MagDiff: Multi-Alignment Diffusion for High-Fidelity Video Generation and Editing

Haoyu Zhao[1,2], Tianyi Lu[1,2], Jiaxi Gu[3], Xing Zhang[1,2], Qingping Zheng[4], Zuxuan Wu[1,2†], Hang Xu[3], and Yu-Gang Jiang[1,2]

[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3] Huawei Noah's Ark Lab [4] Zhejiang University

Abstract. The diffusion model is widely leveraged for either video generation or video editing. As each field has its task-specific problems, it is difficult to merely develop a single diffusion for completing both tasks simultaneously. Video diffusion sorely relying on the text prompt can be adapted to unify the two tasks. However, it lacks a high capability of aligning heterogeneous modalities between text and image, leading to various misalignment problems. In this work, we are the first to propose a unified **M**ulti-**alig**nment **Diff**usion, dubbed as *MagDiff*, for both tasks of high-fidelity video generation and editing. The proposed MagDiff introduces three types of alignments, including subject-driven alignment, adaptive prompts alignment, and high-fidelity alignment. Particularly, the subject-driven alignment is put forward to trade off the image and text prompts, serving as a unified foundation generative model for both tasks. The adaptive prompts alignment is introduced to emphasize different strengths of homogeneous and heterogeneous alignments by assigning different values of weights to the image and the text prompts. The high-fidelity alignment is developed to further enhance the fidelity of both video generation and editing by taking the subject image as an additional model input. Experimental results on four benchmarks suggest that our method outperforms the previous method on each task.

Keywords: Video Generation and Editing · Multi-Alignment Diffusion · Unified Video Diffusion

1 Introduction

Diffusion model (DM) [20] has been widely applied to many visual tasks, including video generation and video editing. Of them, video generation [17] aims to synthesize a video of good visual quality and video editing [47] requires the non-edited regions should remain consistent as the source video. Since the two tasks have their task-specific problem, thereby different diffusion models are leveraged to handle them separately. Besides, a group of video editing methods [47] adopts

† Corresponding author

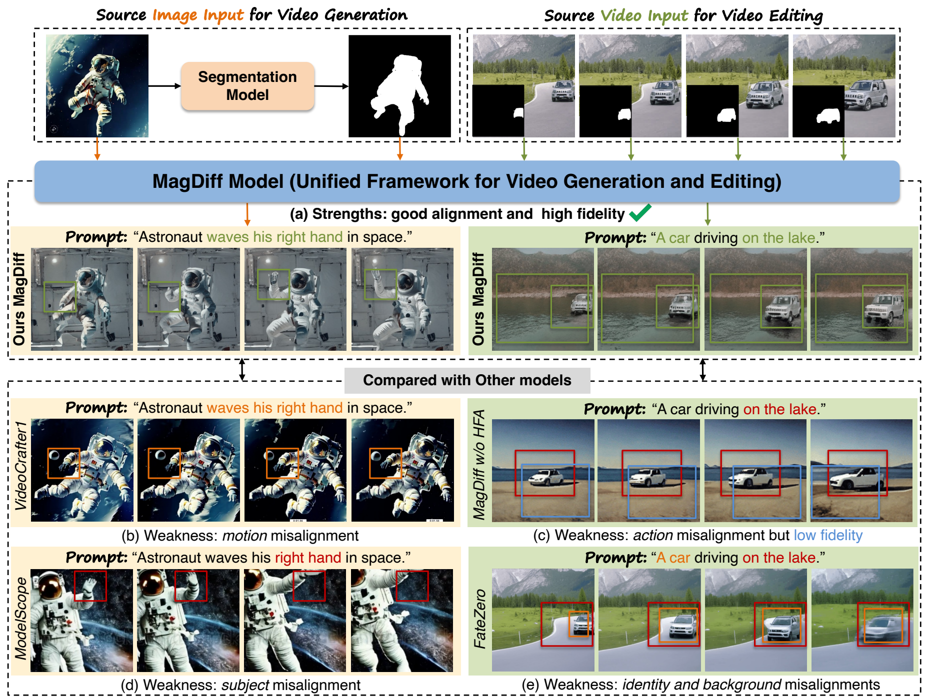

Fig. 1: Comparisons of our proposed unified diffusion model named (a) *MagDiff*, (c) MagDiff *w/o* HFA and other video diffusions, including (b) VideoCrafter1 [9], (d) ModelScope [42], and (e) FateZero [32]. The results show that our proposed MagDiff obtains the best visual performance (*i.e.* good text-and-image alignment and high fidelity) for both tasks of video generation and editing.

the one-shot fine-tuning strategy to improve the performance during the inference. Therefore, it is challenging to employ a unified tuning-free diffusion model to support both tasks at the same time.

Although the traditional video diffusion models [3, 12, 14, 15, 17, 18] which are only conditioned on the textual prompts can be treated as a framework to adapt both tasks, it lacks a capability of aligning the generated image with the given text prompt. Specifically, most video generation methods [9, 12, 14, 17, 37] solely relying on textual prompts cannot precisely control the visual details of the synthesized videos because a specific text description normally maps many various videos. For example, given a prompt "*Astronaut waves his right hand in space*", text-prompt-based video diffusion model suffers from the problem of subject misalignment (see Fig. 1(d) **the astronaut waves his *left hand* rather than the right hand**). Similarly, existing video editing methods [6, 7, 25, 32, 43] leverage the text prompt to edit source video raising a more severe problem, such as identity and background misalignments in Fig. 1(e) (editing with tuning-free inference). The main reason is that it is challenging to align heterogeneous modalities of text and image.

To address this issue, an image prompt has been adopted as complementary information by many researchers [9, 23, 29, 45, 50] to control the video generation. These methods can improve the model's ability to align the text prompt and the generated video since the extra image prompt and the generated image are homogeneous. The existing state-off-the-art VideoCrafter1 [9] considers both image-and-text prompts as conditions to generate a high-fidelity video with a good subject alignment. However, it is found that employing an image prompt as an additional condition is unable to control the action generation of the subject (also known as "action misalignment"), like the astronaut in Fig. 1(b) cannot wave his right hand. This is because the VideoCrafter1 method assigns equal weights on both image and text conditions, neglecting the different video controllability between homogeneous and heterogeneous modalities.

To tackle the aforementioned problems, we propose a unified **M**ulti-**a**li**g**nment Diffusion (*MagDiff*) model, a tuning-free method during the inference stage, for high-fidelity video generation and editing. Our proposed MagDiff introduces three various alignments, including Subject-Driven Alignment (SDA) to integrate both tasks in one framework, Adaptive Prompts Alignment (APA) to trade off the controllability between heterogeneous and homogeneous conditions, and High-Fidelity Alignment (HFA) to maintain the fidelity of the subject image. Specifically, the SDA segments the subject from an image and employs it as the additional condition rather than the whole image, unifying both tasks of video generation and editing in one model. The APA aligns the image prompt and the text prompt with a learnable function in cross-attention blocks, allowing the model to generate more fine-grained content correlated with the subject image and text. The HFA aggregates pixel-level multi-scale information into the latent space via a pyramid encoder, reconstructing the visual details of the subject image in the generated videos. Experimental results on UCF-101, MSR-VTT, DAVIS, and DreamBooth benchmarks show that our proposed MagDiff achieves good performances on generation and editing tasks in both quantitative and qualitative evaluations.

In summary, our work makes the following contributions:

- Our *MagDiff* proposes the subject-driven alignment to unify both tasks of video generation and editing in the single framework by using the subject-driven image as an extra condition rather than the full image.
- Our *MagDiff* develops the adaptive prompts alignment to balance the control strength of homogeneous and heterogeneous conditions, generating fine-grained video well-aligned with both subject image and text.
- Our *MagDiff* introduces the high-fidelity alignment to improve the high-fidelity of the generated or edited videos by aggregating multi-scale contextual information into the latent space.
- Experimental results on UCF-101, MSR-VTT, DAVIS, and DreamBooth benchmarks show that our proposed method achieves good results in both quantitative and qualitative evaluations.

2 Related Work

2.1 Diffusion Models for Video Generation

The great success of diffusion models in image generation [2, 21, 22, 28, 31, 33, 34] has propelled the advancement of video generation. In the early stage, video generation methods use textual prompts as conditions [3, 12, 14, 15, 17, 18] to control the synthesized videos. However, conditioning only on textual prompts makes the synthesized videos limited in visual details [1, 9, 50], such as generating a specific subject or background. To this end, recent methods [13, 23, 29, 45] draw significant attention to integrating image prompts for video generation. The key to image-to-video is adding motion features to the objects in the image. VideoComposer [45] combines the spatial condition (image) and temporal conditions (depth and video) to control the video synthesis. VideoCrafter1 [9] takes both the text and image prompts as the inputs and feeds them into the spatial transformer via cross-attention. I2VGen-XL [50] contains two major stages to get high-resolution videos. The model is trained on large-scale video and image data and then fine-tuned on small high-quality data. Although current generative models incorporate both images and text as control conditions, existing methods often overlook the differences between these multiple modalities, which will limit controllability in video generation. Therefore, we introduce three novel alignment strategies that take into account these heterogeneous modalities.

2.2 Diffusion Models for Video Editing

Before the emergence of text-to-video diffusion models, several studies have explored text-to-image diffusion models for video editing [32, 47], with the incorporation of temporal modules to ensure temporal consistency. Tune-A-Video [47] integrates temporal attention layers into UNet and performs one-shot tuning, while Make-A-Video [37] extends the network with spatial-temporal modules to encompass temporal information. An alternative research direction is influenced by Prompt2Prompt [16] and Plug-and-Play [40], which enable local editing through attention map manipulation. FateZero [32] proposes blending self-attention maps with masks generated by cross-attention maps to facilitate zero-shot video editing. Video-p2p [25] introduces decoupled-guidance attention control to adapt to video scenarios. With the development of high-quality text-to-video diffusion models, a line of work employs text-to-video diffusion models for video editing. Dreamix [27] introduces a mixed fine-tuning strategy with the Imagen Video model [17] for better motion editing. Gen-1 [11] presents a video diffusion model trained with depth information to govern video structure and content. The popularity of the video diffusion model greatly improves this task.

Nevertheless, the visual clues are still under-explored which is essential for identity-preserving during generation. Literature [9] proposes to preserve the content of a reference image while generation. However, without carefully considering the weights of image and text prompts, it still falls short in motion editing and identity preserving.

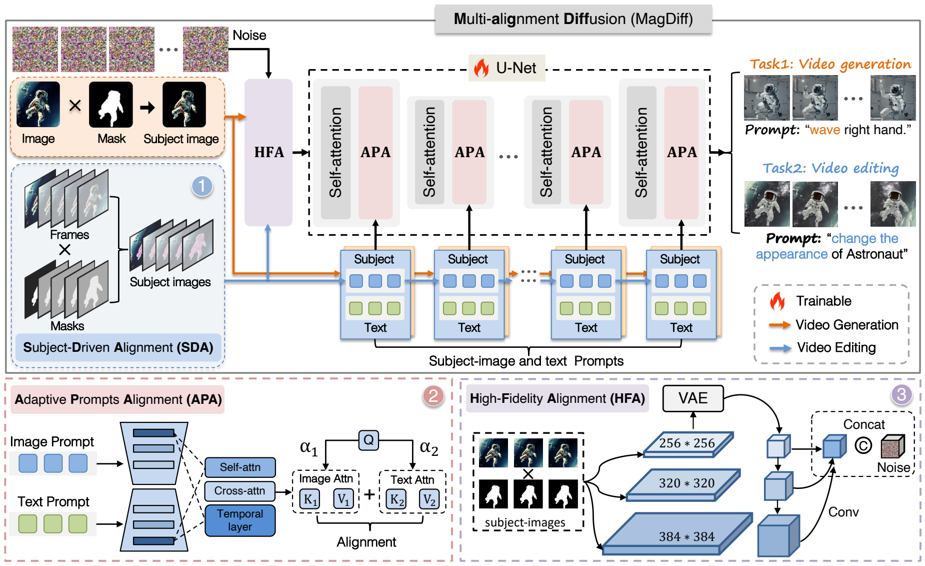

Fig. 2: An overview of our proposed **M**ulti-**a**li**g**nment **Diff**usion (MagDiff), a unified diffusion method supporting both video generation and editing at the same time. Our MagDiff is comprised of three key components: 1) Subject-Driven Alignment (SDA) for unifying two tasks, 2) Adaptive Prompts Alignment (APA) for distinguishing the different controllability between homogeneous and heterogeneous modalities, and 3) High-Fidelity Alignment (HFA) for improving the quality of video generation or editing.

3 Multi-alignment Diffusion (MagDiff)

To achieve high-fidelity video generation and editing tasks in one framework, we propose a **M**ulti-**a**li**g**nment **Diff**usion (MagDiff) model, which solves the multi-alignments among the generated video, text prompt, and subject-image prompt. Fig. 2 shows the overview of our proposed MagDiff conditioned on both text and subject-image prompts to guide video generation and editing using various alignments, including Subject-Driven Alignment, Adaptive Prompts Alignment, and High-Fidelity Alignment. The details of each part are depicted below.

3.1 Subject-Driven Alignment

Video generation and editing are two similar tasks, aiming to generate corresponding content based on text prompts. Differently, the generation task creates a video from pure noise, while the editing task requires the entire video sequence as input and keeps the unchanged parts constant. This difference makes it challenging to unify both tasks in a single model. In this paper, our proposed MagDiff introduces the subject-driven alignment to accommodate both tasks.

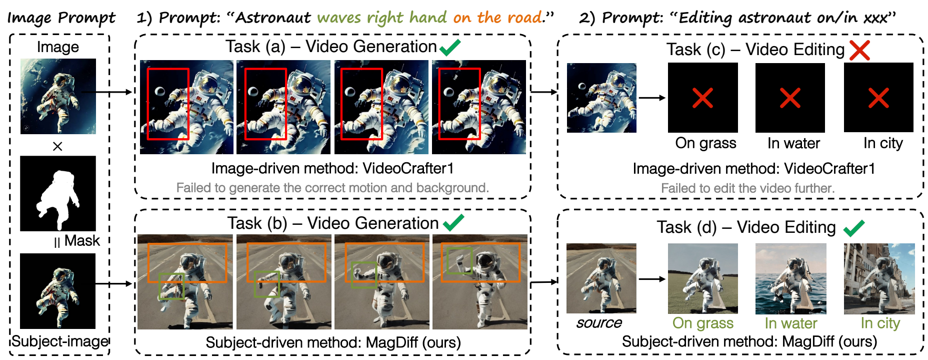

Fig. 3: The comparison between the image-driven method (VideoCrafter1 [9]) and our subject-driven method MagDiff. The subject-driven method can unify two tasks of video generation and editing but the image-driven method does not have this ability.

Image-Driven Alignment. Although existing text-to-video generative models [3, 14] can create videos that precisely describe the content of the given prompts, they cannot achieve a customized appearance and keep the identity of the subjects, as shown in Fig. 1(d). Facing these problems, image-driven video diffusions [9,10,45,50] are proposed to generate videos conditioned on whole reference images and texts. Specifically, given the image prompt $\mathbf{c_i}$ and text prompt $\mathbf{c_t}$, the t-step denoising process in Diffusion Models (DMs) is denoted as:

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\|\mathbf{y} - f_\theta(\mathbf{x}_t; \mathbf{c_i}, \mathbf{c_t}, t)\|_2^2], \tag{1}$$

where the data distribution p_{data} is determined by $\mathbf{c_i}$ and $\mathbf{c_t}$ at the same time. However, existing image-driven methods create videos strictly based on the reference images, leading to misalignment between the generated video and text prompts. For instance, given a prompt of "an astronaut on the road", Fig. 3(a) illustrates that VideoCrafter1 [9] can not follow the right prompt. Thereby, the image-driven method cannot edit video content based on the text prompt, failing to unify video generation and editing in one framework.

Subject-Driven Alignment. To alleviate this problem, we propose subject-driven alignment to enhance the editability condition on both text and image prompts by balancing the tradeoff between them. Specifically, we adopt a segmentation method to extract the subject from the full image, obtaining a subject-driven image. Instead of leveraging the full image as an extra condition, we employ the subject-driven image as the subject-image prompt $\mathbf{c_s}$. Conditioned on the $\mathbf{c_s}$ and the $\mathbf{c_t}$, our proposed MagDiff can align both the subject image and text prompt with the synthesized videos, unifying tasks of video generation and editing in a single model. Therefore, the denoising process in Eq. 1 is changed to

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\|\mathbf{y} - f_\theta(\mathbf{x}_t; \mathbf{c_s}, \mathbf{c_t}, t)\|_2^2]. \tag{2}$$

Through this manipulation, the model can pay more attention to the spatial dimension, not only the temporal information. As illustrated Fig. 3(d), utilizing the subject-image prompt can treat the non-editable parts as the "subject" and preserve them in the edited videos and utilize text prompt to edit the remaining parts, achieving a unified framework for both tasks. Moreover, to unify the inputs of two tasks, we also propose a standardized data input in latent space, which is introduced in Section 3.3.

3.2 Adaptive Prompts Alignment

The subject-image prompt is a homogeneous modality as the generated or edited videos, while the text prompt is a heterogeneous modality. Therefore, the subject image and text prompts have different strengths of alignment to control the video generation in the denoising process. To make a trade-off between them, we introduce an Adaptive Prompts Alignment (APA) module which can well align both prompts, improving the visual performance.

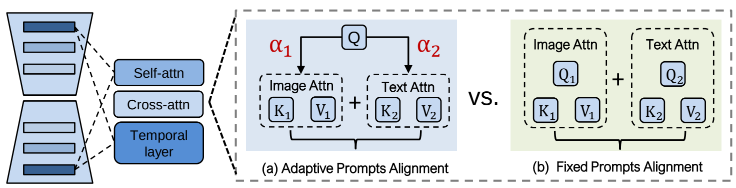

Fig. 4: Comparison of our Adaptive Prompts Alignment (APA) vs. Fixed Prompts Alignment. The APA uses two learnable parameters α_1 and α_2 to adaptively balance the trade-off of alignments between homogeneous and heterogeneous modalities.

Fixed Prompts Alignment. To bring the image prompt into the diffusion model, existing frameworks [10, 26, 27] mainly extract the image feature via the CLIP model and directly inject it into the cross-attention structure in U-Net. As described in Fig. 4(b), we first use CLIP to encode the text prompt and the subject-image prompt into two feature groups Q_1, V_1, K_1 and Q_2, V_2, K_2, separately. Then, we calculate the cross-attention of the two prompts and add them together for feature fusion. The process is denoted as:

$$\text{Attention} = \text{Softmax}(\frac{\mathbf{Q_1}\mathrm{K}_1^\top}{\sqrt{\mathrm{d}}})\mathrm{V}_1 + \text{Softmax}(\frac{\mathbf{Q_2}\mathrm{K}_2^\top}{\sqrt{\mathrm{d}}})\mathrm{V}_2. \tag{3}$$

However, such a function allocates equal weights of controllability or editability to both text prompt and image prompt, failing to consider the inherent differences between the two modalities. The misalignment between them leads to generating or editing the low-quality and low-fidelity video since image and text prompts are flexible and variable. Besides, prior multi-modal works [8, 46] have found that different cross-attentions bring varying results for specific tasks.

Adaptive Prompts Alignment. To address the homogeneous and heterogeneous modality misalignment, we propose the **A**daptive **P**rompts **A**lignment (APA) module to align the image and text prompts in the denoising process. As shown in Fig. 4(a), in the cross-attention block, we keep K_1, V_1 for the text prompt and K_2, V_2 for the subject-image prompt, sharing the query Q. We then combine the two cross-attention features to obtain the adapted feature.

$$\text{Attention} = \alpha_1 * \text{Softmax}(\frac{\mathbf{QK}_1^\top}{\sqrt{\mathrm{d}}})\mathbf{V}_1 + \alpha_2 * \text{Softmax}(\frac{\mathbf{QK}_2^\top}{\sqrt{\mathrm{d}}})\mathbf{V}_2, \tag{4}$$

where α_1, α_2 are two learnable parameters to dynamically control the visual part and textual part, respectively. Different from [9], which only assigns equal weights on both image and text conditions, we find that it is important to give different control for the two modalities. Through cross-attention reassignment, the adapted feature becomes more semantically correlated with the paired subject-image and text prompts.

Furthermore, we compare the experimental results in Table. 6 between the fixed and adaptive prompt alignments and find the APA module has better performance. To prove such a conclusion, we also visualize the cross-attention map in Fig. 7. It can be found that with the APA module, the refinement is learned after aligning the text prompt and subject-image prompt which can bring more fine-grained generative controllability.

3.3 High-Fidelity Alignment

Although aligning the subject-image prompt and the text prompt can achieve more fine-grained control, the fidelity of the subject image is still overlooked, resulting in a loss of detailed appearances. The inherent reason is that the CLIP model encodes the subject-driven image into the features of the high-dimensional semantics rather than the visual details. To achieve the high fidelity of the subject image, our MagDiff proposes a High-Fidelity Alignment (HFA) module which focuses on aligning the generated video with the subject-image prompt.

Compared with the CLIP encoder [26,45], VAE can encode the image to the latent space and decode the latent into the image. Therefore, the HFA shown in Fig. 2 is built based on the VAE model and designed as a pyramid structure. Specifically, given the noise latent $\mathbf{z}_n$ and the subject image x_s, the x_s is sampled into three kinds of sizes. Here, 384×384, 320×320 and 256×256 are chosen empirically. Using the VAE model, the subject images with three resolutions are projected into the latent features $\{\mathbf{z}_s^0, \mathbf{z}_s^1, \mathbf{z}_s^2\}$ respectively. After passing through the VAE, we use convolutional layers (*convs.*) to align VAE features at three different scales, obtaining the output feature $\mathbf{z}_s$ for the subject image. Finally, the latent $\mathbf{z_0}$ for denoising is concatenated by $\mathbf{z}_n \oplus \mathbf{z}_s$. Such a pyramid structure allows it to accept inputs with different resolutions, which can fit contexts with different scales and increase the robustness of the input.

Additionally, our framework also uses VAE to get the video latent representation. Specifically, we encode the reference image with VAE and don't add noise to it, which can inject the appearance into the denoising step. Notably, the structure of the HFA can unify the inputs for video generation and editing tasks. For generation, we use one subject image while the other images are masked, as shown by the orange line in Fig. 2. For editing, we use all the frames from the original videos as input, which helps maintain the content of the video, as indicated by the blue line in Fig. 2.

4 Experiments

4.1 Experimental Setups

Data for Training. Considering the quality of the video content, we select the Pexel Videos dataset [5] to serve as the source data. The dataset contains abundant videos with high quality, each averaging 19.5 seconds in duration. Owing to the lack of subject labels within the original data, we apply the processing approach detailed in *Supplementary* to enhance the dataset's utility for our purposes. We clean up the videos and collect around 76K videos with subjects for training. We initialize the parameters of U-Net from VidRD [14] (5.3M pre-training data).
Evaluation Datasets. We evaluate our MagDiff on four public benchmarks, including UCF-101, MSR-VTT, DreamBooth, and DAVIS. UCF-101 [38] has 101 brief class names (10,000 videos for test), which is commonly employed to assess the generation performance of various methods [18,37,44]. MSR-VTT [48] contains 2,990 videos for testing. DAVIS [19] dataset is proposed for video editing task and DreamBooth [35] dataset is built to evaluate the fidelity of the subject.
Evaluation Metrics. We mainly use two aspects of evaluation metrics:

(i) Metrics for video quality evaluation. Previous works like [3, 12, 37] use two metrics for quantitative evaluation, i.e., **Fréchet Video Distance (FVD)** [41] and Video **Inception Score (IS)** [36]. FVD is a video quality evaluation metric based on FID [30]. Following [37], we use a trained I3D model [5] for calculating FVD. Following previous works [3,18,37], a trained C3D model [39] is used for calculating the video version of IS.

(ii) Metrics for identity consistency and video-prompt alignment. a) We compute the **DINO score** [35] between the generated subject and the given subject image to evaluate the fidelity. b) Following [47], we calculate the average cosine similarity between all pairs of video frames to evaluate the **Frame-consistency**. We calculate the average CLIP score between all frames of generated videos and corresponding prompts to evaluate the **Textual-alignment**.

4.2 Comparison with State-of-the-Art

To evaluate the effectiveness of our unified MagDiff on both video generation and editing, we conducted comparative analyses with various state-of-the-art methods for two tasks separately. For video generation, we compare our method with

[5] https://huggingface.co/datasets/Corran/pexelvideos

the existing video generation methods including text-to-video and text&image-to-video. For video editing, we compare our method with the current video editing methods containing tuning-free and fine-tuning ways during inference.

Table 1: For video generation, quantitative comparison of MagDiff and other methods on UCF-101 and MSR-VTT. All the videos are generated in a zero-shot manner.

Models	**Input Type**	**Training**	**UCF-101**		**MSR-VTT**
		Videos	IS ↑	FVD ↓	FVD ↓
LVDM [15]	text to video	2.0M	-	641.80	-
ModelScope [42]	text-to-video	10M	-	-	550
Make-A-Video [37]	text-to-video	20.0M	33.00	367.23	-
VideoFactory [44]	text-to-video	140.7M	-	410.00	-
PYoCo [12]	text-to-video	22.5M	47.76	355.19	-
I2VGen-XL [50]	text&image-to-video	10M	18.90	597.49	-
VideoComposer [45]	text&image-to-video	10.3M	-	-	580
VideoCrafter1 [9]	text&image-to-video	10.3M	44.53	415.87	465
MagDiff (Ours)	text&image-to-video	**5.3M+76K**	**48.57**	**339.62**	**245**

Table 2: For video generation, quantitative comparison of our MagDiff and other methods from the additional metrics on UCF-101, MSR-VTT, and DreamBooth.

Methods	**UCF-101**			**MSR-VTT**			**DreamBooth**		
	DINO	Textual align	Frame consist	DINO	Textual align	Frame consist	DINO	Textual align	Frame consist
AnimateDiff (V3) [49]	46.2	24.1	89.8	47.3	22.4	89.6	59.1	23.2	90.4
I2VGen-XL [50]	44.1	21.3	89.4	42.7	18.4	**89.8**	58.6	22.9	89.7
MagDiff (Ours)	**50.8**	**25.4**	**90.2**	**50.2**	**23.2**	88.4	**61.4**	**25.4**	**92.2**

Evaluation of Video Generation. Table 1 exhibits the best performance of our proposed MagDiff in terms of all metrics on the UCF-101 and MSR-VTT benchmark datasets. Specifically, our method surpasses all text-to-video generation methods, since it utilizes the subject within the image, which contains less visual information. Besides, our MagDiff also outperforms other text&image-to-video methods on both FVD and IS metrics, because we use the subject-driven image as a condition rather than the full image. To demonstrate the fidelity and identity consistency of subject images within the videos synthesized by MagDiff, we employ the DINO score, Textual-align, and Frame-consistency for evaluative purposes on the UCF-101, MSR-VTT, and DreamBooth compared with [49,50]. The results in Table 2 indicate that the videos we generate have good text alignments, and the continuity between frames is also commendable.

Evaluation of Video Editing. Table 3 compares our MagDiff and other video editing methods on the popular DAVIS dataset with CLIP-Score metrics of text and image. The results show that our MagDiff surpasses the tuning-free method Framewise IP2P [4], with improvements of 2.54 and 4.10 on the textual-align and frame-consistency metrics respectively. Compared to fine-tuning methods during inference, our tuning-free MagDiff method is much more robust and performs competitively with the Tune-A-Video [47] and FateZero [32] methods, verifying the superiority of our proposed MagDiff.

Table 3: For video editing, quantitative comparison of our MagDiff and other video editing methods on DAVIS benchmark.

Methods	Inference method	Textual-align	Frame-consistency
Tune-A-Video [47]	Fine-Tuning	28.33	90.45
FateZero [32]	Fine-Tuning	23.81	92.92
Framewise IP2P [4]	Tuning-Free	25.11	86.76
MagDiff (Ours)	Tuning-Free	**27.65** (+2.54)	**90.86** (+4.10)

Human Evaluations. We perform human evaluations by using a panel of 34 human raters, over 15 videos with corresponding prompts for each method. We adopt the Likert Scale [24] to evaluate subject fidelity, prompt alignment, and the quality of the generated videos. The range of these three scores is between 1 and 5, which represents from very dissatisfied to very satisfied. Table 4 compares our MagDiff with both VideoCrafter1 [9] and AnimateDiff (V3) [49] on human evaluations. We maintain the original input format of these models, which is the complete image without the mask. Our method achieves the highest scores on both subject fidelity and text alignment, as well as producing high-quality videos. These results indicate that our method performs well under human evaluation and demonstrates its superiority over existing methods.

Table 4: Human-preference aligned results from three different aspects, with the rank of each aspect in the brackets.

Methods	Image-prompt Alignment	Text-prompt Alignment	Quality
VideoCrafter1 [9]	3.2	2.8	3.2
AnimateDiff (V3) [49]	3.5	3.6	3.3
MagDiff (Ours)	**4.4**	**4.1**	**3.7**

Qualitative Evaluation. Fig. 5 offers a visual comparison between our MagDiff model and other methods for both tuning-free video generation and editing. Examples in Fig. 5(a) present that our method can synthesize smoother and higher fidelity videos than other video generation methods. Fig. 5(b) demonstrates that

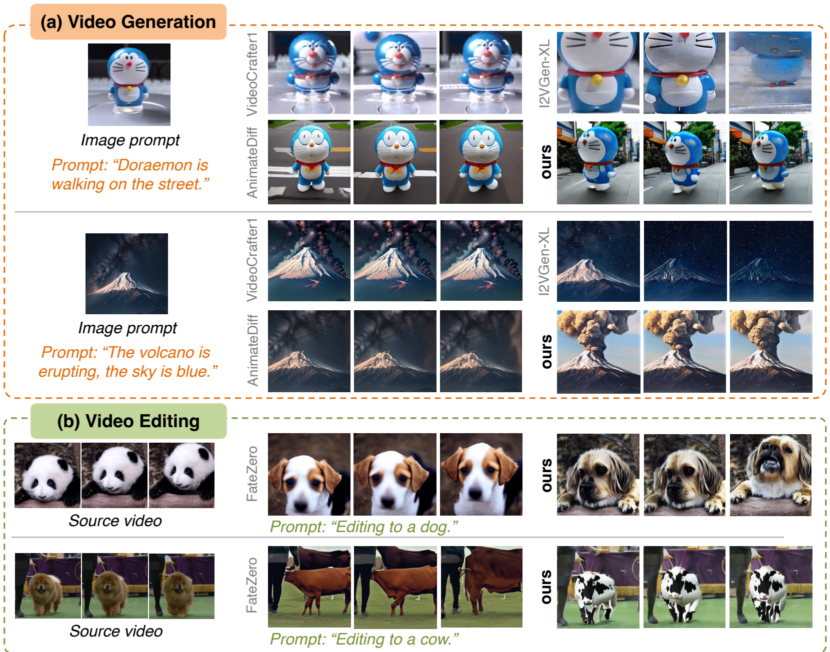

Fig. 5: For qualitative evaluation, we compare our MagDiff with VideoCrafter1 [9], I2VGen-XL [50], and AnimateDiff (V3) [49] on the generation task (orange dotted box) and compare with FateZero [32] on the editing task (green dotted box).

our method can edit video contents more aligned with the user-defined text prompts. This further verifies the effectiveness of our unified framework.

4.3 Ablation Studies

Baseline analysis of three alignments. Table 5 analyzes the influences of the SDA, APA, and HFA modules for our MagDiff. The base model is the VidRD [14], which is a T2V method. We find that using the SDA can enhance the overall effect of our method. We analyze that the mask can help the model better implement text control capabilities. Meanwhile, we suppose that the HFA can introduce specific pixel information directly into the latent space, it also plays a significant role in improving the effectiveness of the model.

Additionally, as presented in Fig. 6, **1)** compared to the first and last rows in Fig. 6, it can be observed that lacking subject-driven prompts generates videos unmatching the user-defined text prompt (see Fig. 6(a)) and is also unable to edit the source videos conditioned on the given text description (Fig. 6(d)). This validates the effectiveness of our proposed SDA. Besides, **2)** compared to the second and last rows in Fig. 6, we find that while missing APA module in MagDiff can generate or edit a video, better aligning the text condition than the first rows, it suffers from a low-fidelity of visual results. This is due to overlook the fidelity

Table 5: Baseline analysis of SDA, APA, and HFA within MagDiff. We use the VidRD [14] as the basemodel, which is a text-to-video model.

Methods	HFA	APA	SDA	UCF-101		MSR-VTT
				IS ↑	FVD ↓	FVD ↓
MagDiff	✓	✗	✗	42.11	530.26	372
	✗	✓	✗	40.47	534.84	394
	✓	✓	✗	43.39	444.67	311
	✓	✗	✓	45.74	367.23	274
	✗	✓	✓	46.85	388.41	286
	✓	✓	✓	**48.57**	**339.62**	**245**

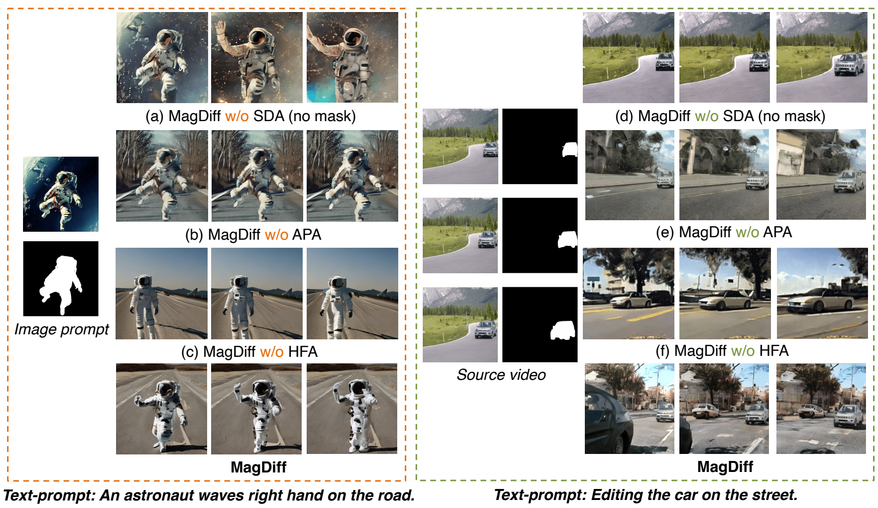

Fig. 6: Visualization of the three alignments in MagDiff, including SDA, APA, and HFA. The orange and green dotted boxes show the generated and edited results.

of the subject image. Consequently, **3)** compared to the last two rows, it can be found that adding the HFA is useful to improve the video's fidelity because the pixel-level multiscale information is aggregated into the latent space.

Effect of APA. To validate the effectiveness of our proposed APA, we compare it with the Fixed Prompts Alignment (FPA). Table. 6 illustrates that "MagDiff w/ APA" outperforms "MagDiff w/ FPA" by a clear margin. On the UCF-101 and MSR-VTT datasets, "MagDiff w/ APA" has FVD scores of 48.79 and 46 lower than "MagDiff w/ FPA", respectively. The results show the superior performance of our proposed APA, verifying that better alignment between the subject-image prompt and the text prompt can help the model understand fine-grained text. To further prove this, in Fig. 7, conditioned on the text prompts, we visualize the APA's cross-attention maps between the generated video frame and the words "right hand", and "wings". Results demonstrate that FPA wrongly activates both

Table 6: Performance comparison of FPA and APA within MagDiff. "MagDiff w/ APA" denotes the use of APA, while "MagDiff w/ FPA" denotes the use of FPA.

Methods	UCF-101		MSR-VTT
	IS ↑	**FVD ↓**	**FVD ↓**
MagDiff w/o FPA	44.26	388.41	291
MagDiff w/ APA	**48.57**	**339.62**	**245**

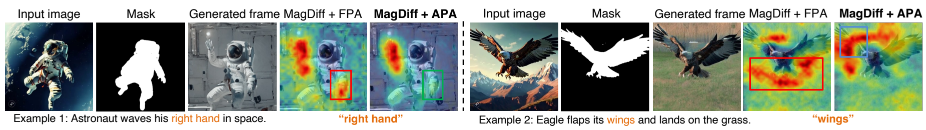

Fig. 7: Visualization of cross-attention maps between the "MagDiff +" FPA and APA.

the astronaut's left and right hands, while APA has a better understanding and activates the right hand corresponding to the text description.

Effect of different values for α_1 and α_2. We also explore the impact of different values for α_1 and α_2 on the APA module. Table 7 compares the effects of fixing the two parameters with trainable. We conduct comparisons for three different group values of α_1 and α_2. The results show that the model achieves optimal performance when using a trainable way. This is because trainable parameters better balance image-text during training, improving model performance.

Table 7: Effects of different values for α_1 and α_2 in APA cross-attention.

Methods	UCF-101		SR-VTT
	IS ↑	**FVD ↓**	**FVD ↓**
$\alpha_1 = 0.3$ & $\alpha_2 = 0.7$	46.45	391.72	283
$\alpha_1 = 0.5$ & $\alpha_2 = 0.5$	45.11	364.89	296
$\alpha_1 = 0.7$ & $\alpha_2 = 0.3$	47.20	386.32	267
Trainable	**48.57**	**339.62**	**245**

5 Conclusion

In this work, we first propose a unified diffusion model named MagDiff, for both the video generation and editing tasks. In our framework, we mainly solve the three kinds of alignments to achieve high-fidelity video generation, including subject-driven alignment, adaptive prompts alignment, and high-fidelity alignment. Experimental results based on four benchmarks show that our method achieves good results in both quantitative and qualitative evaluations.

Acknowledgements

This project was supported by NSFC under Grant No. 62032006 and No. 62102092. Qingping Zheng was supported by the Postdoctoral Fellowship Program of CPSF under Grant No. GZC20241489.

References

1. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945 (2024)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Fidler, S.W.K.S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR. pp. 22563–22575 (2023)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR. pp. 18392–18402 (2023)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
6. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: ICCV. pp. 23206–23217 (2023)
7. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: ICCV. pp. 23040–23050 (2023)
8. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: ICCV. pp. 357–366 (2021)
9. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
10. Chen, H., Wang, X., Zeng, G., Zhang, Y., Zhou, Y., Han, F., Zhu, W.: Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. arXiv preprint arXiv:2311.00990 (2023)
11. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. arXiv preprint arXiv:2302.03011 (2023)
12. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: CVPR. pp. 22930–22941 (2023)
13. Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S.S., Shah, A., Yin, X., Parikh, D., Misra, I.: Emu video: Factorizing text-to-video generation by explicit image conditioning. arXiv preprint arXiv:2311.10709 (2023)
14. Gu, J., Wang, S., Zhao, H., Lu, T., Zhang, X., Wu, Z., Xu, S., Zhang, W., Jiang, Y.G., Xu, H.: Reuse and diffuse: Iterative denoising for text-to-video generation. arXiv preprint arXiv:2309.03549 (2023)
15. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)

16. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
17. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
18. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In: ICLR (2023)
19. Jay, Z., Gao, D., Bai, J., Mike, S., Li, X., Zhen, D., Aishani, S., Kurt, K., Forrest, L.: Loveu@cvpr'23 - track4 (2023), https://sites.google.com/view/loveucvpr23/track4
20. Jonathan, H., Ajay, J., Pieter, A.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
21. Jonathan, H., Tim, S.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
22. Konpat, P., Nattanat, C., Suttisak, W., Supasorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: CVPR. pp. 10619–10629 (2022)
23. Li, Z., Tucker, R., Snavely, N., Holynski, A.: Generative image dynamics. arXiv preprint arXiv:2309.07906 (2023)
24. Likert, R.: A technique for the measurement of attitudes. Archives of psychology (1932)
25. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv preprint arXiv:2303.04761 (2023)
26. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410 (2023)
27. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329 (2023)
28. Nan, L., Shuang, L., Yilun, D., Antonio, T., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: ECCV. pp. 423–439 (2022)
29. Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models. In: CVPR. pp. 18444–18455 (2023)
30. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: CVPR. pp. 11410–11420 (2022)
31. Prafulla, D., Alexander, N.: Diffusion models beat gans on image synthesis. NeurIPS **34**, 8780–8794 (2021)
32. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)
33. Quinn, N.A., Prafulla, D.: Improved denoising diffusion probabilistic models. arXiv preprint arXiv:2102.09672 (2021)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
35. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR. pp. 22500–22510 (2023)
36. Saito, M., Saito, S., Koyama, M., Kobayashi, S.: Generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. IJCV **128**, 2586–2606 (2020)

37. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
38. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
39. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
40. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR. pp. 1921–1930 (2023)
41. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. ICLR Workshop (2019)
42. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
43. Wang, W., Jiang, Y., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint arXiv:2303.17599 (2023)
44. Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J., Liu, J.: Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. arXiv preprint arXiv:2305.10874 (2023)
45. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. arXiv preprint arXiv:2306.02018 (2023)
46. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: CVPR. pp. 10941–10950 (2020)
47. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV. pp. 7623–7633 (2023)
48. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR. pp. 5288–5296 (2016)
49. Yuwei, G., Ceyuan, Y., Anyi, R., Yaohui, W., Yu, Q., Dahua, L., Bo, D.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
50. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)