arXiv:2610.17829v2 [cs.CV] 14 Oct 2026

QK-Edit: Revisiting Attention-based Injection in MM-DiT for Image and Video Editing

Tiancheng Shen[1,2,3,4] Zilong Huang[2] Xiangtai Li[2] Zhijie Lin[2] Jiyang Liu[3]
Yitong Wang[3] Jiashi Feng[2] Ming-Hsuan Yang[1] Jun Hao Liew[2]

[1] UC Merced [2] ByteDance Seed [3] ByteDance Intelligent Creation [4] CUHK

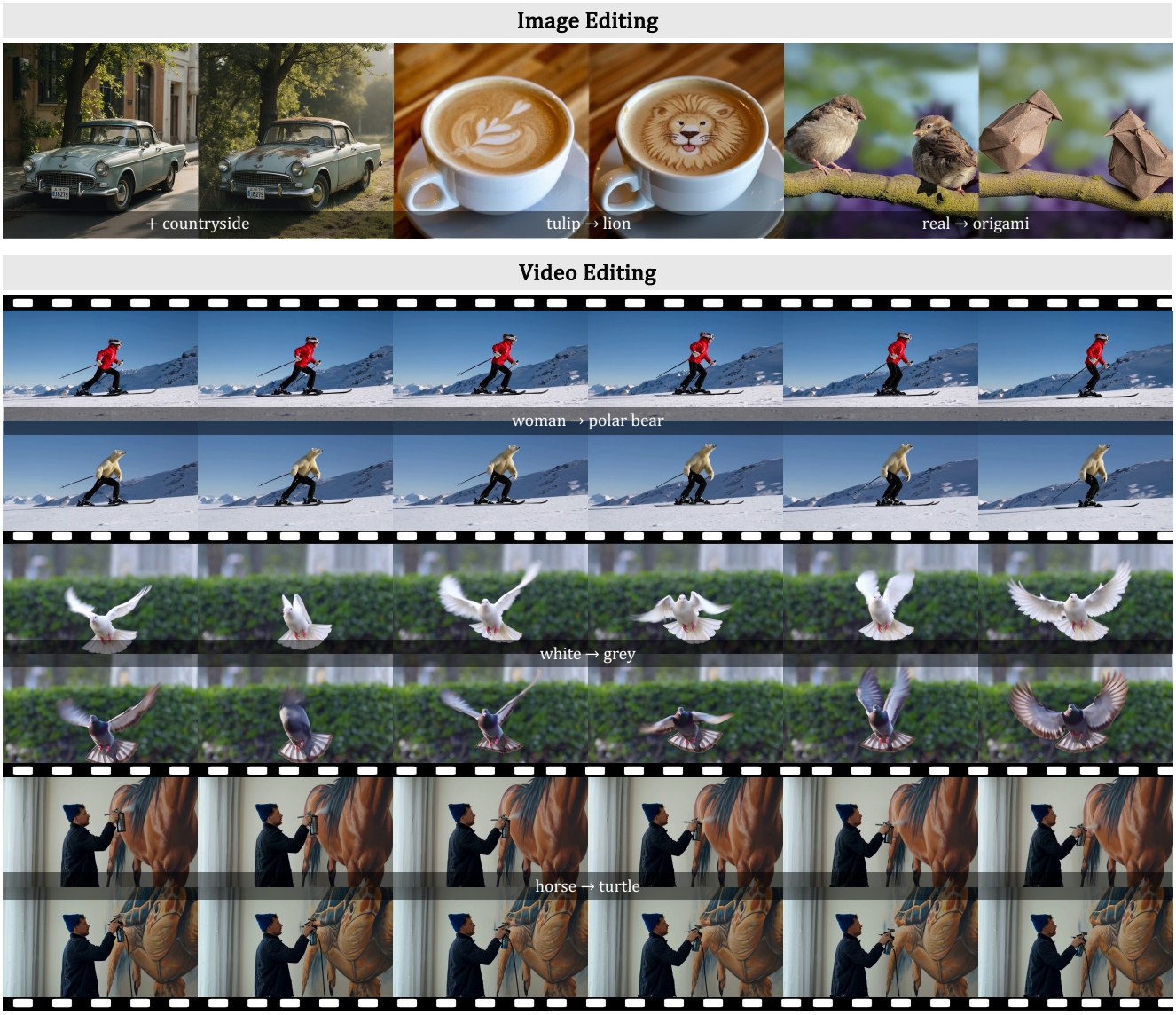


Figure 1. **QK-Edit: A training-free, text-guided framework for high-fidelity image and video editing.** Our method leverages strategic query-key manipulation within the vision components of Multimoda Self-Attention (MM-SA) in MM-DiTs to seamlessly perform diverse edits (*e.g.*, object addition, object removal, object replacement, changing background, changing material, changing color, and style transformation *etc.*), while preserving structural consistency and visual details. Please refer to the supplementary material for more video results at higher resolutions (1K), different aspect ratios (landscape and portrait formats) and longer lengths (129 frames).

Abstract

Multimodal Diffusion Transformers (MM-DiTs) have recently emerged as a powerful framework for unified text-vision synthesis, surpassing traditional U-Net architectures in generative tasks. One key innovation lies in its Multimodal Self-Attention (MM-SA) interaction where image and text tokens are concatenated and processed via self-attention. However, this mechanism poses significant challenges for editing, rendering conventional U-Net-based attention manipulation methods ineffective. To address this limitation, we propose QK-Edit, a training-free framework that exploits the unique attention dynamics of MM-DiTs for precise text-guided image and video editing. By introducing a novel query-key manipulation strategy, our method isolates and adjusts critical attention components to achieve an optimal balance between prompt fidelity and structural consistency. This enables seamless edits across various tasks, including object addition, object removal, object replacement, changing background, changing material, changing color, and style transformation. Notably, it can be easily implemented with feature replacement in inference. QK-Edit demonstrates superior editing performance on state-of-the-art models, such as FLUX and HunyuanVideo, effectively bridging the gap between generative power and editable flexibility in MM-DiTs, and paving the way for scalable multimodal content creation.

1. Introduction

The emergence of Transformer-based diffusion models [50], particularly Multimodal Diffusion Transformers (MM-DiTs)[5, 16, 34, 70], has redefined the landscape of generative models for image and video synthesis. Unlike U-Net architectures[55], which separate text-vision Cross-Attention (CA) and vision-vision Self-Attention (SA), MM-DiTs leverage a Multimodal Self-Attention (MM-SA) mechanism to process multimodal inputs in a unified embedding space. MM-SA concatenates text and vision embeddings and computes self-attention on the joint embedding[1]. This integration enables nuanced and coherent outputs, making MM-DiTs a cornerstone in models such as FLUX [5], Stable Diffusion 3.0 [16], HunyuanVideo [34], and CogVideo-X [70]. However, MM-SA introduces complexities that challenge existing editing paradigms based on attention manipulation [22, 61]. Specifically, due to the joint attention mechanism, textual and visual interactions become deeply intertwined within a single attention matrix, complicating the independent manipulation of textual and visual information. As a result, such editing methods for U-Net models become ineffective in Fig. 2.

[1]In this paper, "vision" refers to video or image data.



Figure 2. Directly applying CA-based and SA-based editing techniques designed for U-Net diffusion models to MM-DiTs can lead to undesirable results.

These complexities arise because such previous editing methods rely heavily on the modular nature of U-Net architectures [54], where cross-attention facilitates precise text-guided modifications, and self-attention ensures spatial consistency. In contrast, MM-DiTs integrate these attention mechanisms into a unified structure, making it challenging to independently control text-vision and vision-vision interactions without one adversely affecting the other. A naive solution is to decompose the MM-SA map into four quadrants, each representing a specific interaction between text and vision embeddings and directly apply previous CA-based or SA-based editing techniques such as P2P [22] or PnP [61] to the respective attention maps. However, this often leads to undesirable outputs, as illustrated in Fig. 2.

Through systematic analysis, we observe: (1) text-to-vision attention has negligible impact on the final generation, (2) vision-to-text attention relates vision region to the text token, and finally (3) vision-vision attention defines both spatial and temporal consistency within visual embeddings. As a result, simultaneous replacement of vision-to-text and vision-to-vision attention is necessary to enable precise control over spatial and temporal properties while adhering to the target prompt. Building upon these insights, we propose QK-Edit, a training-free editing framework that manipulates the attention quadrants of MM-DiTs to achieve high-fidelity edits. In practice, we leverage the fact that text-to-vision attention has minimal impact on the generation and replace the vision component in query and key embeddings of the target image or video with that of the source image or video during the denoising process. In practice, our proposed strategy is straightforward to implement, requiring only minimal code modifications.

Our QK-Edit can be seamlessly extended to image and video domains, demonstrating its versatility in various generative tasks. Extensive experiments validate the effectiveness of QK-Edit on state-of-the-art models such as FLUX [5] and HunyuanVideo [34], achieving superior results in tasks ranging from object-level modifications to global scene transformations. Notably, our method enables high-resolution, long-frame video editing with spatio-temporal consistency, overcoming the limitations of previous editing approaches. Furthermore, the training-free na-

ture of QK-Edit ensures computational efficiency and broad applicability. QK-Edit represents a pivotal step in bridging the gap between the generative power and editable flexibility of Transformer-based diffusion models.

2. Related Work

From U-Net to MM-DiT in Diffusion Models. We revisit diffusion models, which underpin most recent editing methods. The U-Net architecture [51, 55] has been extensively studied in generative models, as seen in Lumiere [4] for text-to-video synthesis and Stable Diffusion [54] for text-to-image generation. Recently, Transformer-based diffusion models have gained attention for capturing long-range dependencies and multimodal interactions. For instance, Peebles and Xie [50] introduce the Diffusion Transformer (DiT), replacing U-Net with a Vision Transformer (ViT) for improved results, while Chen et al.[13] propose GenTron to overcome CNN-based U-Net limitations. Multimodal extensions, such as MM-DiT [5, 16], integrate text and vision embeddings to enable sophisticated multi-prompt generation tasks [10]. For video generative models, following Sora [48], several works [34, 40, 60] have open-sourced video diffusion models like Hunyuan Video [34] and CogVideoX [70]. These advances illustrate an evolution from U-Net to Transformer-based frameworks, offering enhanced scalability, flexibility, and performance. Our training-free technique offers a fast and low-cost solution to unlock the inherent editing potential of these MM-DiT-based new-generation models.

Diffusion-based Image Editing. Recent image editing advances fall into training-free and training-based approaches. Training-free methods use pre-trained generative models for direct image manipulation, avoiding costly training [2, 8, 21, 22, 26, 29, 30, 37, 41, 44, 59, 62, 69]. They mainly include inversion-based and direct manipulation-based techniques. Inversion-based methods, such as Null-Text Inversion [45], StyleDiffusion [65] and KV Inversion [24] encode images into a generative model's latent space, where edits adjust latent representations. Other methods, such as pix2pix-zero [49] and SEGA [7], modify latent features or noise maps using attention and blending, enabling precise and efficient editing. In contrast, training-based approaches [12, 19, 25, 28, 32, 36, 38, 47, 53, 63, 67, 72, 73] enhance editing capabilities by fine-tuning generative models or training auxiliary networks on task-specific datasets [9, 27, 72]. These approaches generally offer stronger semantic fidelity and improved visual quality at the cost of additional computational overhead. Methods like InstructPix2Pix [9] and MagicBrush [72] employ instruction-based editing by fine-tuning generative models to directly map textual instructions and input images to edited outputs. Our QK-Edit, built on MM-DiT architectures, requires no extra training while efficiently unlocking new-generation models' editing potential for high prompt fidelity and structural consistency.

Diffusion-based Video Editing. Diffusion-based Video Editing. Training-free methods [3, 29, 43, 52, 57, 64] leverage pre-trained diffusion models and latent manipulation to avoid extensive training. For example, Video-P2P [43] and Vid2Vid-Zero [64] extend the image-based P2P [22] framework to video by injecting attention maps and using latent inversion for temporal consistency, while Fate-Zero [52] enhances spatial coherence by selectively injecting SA and CA maps from the reconstruction branch. Similarly, Edit-A-Video [57] and Make-A-Protagonist [75] refine latent manipulations with explicit semantic masks or advanced attention mechanisms, improving temporal stability without extra training. In contrast, training-based approaches [6, 14, 18, 42, 58, 76] fine-tune models specifically for video data [1, 14, 18, 20, 23], significantly boosting editing performance at the cost of increased computation. Tune-A-Video [66] fine-tunes spatial and temporal attention layers for consistency, while MotionDirector [74] and Dreamix [46] employ decoupled motion and appearance fine-tuning to reduce artifacts. ControlNet-based methods, such as MagicEdit [39] and Ground-A-Video [71], use depth maps and bounding boxes to guide edits for improved semantic and temporal coherence, and InsV2 [14] leverages motion compensation with instruction-based fine-tuning to handle longer videos with enhanced continuity. Our QK-Edit can easily generalize to video editing for MM-DiT-based video generative models [34, 70].

3. Method

We aim to design a general method to edit images and videos based on text prompts for MM-DiTs models [5, 34, 35]. Given a source image I^{src} or video V^{src} to edit, a source prompt P^{src} describing the source vision content, and a target prompt P^{tgt} describing the desired editing target, our goal is to generate a modified image I^{tgt} or video V^{tgt} that exhibits the desired changes specified by P^{tgt} while preserving the unedited region. To achieve it, our QK-Edit leverages the vision part of the query and key embeddings in the Multimodal Diffusion Transformer (MM-DiT) [17]. For synthesized images and videos, the query and key embeddings are preserved throughout the synthesis process, whereas for real images and videos, these embeddings are derived from the inversion process.

3.1. Preliminaries

U-Net-based diffusion models [4, 54] employ an encoder-decoder framework with skip connections, where text-vision cross-attention and vision-vision self-attention are applied sequentially. In particular, self-attention in Eq. 1 facilitates the pixel-wise dependencies within the image, en-

suring the ability to control spatial relationships:

$$\text{Self-Attention} = \text{Softmax}\Big(\frac{Q_{\text{vision}}K_{\text{vision}}^{\top}}{\sqrt{d}}\Big)V_{\text{vision}}. \quad (1)$$

On the other hand, cross-attention (Eq. 2) enables the incorporation of external guidance, such as textual prompts:

$$\text{Cross-Attention} = \text{Softmax}\Big(\frac{Q_{\text{vision}}K_{\text{text}}^{\top}}{\sqrt{d}}\Big)V_{\text{text}}. \quad (2)$$

Existing editing methods typically make use of the modular nature of U-Net architecture for editing. For example, P2P [22] edits images by replacing cross-attention maps, allowing for semantic changes guided by new textual prompts while preserving the original layout. PnP [61] injects the features and self-attention maps at specific denoising steps with those from a reference image to encourage the generation of images that closely resemble the reference image.

Multimodmal Diffusion Transformers (MM-DiTs), on the contrary, as implemented in models such as FLUX [5] and HunyuanVideo [34], adopt a text-vision Multimodal Self-Attention (MM-SA) mechanism. In this setup, text and visual inputs are concatenated into a unified embedding space, and a single self-attention operation is performed over this combined representation. Consequently, text-vision cross-attention and vision-vision self-attention are integrated within the same attention map, influencing each other through the softmax operation. Mathematically, the MM-SA operation is formulated as

$$\text{MM-SA} = \text{Softmax}\Big(\frac{QK^{\top}}{\sqrt{d}}\Big)V, \quad (3)$$

where Q, K and V are the concatenated query, key and value matrices of text and vision embeddings, respectively. $[\cdot]$ denotes concatenation operation.

$$Q = [Q_{\text{text}}, Q_{\text{vision}}],\ K = [K_{\text{text}}, K_{\text{vision}}],\ V = [V_{\text{text}}, V_{\text{vision}}], \quad (4)$$

This integrated approach enables more nuanced interactions between textual and visual modalities, enhancing the model capability to generate coherent and contextually relevant outputs. However, it also introduces complexity, as the intertwined attention mechanisms can lead to intricate dependencies that affect the interpretability and controllability of the model behavior.

3.2. Analysis of MM-DiT

Taking FLUX [5] as an example, if we directly apply the cross-attention and self-attention manipulation techniques, the results are not as satisfactory as in U-Net diffusion [22, 61], as shown in Fig.2. Understanding these distinctions is crucial for developing methods that leverage the strengths of each architecture, particularly in tasks requiring precise control over the generated content.

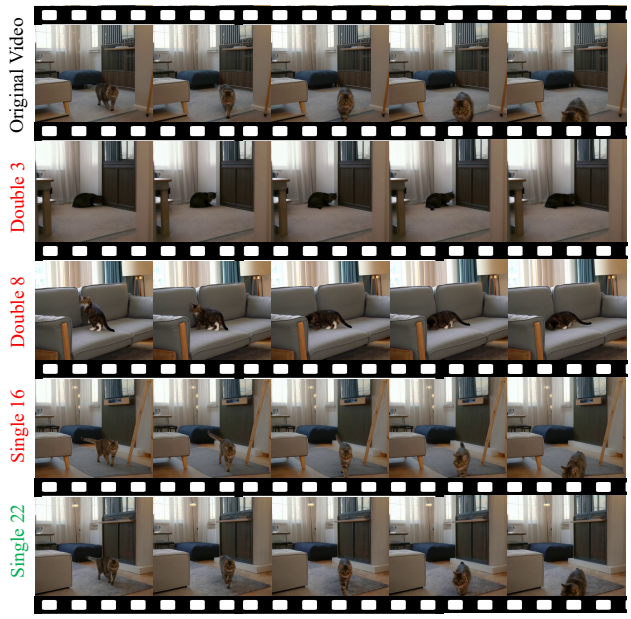


Figure 3. Removing individual blocks in the MM-DiT architecture affects different aspects of the generated video.

Next, we conduct a progressive analysis of the model, from the main modules to the finer details, to fully understand the design and functionality of MM-DiT. In the following experiments, we take FLUX [5] and HunyuanVideo [34] as representative image and video generative models, respectively.

Principal and non-principal blocks. We first examine the role of each MM-DiT block on HunyuanVideo [34]. Given the same prompt, we skip one block at a time and observed that skipping certain blocks significantly affects the generated videos, while others have different or minimal impacts. For example, as shown in Fig. 3, skipping the 22^{nd} single block does not lead to visible change, whereas bypassing other blocks causes significant changes. This phenomenon is also consistent with the observations in Stable Flow [2]. Next, we investigate these blocks to better understand the underlying mechanism.

MM-SA decides the final spatial arrangement. Each MM-DiT block comprises a MM-SA and a Feed-forward Network (FFN), connected through residual connections. As before, we skip one component at a time to inspect its impact on the final generation. As shown in Fig. 4, MM-SA decides the final spatial arrangement, while skipping FFN has negligible effect. This inspires us to focus on the MM-SA for potential editing. Next, we analyze each attention quadrant of the MM-SA attention map.

Attention quadrants. The attention map in MM-SA can be

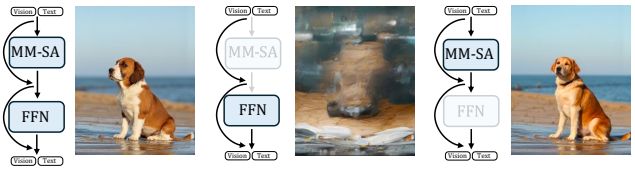


Figure 4. We study the role of the Multimodal Self-Attention (MM-SA) and Feed-Forward Network (FFN) components in MM-DiT blocks by bypassing one at the time. The results underscore that MM-SA predominantly determines spatial and temporal consistency, while FFN has a negligible impact.

decomposed into 4 quadrants as shown in Fig. 5(a). Based on this, we conduct two pilot studies: in Fig. 5(b), (c), (d), and (e), we analyze the effect of each attention quadrant by setting its values to 0; and in Fig. 5(f), (g), and (h), we visualize the attention map of the 1st and 3rd quadrants and the PCA feature of the 4th quadrant.

(i) 1st Quadrant: $Q_{\text{text}}K_{\text{text}}^{\top}$ controls the relationship among the text tokens of prompt $P^j, \forall j \in \{src, tgt\}$ as shown in Fig. 5(f). It is not symmetrical. And the attention value in this quadrant is usually $\gg$ the other 3 vision-related counterparts.

(ii) 2nd Quadrant: $Q_{\text{text}}K_{\text{vision}}^{\top}$ is not as important as the others, although both the 2nd and 3rd Quadrant represent the text-vision relationship in theory. When removing this quadrant (*i.e.* setting this quadrant to 0), there is a negligible effect in the generated result (Fig. 5(c)).

(iii) 3rd Quadrant: $Q_{\text{vision}}K_{\text{text}}^{\top}$ is responsible for deciding the text-vision relationship. Different from the 2nd Quadrant, FLUX fails to generate a proper image when setting the 3rd Quadrant to 0 (see Fig. 5(d)). Every column is related to the region corresponding to a certain text token as shown in Fig. 5(g).

(iv) 4th Quadrant: $Q_{\text{vision}}K_{\text{vision}}^{\top}$ is responsible for deciding the spatial layout since it defines the pixel affinity (similarity between pixel i and j) (Fig. 5(h).

3.3. Editing using Multimodal Self-Attention

In the previous section, we empirically found that the MM-SA determines the spatial arrangement (spatial layout in image, motion in video). Therefore, to achieve desirable editing, we need to "fix" the elements in MM-SA to preserve the unchanged elements in the source video V^{src} and modify the region to change following the target prompt P^{tgt}.

Based on these observations, we show that text-guided editing can be performed by manipulating **3rd Quadrant** and **4th Quadrant**, *i.e.* $Q_{\text{vision}}K_{\text{text}}^{\top}$ and $Q_{\text{vision}}K_{\text{vision}}^{\top}$. Specifically, we begin with a source prompt P^{src} and a source video V^{src} to edit. In the process of generating source video from source prompt or inversing real video to initial noise, we record Q_{vision}^{src} and $K_{\text{vision}}^{src\top}$ of selected time steps. To edit the source video according to the target prompt, we manipulate the 3rd and 4th Quadrant during the denoising process by:

$$Q_{\text{vision}}^{tgt}K_{\text{text}}^{tgt\top} \leftarrow Q_{\text{vision}}^{src}K_{\text{text}}^{tgt\top}, \tag{5}$$

$$Q_{\text{vision}}^{tgt}K_{\text{vision}}^{tgt\top} \leftarrow Q_{vision}^{src}K_{\text{vision}}^{src\top}. \tag{6}$$

However, in practice, attention operation is implemented in a highly optimized and integrated manner, where attention maps are often encapsulated within a package (*e.g.*, `flash-attention`[2]). As a result, it is challenging to directly modify the attention maps.

Building on our previous observation that 2nd Quadrant has little impact on the final output, we can circumvent the constraints imposed by this encapsulation. Consequently, manipulating 3rd and 4th Quadrant becomes equivalent to substituting the vision component of Q^{tgt} and K^{tgt}, i.e.

$$[Q_{\text{text}}^{tgt}, Q_{\text{vision}}^{tgt}] \leftarrow [Q_{\text{text}}^{tgt}, Q_{\text{vision}}^{src}], \tag{7}$$

$$[K_{\text{text}}^{tgt}, K_{\text{vision}}^{tgt}] \leftarrow [K_{\text{text}}^{tgt}, K_{\text{vision}}^{src}]. \tag{8}$$

This substitution is illustrated in Fig. 6. Finally, the manipulated QKV is used to generate the target image/ video.

Easy Implementation. QK-Edit is straightforward to implement. It involves only two steps: (1) during real video inversion or generative video generation, save the Q_{vision}^{src} and K_{vision}^{src}; and (2) during target video generation, use the saved ones to replace the Q_{vision}^{tgt} and K_{vision}^{tgt}.

4. Experiments

4.1. Implementation details

We implement QK-Edit on state-of-the-art MM-DiT-based generative models. Specifically, we use FLUX.1-dev [5] and HunyuanVideo [34] as base model for image and video editing, respectively. For image editing, we apply our query-key manipulation during the first 30% diffusion inference steps. For video editing, resolutions vary between 320×320 (with 65 frames) and 1280×720 (up to 129 frames). We apply query-key manipulation during the first 20% of inference steps, which empirically provides optimal temporal consistency and editing fidelity. All experiments are conducted on NVIDIA A100 GPUs.

4.2. Quantitative Evaluation Metrics

To rigorously assess editing performance, we utilize the following metrics following previous works [18, 33, 58, 68].

Specifically, for image editing, we employ **CLIP$_{\text{text}}$** (measures semantic alignment between the edited image and the text prompt), **DINO** (perceptual similarity for structure-alignment) and **PickScore** (assesses overall editing quality and realism in a human-like manner).

[2] https://github.com/Dao-AILab/flash-attention

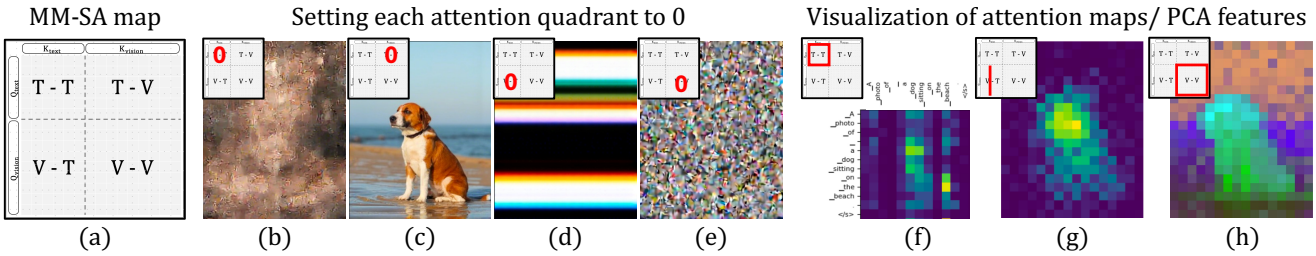


Figure 5. **Pilot experiments.** (left) MM-SA attention map. (middle) prompting the influence of each attention quadrant by setting it to **0**. (right) visualization of attention map of the 1st and 3rd quadrants and PCA feature of the 4th quadrant.

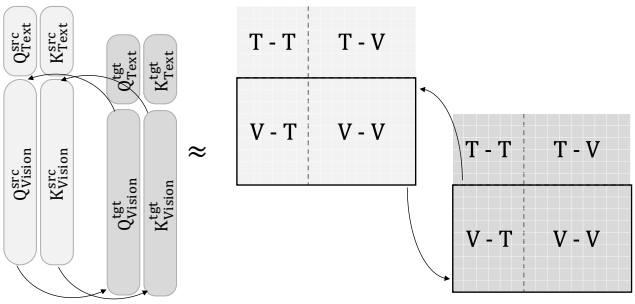


Figure 6. **Query-key manipulation strategy.** By substituting the vision components in Q and K, the approach effectively guides the editing process to balance prompt fidelity with preservation of the source structure.

Method	CLIP-T↑	DINO↑	PickScore↑
LEDITS++ [8]	0.207	0.571	19.88
MasaCtrl [11]	0.235	0.696	20.68
RF-Inversion [56]	0.252	0.523	21.31
RF-Solver [62]	0.257	0.769	21.98
QK-Edit	**0.264**	**0.776**	**22.33**

Table 1. **Real image editing evaluation on PIE-Bench**.

On the other hand, to evaluate video editing performance, we use $\textbf{CLIP}_{\textbf{frames}}$ (computes frame consistency through the average cosine similarity across all video frames), **PickScore** (assesses the predicted human preference averaged over all video frames), $\textbf{ViCLIP}_{\textbf{dir}}$(measures directional video-text alignment) and $\textbf{ViCLIP}_{\textbf{out}}$ (evaluates overall video-text alignment).

4.3. Quantitative Results

Real image editing. For image editing, we compare our QK-Edit against LEDITS++ [8], MasaCtrl [11], RF-Inversion [56], and RF-Solver [62] on the PIE-Bench [31] (Tab. 1). As shown in Tab. 1, QK-Edit achieves the highest scores in all metrics, indicating its superiority in aligning with the text prompt while preserving structural details.

Video editing. For video editing, we evaluate on the TGVE+ benchmark [58] with two settings. In the first setting, we evaluate on the generative video editing task (without video inversion. Instead, source video is generated as the first step). We compare our QK-Edit with DiTCtrl [10]. Both methods are based on HunyuanVideo [34] for fair comparison. As shown in Tab. 2, QK-Edit outperforms DiTCtrl across all metrics, demonstrating better frame consistency and stronger video-text alignment.

Method	$\text{CLIP}_{\texttt{frames}}$↑	PickScore↑	$\text{ViCLIP}_{\texttt{dir}}$↑	$\text{ViCLIP}_{\texttt{out}}$↑
DiTCtrl [10]	0.904	20.01	0.141	0.242
QK-Edit	**0.951**	**21.24**	**0.217**	**0.268**

Table 2. **Generative video editing performance on TGVE+ dataset.**

Method	$\text{ViCLIP}_{\texttt{dir}}$↑	$\text{ViCLIP}_{\texttt{out}}$↑	$\text{CLIP}_{\texttt{frames}}$↑	PickScore↑
CCEdit [18]	0.159	0.244	0.877	19.73
Flatten [15]	0.204	0.236	0.887	19.91
AnyV2V [35]	0.117	0.228	0.906	20.28
QK-Edit	**0.196**	**0.258**	**0.921**	**20.52**

Table 3. **Real video editing performance on TGVE+ dataset.**

In the second setting, we evaluate the performance of real video editing (requires video inversion as the first step). We evaluate QK-Edit against CCEdit [18], Flatten [15], and AnyV2V [35]. As shown in Tab. 3, QK-Edit outperforms all the baselines, validating that QK-Edit not only produces visually coherent edits but also maintains semantic consistency across video frames.

4.4. Qualitative Comparisons

In addition to the quantitative evaluation, we also present some qualitative comparisons. As shown in Fig. 7, we can see that QK-Edit not only preserves fine-grained details and structures, but also is capable to integrate target textual cues into the edited image. On the other hand, for video editing, QK-Edit maintains excellent frame-to-frame consistency and temporal continuity. These visual examples underscore the method's ability to deliver high-fidelity edits while remaining temporally coherent.

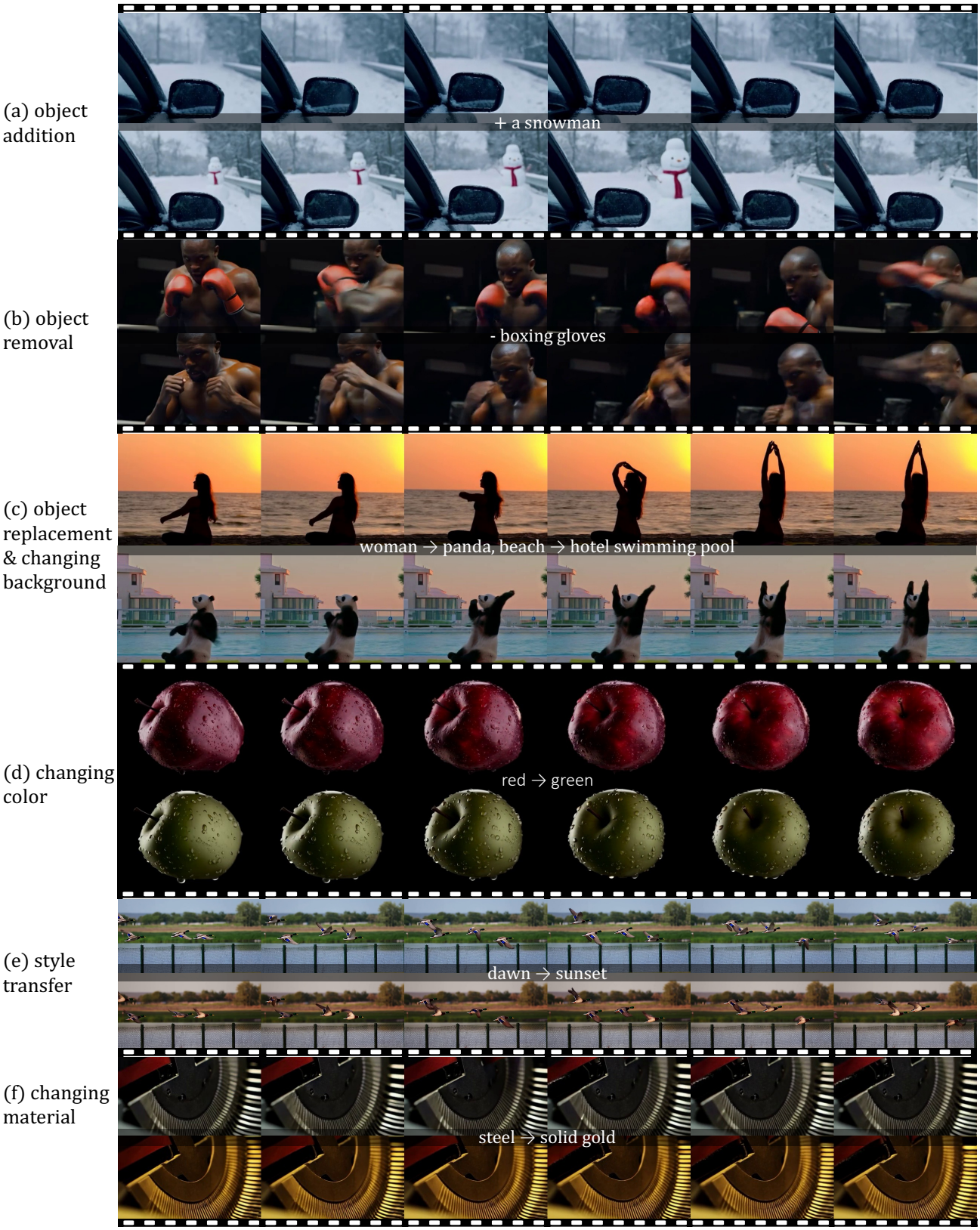


Figure 7. Visualizations showcasing QK-Edit's capabilities in various video editing tasks.

Method	KV_{concat}	$V_{replace}$	$KV_{replace}$	$QK_{replace}$
$CLIP_{frames}$↑	0.896	0.900	0.901	**0.908**
PickScore↑	19.87	19.94	19.92	**20.08**

Table 4. **Quantitative results on HiLo-Benchmark**. Our proposed QK-Edit via $QK_{replace}$ method (last column) works the best.

Besides, we also found that QK-Edit excels in multiple editing scenario, including object addition (*e.g.*, add a snowman in Fig. 7(a)), object removal (*e.g.*, remove the boxing gloves in Fig. 7(b)), object replacement (*e.g.*, woman to panda in Fig. 7(c)), changing background (*e.g.*, beach to hotel swimming pool in Fig. 7(c)), changing color (*e.g.*, red to green in Fig. 7(d)), style transformation (*e.g.*, from dawn to sunset in Fig. 7(e)), and changing material (*e.g.*, from steel to solid gold in Fig. 7(f)). These examples demonstrate the flexibility of QK-Edit across a wide range of editing tasks.

4.5. Ablation Study

We perform extensive ablation studies to validate our design choices and more challenging scenario.

Manipulation on QKV combinations. We perform an ablation study to understand the effect of utilizing different combinations of Q_{vision}, K_{vision} and V_{vision} in Fig 8. We observe that (1) replacing QK works the best in preserving the source video structure while following the target prompt, (2) replacing KV or V preserve source video structure better, but sometimes fails to follow the target prompt. (3) replacing Q or K can better follow the prompt, but often fails to preserve the source video structure.

High-fidelity video editing. Recognizing the growing demand for high-fidelity video editing, we introduce a new benchmark, HiLo-Benchmark, consisting of 43 carefully selected challenging videos with resolutions exceeding 1K and sequences longer than 100 frames. This benchmark specifically evaluates the ability of video editing methods to handle extended and high-resolution content. For baselines, considering compatibility with high-resolution and long-duration video generation, we adapt ideas from previous image/video synthesis frameworks and reimplement them based on HunyuanVideo [34]. Our proposed benchmark includes comparisons with methods such as KV_{concat} (MasaCtrl [11] and Tune-a-video [66]), $V_{replace}$ (RF-Solver [62]) and $KV_{replace}$ (Stable Flow [2]). Tab. 4 demonstrates that QK-Edit consistently achieves higher editing fidelity, structural consistency, and semantic alignment across both images and videos compared to existing approaches.

5. Conclusion

In this work, we present QK-Edit, a simple yet effective training-free framework that exploits the unique attention dynamics of MM-DiTs for precise text-guided image and video editing. Extensive experiments demonstrate the versatility of QK-Edit across a wide range of editing tasks, *e.g.*, object addition, object removal, object replacement, changing background, changing material, changing color, and style transformation *etc*. More importantly, it can be easily implemented with a few lines of code.

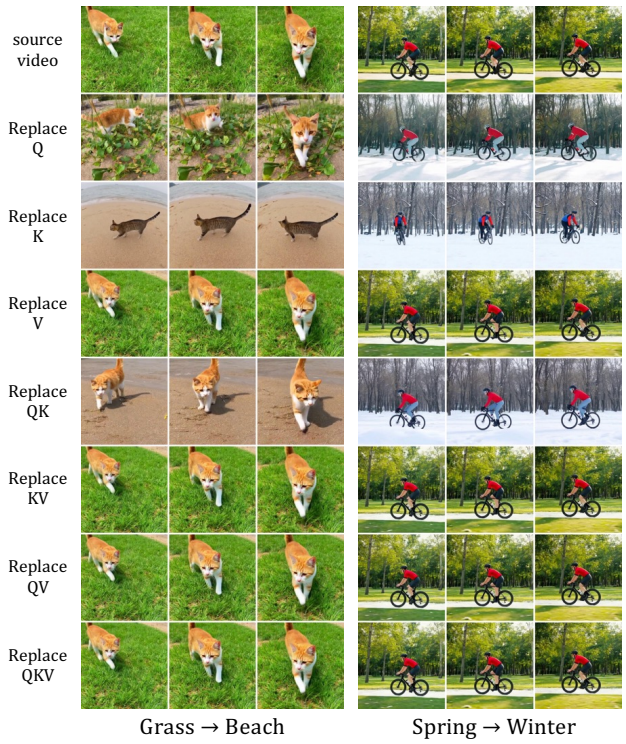


Figure 8. **Manipulation of different QKV combinations.** Replacing KV or V preserves original structure better, but sometimes fails to follow the target prompt; replacing Q or K, on the hand, adheres to the editing prompt but often fail to preserve source video structure; replacing QK achieves the best trade-off.

Limitation and future work. While QK-Edit significantly advances training-free text-guided editing for MM-DiTs, it still exhibits several limitations. First, the accuracy of real video inversion remains challenging, particularly in complex scenes with intricate motions, which can compromise subsequent editing results. Second, QK-Edit inevitably introduces minor unintended modifications in regions that should remain unaltered; future work could incorporate masks to better preserve these areas. Third, as a training-free approach, QK-Edit inherits the limitations of its base generative model, such as suboptimal text rendering and structural deficiencies in handling complex motions (e.g., gymnastic videos). These issues might be alleviated by employing more advanced generative models, which we leave as a promising direction for future research.

References

[1] Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing. In *ECCV*, 2022.

[2] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. *arXiv preprint arXiv:2411.14430*, 2024.

[3] Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. Uniedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185*, 2024.

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia*, 2024.

[5] Black Forest Labs. The announcements page of black forest labs website. https://blackforestlabs.ai/announcing-black-forest-labs/, 2023. Accessed: 2024-4.

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[7] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *NeurIPS*, 2023.

[8] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *CVPR*, 2024.

[9] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[10] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. *arXiv preprint arXiv:2412.18597*, 2024.

[11] 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 *ICCV*, 2023.

[12] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *CVPR*, 2024.

[13] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *CVPR*, 2024.

[14] Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. In *ICLR*, 2024.

[15] Yuren Cong, Mengmeng Xu, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, Sen He, et al. Flatten: optical flow-guided attention for consistent text-to-video editing. In *The Twelfth International Conference on Learning Representations*, 2024.

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[17] 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 *ICML*, 2024.

[18] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. In *CVPR*, 2024.

[19] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *CVPR*, 2024.

[20] Xin Gu, Libo Zhang, Fan Chen, Longyin Wen, Yufei Wang, Tiejian Luo, and Sijie Zhu. Edit3k: Universal representation learning for video editing components. *arXiv preprint arXiv:2403.16048*, 2024.

[21] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *CVPR*, 2024.

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

[23] Jiahao Hu, Tianxiong Zhong, Xuebo Wang, Boyuan Jiang, Xingye Tian, Fei Yang, Pengfei Wan, and Di Zhang. Vivid-10m: A dataset and baseline for versatile and interactive video local editing. *arXiv preprint arXiv:2411.15260*, 2024.

[24] Jiancheng Huang, Yifan Liu, Jin Qin, and Shifeng Chen. Kv inversion: Kv embeddings learning for text-conditioned real image action editing. In *PRCV*, 2023.

[25] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*, 2024.

[26] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *CVPR*, 2024.

[27] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

[28] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *ICLR*, 2024.

[29] Yueru Jia, Yuhui Yuan, Aosong Cheng, Chuke Wang, Ji Li, Huizhu Jia, and Shanghang Zhang. Designedit: Multi-

layered latent decomposition and fusion for unified & accurate image editing. *arXiv preprint arXiv:2403.14487*, 2024.

[30] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *ICLR*, 2024.

[31] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *ICLR*, 2024.

[32] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023.

[33] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *NeurIPS*, 2023.

[34] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[35] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*, 2024.

[36] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *NeurIPS*, 2023.

[37] Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu, Jianzhuang Liu, et al. Zone: Zero-shot instruction-guided local editing. In *CVPR*, 2024.

[38] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024.

[39] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023.

[40] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[41] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *CVPR*, 2024.

[42] Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. Stablev2v: Stablizing shape consistency in video-to-video editing. *arXiv preprint arXiv:2411.11045*, 2024.

[43] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *CVPR*, 2024.

[44] Zichong Meng, Changdi Yang, Jun Liu, Hao Tang, Pu Zhao, and Yanzhi Wang. Instructgie: Towards generalizable image editing. In *ECCV*, 2025.

[45] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023.

[46] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

[47] Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Dreamcreature: Crafting photorealistic virtual creatures from imagination. *arXiv preprint arXiv:2311.15477*, 2023.

[48] Open-AI. Video generation models as world simulators. 2024.

[49] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH 2023*, 2023.

[50] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[51] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[52] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, 2023.

[53] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: a unified diffusion model for controllable visual generation in the wild. In *NeurIPS*, 2023.

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[56] 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.

[57] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. In *ACML*, 2024.

[58] Uriel Singer, Amit Zohar, Yuval Kirstain, Shelly Sheynin, Adam Polyak, Devi Parikh, and Yaniv Taigman. Video editing via factorized diffusion distillation. In *ECCV*, 2024.

[59] Xue Song, Jiequan Cui, Hanwang Zhang, Jingjing Chen, Richang Hong, and Yu-Gang Jiang. Doubly abductive counterfactual inference for text-based image editing. In *CVPR*, 2024.

[60] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.

[61] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.

[62] 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.

[63] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[64] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023.

[65] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *ICCV*, 2023.

[66] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023.

[67] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 2023.

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

[69] Zhen Yang, Ganggui Ding, Wen Wang, Hao Chen, Bohan Zhuang, and Chunhua Shen. Object-aware inversion and reassembly for image editing. *arXiv preprint arXiv:2310.12149*, 2023.

[70] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[71] Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. In *ICLR*, 2024.

[72] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS*, 2023.

[73] 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 *CVPR*, 2024.

[74] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *ECCV*, 2024.

[75] Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850*, 2023.

[76] Zhichao Zuo, Zhao Zhang, Yan Luo, Yang Zhao, Haijun Zhang, Yi Yang, and Meng Wang. Cut-and-paste: Subject-driven video editing with attention control. *Neural Networks*, 2025.

QK-Edit: Revisiting Attention-based Injection in MM-DiT for Image and Video Editing

Supplementary Material

Algorithm 1: QK-Edit

Input: 1. T_list: timesteps to replace the Q_{Vision} and K_{Vision}
2. QK_list: record $Q^{src\cdot t}_{Vision}$ and $K^{src\cdot t}_{Vision}$ of every timestep t
Output: an edited latent embedding $z^{tgt\cdot 0}$

Part 1: Inversion to record QK from source
for t = 1, ..., T-1, T **do**
$Q^{src\cdot t-1} = [Q^{src\cdot t-1}_{Text}, Q^{src\cdot t-1}_{Vision}]$
$K^{src\cdot t-1} = [K^{src\cdot t-1}_{Text}, K^{src\cdot t-1}_{Vision}]$
$V^{src\cdot t-1} = [V^{src\cdot t-1}_{Text}, V^{src\cdot t-1}_{Vision}]$
$[Q^{src\cdot t}_{Vision}, K^{src\cdot t}_{Vision}],\ z^{src\cdot t} = \text{Inversion}([Q^{src\cdot t-1}, K^{src\cdot t-1}, V^{src\cdot t-1}], z^{src\cdot t-1}, \text{t-1})$
QK_list.append($[Q^{src\cdot t}_{Vision}, K^{src\cdot t}_{Vision}]$)
end

Part 2: Performing editing with QK replacement
for t = T, T-1, ..., 1 **do**
if t in T_list **do**
$[Q^{src\cdot t}_{Vision}, K^{src\cdot t}_{Vision}]$ = QK_list[t]
$Q^{tgt\cdot t} = [Q^{tgt\cdot t}_{Text}, Q^{src\cdot t}_{Vision}]$
$K^{tgt\cdot t} = [K^{tgt\cdot t}_{Text}, K^{src\cdot t}_{Vision}]$
$V^{tgt\cdot t} = [V^{tgt\cdot t}_{Text}, V^{tgt\cdot t}_{Vision}]$
$z^{tgt\cdot t} = \text{Forward}([Q^{tgt\cdot t}, K^{tgt\cdot t}, V^{tgt\cdot t}], z^{tgt\cdot t-1}, t)$
end

Return $z^{tgt\cdot 0}$

Figure 9. The pseudo code of QK-Edit.

6. Pseudo Code

To illustrate the simplicity of our QK-Edit, we provide the pytorch-style pseudo code for our algorithm in Fig. 9. This pseudo code demonstrates the straightforward implementation of QK-Edit.

7. Additional Image Editing Results

Additional qualitative results for image editing are provided in Fig. 11. These examples further demonstrate QK-Edit's ability to preserve fine-grained details, maintain structural consistency, and achieve precise semantic alignment across diverse editing scenarios.

8. Video Editing Results for High-Resolution

Please refer to the accompanying MP4 files for video editing results at higher resolutions, longer frames, and two different aspect ratios. Specifically, the two files contain results for videos with dimensions of 1280×720 with 129 frames (`1280_720_129.mp4`) and 720×1280 with 65 frames (`720_1280_65.mp4`), showcasing the flexibility of our method on both landscape and portrait formats.

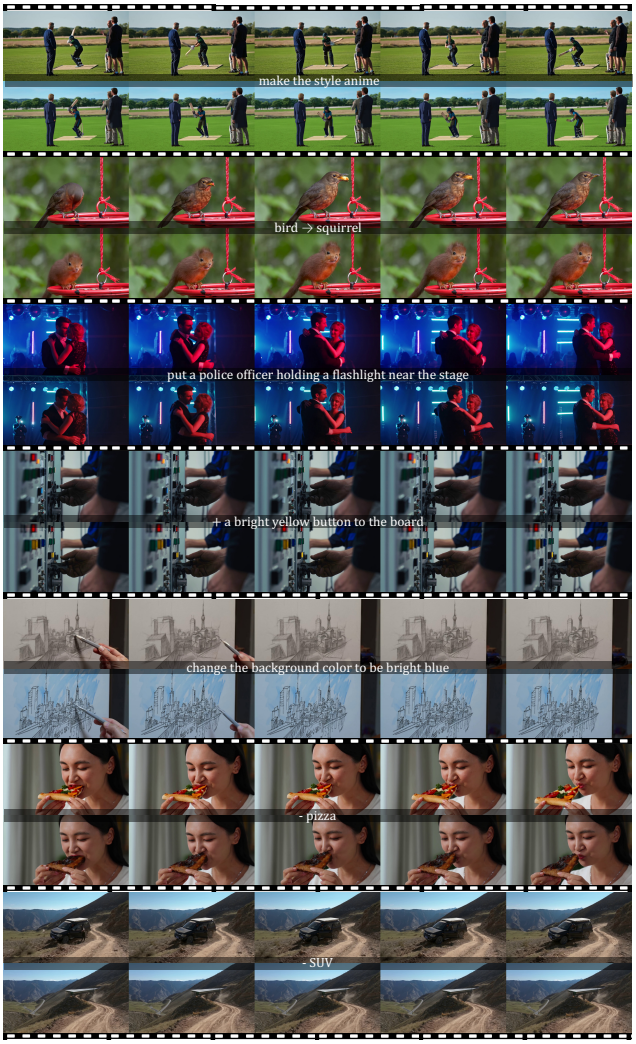


Figure 10. Failure cases and artifacts.

9. Failure Cases and Artifacts

As a training-free method, our performance is inherently bounded by the capabilities of the base model. We have discussed some failure cases in the Limitation section. More failure cases are illustrated in Fig. 10.

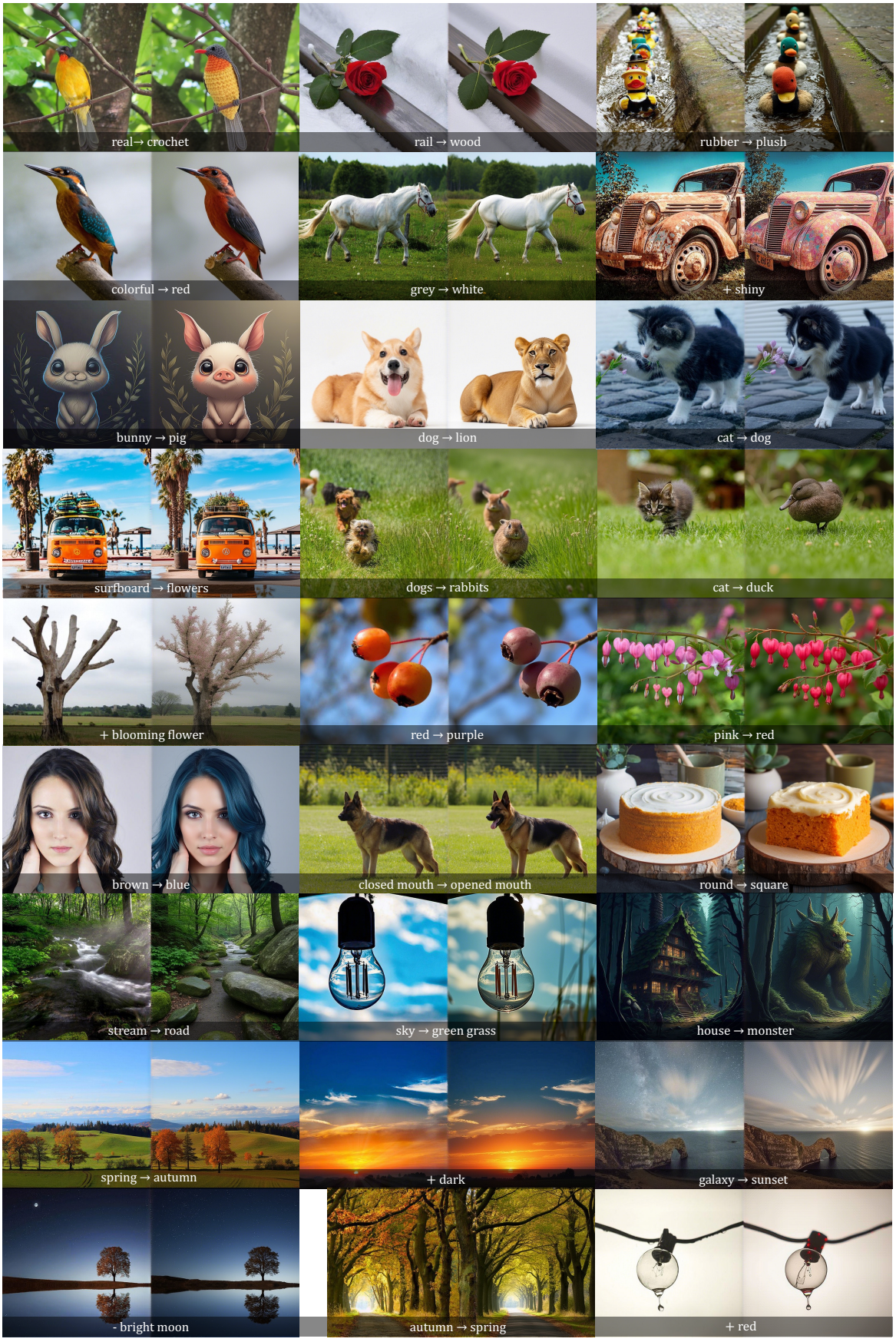


Figure 11. Additional image editing results.