of methods on structure-inconsistent editing tasks.

Fig. 9. Effect of consistency strength on structural consistency. High strength strictly enforces structural preservation, while low strength permits prompt-driven shape changes. Texture editing remains consistent, highlighting effective disentanglement.

Fig. 10. Qualitative comparison on consistency strength adjustment.

et al. 2024], evaluating how well different methods preserve content consistency in non-edited regions. All methods use the same binary mask, extracted using our mask extraction method. According to the results in Fig. 12, we can see a hard replacement strategy described in DiffEdit [Couairon et al. 2023] introducing visible artifacts at transition boundaries. Secondly, swapping only the vision tokens of Q and K maintains structural consistency but introduces slight color shifts, which degrade metric scores in Tab. 3. In contrast, swapping only the vision part of the V tokens yields a more stable preservation.

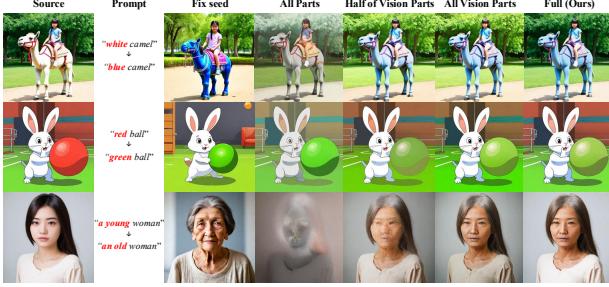Finally, Tab. 3 and Fig. 3 shows that combining vision-token swaps

Fig. 11. Ablation study on attention control for structure consistency. We compare (1) fixed seed results, (2) swapping all Q and K tokens across all blocks, (3) swapping only the vision part of Q and K tokens in the last half of the blocks, (4) swapping only the vision part of Q and K tokens in all blocks, and (5) adding our non-editing region consistency module.

Fig. 12. Ablation study of non-edited region preservation. The edit prompt is "a head" → "a dog head".

Fig. 13. Examples of editing results with FLUX.

of Q, K, and V achieves the best results in both quantitatively and qualitatively, as it preserves more details.

4.6 Compatibility and Application

4.6.1 Generalization to MM-DiT variants. Our method not only works effectively with SD3 but also generalizes well to other MM-DiT variants such as FLUX.1-dev [Labs 2024]. In Fig. 13, the consistent preservation of fine-grained details and the accurate adaptation of lighting-related reflections further highlight the potential of our approach when applied to more powerful future models.

4.6.2 Generalization to video editing. While our method has already been demonstrated to be agnostic to specific samplers, we further showcase its broad applicability across generation methods (*e.g.*, diffusion models) and domains (*e.g.*, video) by applying it to CogVideoX-2B [Yang et al. 2024], a diffusion-based video generation model. As shown in Fig. 14, our approach enables consistent and controllable editing in both the spatial and temporal domains. Importantly, small inconsistencies that may go unnoticed in static images
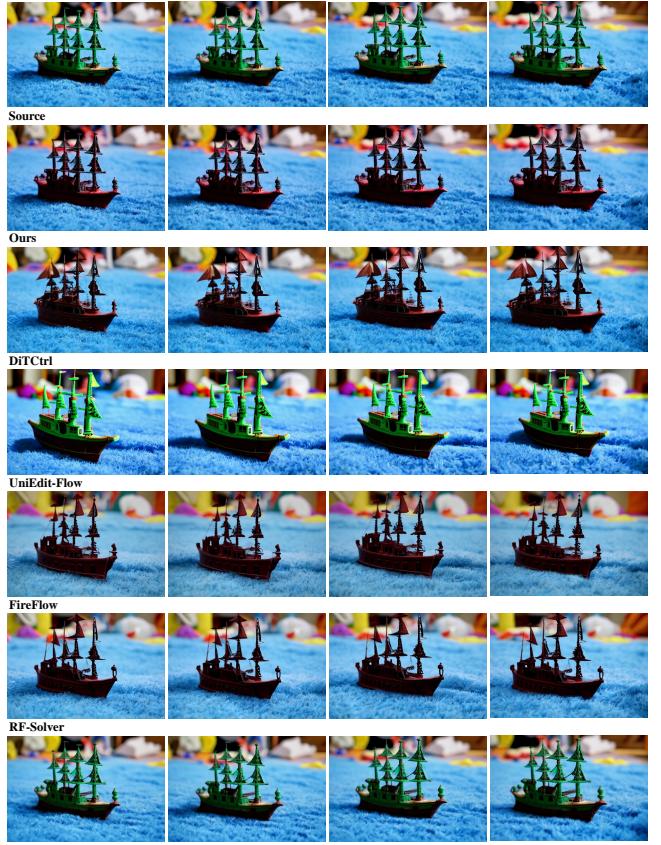

Fig. 14. Qualitative comparison of methods on video editing tasks. The edit prompt is "green toy ship" → "dark red toy ship".

Fig. 15. Examples of applications.

often become amplified and distracting in videos. Our method effectively highlights its robustness and generalizability. Additional results are provided in Appendix B.1.

4.6.3 Application. Fig. 15 showcases our method's versatility across several challenging editing tasks, including recoloring, relighting, animation, shape deformation, and material change. Extending these capabilities to video further amplifies creative possibilities by enabling temporally consistent and detailed edits. The strong editing

power and ease of use highlight the broad potential of our approach for practical and scalable content creation.

5 Conclusion

In this work, we identify key limitations of existing training-free editing methods, including their inability to achieve both strong and consistent text-guided editing, as well as their lack of fine-grained control, where most prior approaches were designed for U-Net or naively applied to MM-DiT without architectural adaptation. To address this, we conduct a detailed analysis of the attention mechanism in MM-DiT and uncover three critical insights that reveal why existing methods fall short. Building on these findings, we propose ConsistEdit, a novel attention control method that operates exclusively on vision tokens. By separating editing and non-editing regions and applying differentiated attention manipulation, ConsistEdit achieves precise, structural consistent edits in edited regions while preserving content in non-edited regions.

Extensive experiments demonstrate that ConsistEdit achieves state-of-the-art performance across diverse image and video editing tasks, without requiring manual tuning. It delivers reliable performance out of the box while offering users fine-grained control over structural consistency. These findings highlight the potential of MM-DiT when paired with our attention control strategies.

Acknowledgments

References

Stability AI. 2024. Stable Diffusion 3.5. https://github.com/Stability-AI/sd3.5. Accessed: May 2025.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. 2025. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7763–7772.

John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*. 22560–22570.

Minghao Chen, Iro Laina, and Andrea Vedaldi. 2024. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 5343–5353.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2023. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*.

Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. 2025. FireFlow: Fast Inversion of Rectified Flow for Image Semantic Editing. In *Forty-second International Conference on Machine Learning*.

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

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. ANIMATEDIFF: ANIMATE YOUR PERSONALIZED TEXT-TO-IMAGE DIFFUSION MODELS WITHOUT SPECIFIC TUNING. In *12th International Conference on Learning Representations, ICLR 2024*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.

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

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. 2024. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12469–12478.

Bernd Jähne. 2005. *Digital image processing*. Springer Science & Business Media.

Guanlong Jiao, Biqing Huang, Kuan-Chieh Wang, and Renjie Liao. 2025. UniEdit-Flow: Unleashing Inversion and Editing in the Era of Flow Models. *arXiv preprint arXiv:2504.13109* (2025).

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. 2024. PnP Inversion: Boosting Diffusion-based Editing with 3 Lines of Code. In *The Twelfth International Conference on Learning Representations*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *CoRR* (2024).

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

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux. Accessed: May 2025.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. 2024a. 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*. 7817–7826.

Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, et al. 2025. Generative video propagation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 17712–17722.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2024b. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8599–8608.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.

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

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15932–15942.

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

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.

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

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. 2025. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations. In *The Thirteenth International Conference on Learning Representations*.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16515–16525.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10748–10757.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. 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*. 1921–1930.

Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. 2023. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17979–17989.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. 2025. Taming Rectified Flow for Inversion and Editing. In

Forty-second International Conference on Machine Learning.

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

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. 2023. Inversion-Free Image Editing with Natural Language. *CoRR* (2023).

Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. 2023. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems* 36 (2023), 26291–26303.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *CoRR* (2024).

Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. 2023. Talking head generation with probabilistic audio-to-visual diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 7645–7655.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 11127–11150.

A Implementation Details

A.1 Inference Settings

We use 28 inference steps for SD3 [Esser et al. 2024], 50 for CogVideoX-2B [Yang et al. 2024], and 20 for FLUX.1-dev [Labs 2024]. The classifier-free guidance (CFG) scale [Ho and Salimans 2022] is set to 7.5 for editing generated sources and 2.0 for real-image editing. All images are generated at 1024×1024, and all videos at 720×480. The inference device is $1 \times$ RTX-4090 GPU.

Our method is compatible with various samplers, with the Euler sampler adopted as default unless otherwise specified. It also supports various inversion techniques; in all experiments, we use the latest inversion method from UniEdit-Flow [Jiao et al. 2025]. We fix the consistency strength to $\alpha = 1$ for tasks requiring structural preservation, and set it to $\alpha = 0.3$ for other tasks.

We adopt a default mask threshold of 0.1, which consistently performs well across our experiments. This relatively coarse masking suffices thanks to the generation models' strong global adaptation, allowing it to propagate edits from partial color cues to semantically aligned regions. The target object for mask extraction is identified either using "blended_word" keywords from PIE-Bench [Ju et al. 2024], or simply by extracting the noun of the object to be edited. Furthermore, our method supports externally provided masks, enabling users to integrate masks generated from other pipelines for more flexible control.

A.2 Sampling Details

To accelerate inference and reduce the number of function evaluations (NFE) during sampling, similar to the approach in Wang et al. [2025], we first run the source prompt branch and cache the Q, K, and V tokens at each step and block for later use. During this stage, we also compute and store the final averaged editing mask. When editing with the target prompt, we load the stored Q, K, and V tokens from the source and apply the editing mask through Mask-Guided Attention Fusion as needed. This strategy ensures that the mask extraction and editing process introduces no additional NFE, maintaining the same efficiency as standard sampling methods.

A.3 Implementation of Compared Methods

Since some compared methods do not provide implementations for SD3 or CogVideoX-2B, or compatible sampling code, we re-implement them within the SD3 and CogVideoX-2B framework by faithfully following their original logic and carefully tuning hyper-parameters to match the reported performance. Implementation details are as follows:

- **DiTCtrl [Cai et al. 2025]**: For image editing, we set the start timestep to 2 and the end timestep to 17, applying edits to the last 5 blocks. For video editing, we use the official implementation. During the editing steps, K and V tokens are copied from the source branch to the target branch in the attention layers. For this method, consistency strength is controlled by increasing the number of end step during which K and V tokens are shared.
- **UniEdit-Flow [Jiao et al. 2025]**: The official implementation is based on SD3, but only provides the parameter ω for CFG = 1. Following the similarity transformation described in the paper, we adopt $\omega = 5 \div 7.5 \approx 0.6$ and set $\alpha = 0.6$, which yields performance

Fig. 16. Examples of video editing.. The prompt is "red SUV" → "blue SUV".

comparable to the original. The same settings are used for video generation. Under this setting, consistency strength is controlled by decreasing the value of α.

- **FireFlow [Deng et al. 2025]**: We observe a significant drop in source–target consistency as CFG increases. Due to performance degradation when the number of edited timesteps increases (as shown in Fig. 4), we limit editing to timesteps from 0 to 3 across all blocks. For video generation, the end timestep is set to 9. During editing, V tokens are copied from source to target. In this approach, consistency strength is modulated by increasing the number of final step in which the V features are shared.
- **RF-Solver [Wang et al. 2025]**: Similar to FireFlow, we set the editing range from timestep 0 to 7 for the latter half of the blocks. During editing, V tokens are copied from the source to the target branch. For video generation, the end timestep is set to 9. The end step of sharing of V tokens serves as the control mechanism for consistency strength in this method.
- **SDEdit [Meng et al. 2022]**: We set $t_0 = 0.6$ and apply editing to either generated source content or real input content, for both image and video generation tasks. For this method, consistency strength is controlled by decreasing the value of t_0.

A.4 Implementation of FLUX

FLUX [Labs 2024] is composed of several double blocks and single blocks. As noted by Wang et al. [2025], single blocks primarily encode general information relevant to generation. Therefore, we apply our editing methods specifically to the single blocks.

B Results and Analysis

B.1 More Results

Additional image editing comparisons are presented in Fig. 21, covering both structure-consistent and structure-inconsistent editing tasks. The results demonstrate that our method achieves superior structural consistency, better preservation of non-edited regions, and enhanced editability compared to existing approaches.

We present additional results on video editing tasks in Fig. 16 and 18. Fig. 16 showcases examples generated by our method alone, while Fig. 18 provides comparisons with existing approaches, demonstrating our superior performance, particularly in scenarios with complex motion.

Fig. 17 presents additional cases of multi-region editing, demonstrating that our method can handle multi-object editing even in the presence of occlusion or complex geometric relationships. Notably, even when multiple regions exhibit intertwined textures, our method accurately identifies the target color for each region and

"green T-shirt … white car" *"yellow dog … purple ball"* *"blue and white stripes"* *"red sweater … black pants"* *"red and blue pill"*
↓ ↓ ↓ ↓ ↓
"blue T-shirt … red car" *"red dog … grey ball"* *"green and yellow stripes"* *"yellow sweater … blue pants"* *"green and yellow pill"*

Fig. 17. Examples of multi-region editing.

"blue paints" *"red golf ball"* *"yellow rose"* *"light blue"* *"blue spoon"*
↓ ↓ ↓ ↓ ↓
"green paints" *"white golf ball"* *"green rose"* *"dark blue"* *"green spoon"*

Fig. 19. Examples of real input image editing. The first row shows the source images, the second row presents the reconstructed images via inversion, and the third row displays the editing results based on the target prompts.

Method	Preference (%)
RF-Solver [Wang et al. 2025]	0.74
SDEdit [Meng et al. 2022]	5.19
FireFlow [Deng et al. 2025]	5.93
UniEdit-Flow [Jiao et al. 2025]	6.67
DiTCtrl [Cai et al. 2025]	10.37
Ours	**71.11**

Table 4. User study preferences over different methods.

B.2 User Study

We conducted a user study involving 18 participants to evaluate editing quality across different methods. Each participant was presented with 30 randomly selected pairs of structure-consistent and structure-inconsistent edits, and was asked to choose the preferred result in each pair. As summarized in Tab. 4, Ours achieved a dominant preference rate of **71.11%**, substantially outperforming all competing approaches.

B.3 Consistency Strength

Fig. 20. Qualitative comparison of different consistency strength settings. "Ours" denotes the method proposed in the main paper, while "Ours*" refers to a modified version of our method.

The main text demonstrates that our method offers fine-grained control over structural alignment with the source image through the consistency strength, while preserving the ability to edit texture
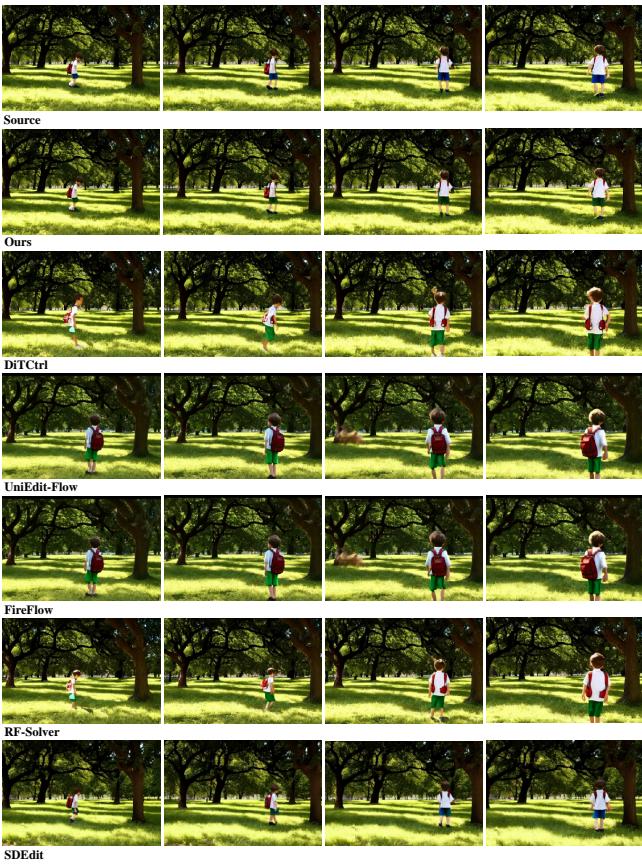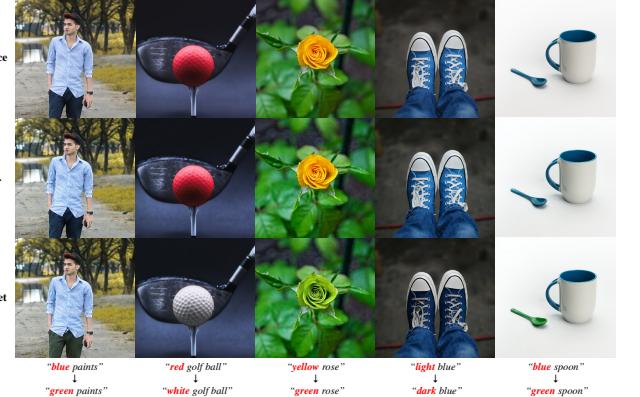

Fig. 18. Additional qualitative comparison of methods on video editing tasks. The edit prompt is "blue shorts" → "green shorts".

performs the corresponding edits. These results highlight the precise text-driven control of our method, fine-grained understanding of visual structure, and strong structure preservation capabilities.

Additional real-image editing examples are shown in Fig. 19. Our method preserves the structural integrity within the edited regions while maintaining the original content in the non-edited regions, achieving performance on par with editing generated images.

Fig. 21. Additional qualitative comparison of methods on structure-consistent and structure-inconsistent editing tasks.

according to the prompt. However, in certain downstream applications, users may prefer a binary behavior in which a consistency strength of 1 results in an output identical to the source image, and a strength of 0 produces results fully aligned with the edited prompt. Although such scenarios are beyond the primary focus of this work, we provide a simple mechanism to enable this behavior, which may serve as a basis for future research in this direction.

To support this behavior, we apply a small modification to our method: within the editing region, in addition to transferring the vision part of Q and K tokens, we also transfer that of V tokens.

As shown in Fig. 20, this simple adjustment successfully achieves the desired behavior between unedited and fully edited results.

B.4 Limitation

The generation quality and the precision of text-guided localization in our method are ultimately constrained by the capabilities of the base generative models. Two representative failure modes are illustrated in Fig. 22:

- **Localization Failure**: Small or abstract objects may not be edited when the attention map lacks a clear activation, leading to no

Fig. 22. Examples of typical failure cases.

visible change. For example, in the top case of Fig. 22, although the overall color, including some very small holes, is edited correctly,

the model struggles to distinguish between intertwined hair and veil.

- **Semantic Ambiguity**: Given a prompt to change lipstick color, the model may instead edit the lipstick case rather than the lipstick itself.

In a different aspect, compared with image models, current video models still lag considerably in generation fidelity. Nevertheless, as foundation models continue to improve, we expect our method to benefit correspondingly and expand its applicability.

Furthermore, our ability to edit real images and videos is inherently constrained by the limitations of current inversion and reconstruction techniques. Although our method performs reliably on data within the distribution of the generative model, editing real-world inputs requires accurately mapping them into the latent space of the model, a task that remains challenging and highly dependent on the quality of the inversion process.