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

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

A Overview

We provide more details about MagDiff including:

- We demonstrate the training details of MagDiff in Section B.
- We introduce our method of training data process in Section C, and the details of the model and experiments in Section D.
- We provide more qualitative results of video generation and editing in Section E.
- We conclude the limitation in Section F.

B Training Details

During the model training, we sample eight frames for one video to train the denoising network. Our MagDiff is initialized with weights from VidRD model [3]. For model training, we utilize high-quality data from existing video dataset. For the frame input, we utilize the widely used data augmentation strategy for training, including center crop and random shuffle. We also set a 15% probability of randomly dropping the prompt during training. Noticed that we do not train all the parameters, only the temporal layers, transformer blocks of the spatial layers, and the newly added projections are trainable. The learning rate is set at 5×10^{-5} for all training tasks. The VAE model and CLIP model are frozen. When doing inference, the classifier-free guidance is set as 7.5. For latent diffusion model sampling, we use DDIM in all our experiments. All experiments are conducted using eight Nvidia Tesla V100 GPUs for $50K$ iterations with a batch size of 64.

† Corresponding author

C Data Process for Training

Our proposed MagDiff is a tuning-free method that needs training on the video dataset, which encompasses paired text-video data along with their corresponding segmentation masks. Due to the lack of existing suitable training data, we propose an automated data processing methodology that segments the subject within a video to construct the requisite training set, shown in Fig. A.

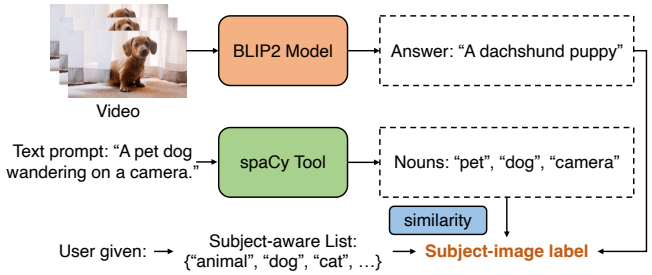

Fig. A: The overview of the method for training data processing.

We first analyze the existing datasets carefully to get the subject in the image. Due to the low quality of the existing dataset, we find that the misalignment between the video data and the caption makes it difficult to get the correct mask area of the subject. To address this issue, we utilize BLIP-2 [4] to generate more precise captions. Specifically, given a video containing m frames $\mathcal{V} = \{v_i \mid i \in [1, m]\}$ with its original caption $\mathcal{P}$, we select the first frame v_1 and use BLIP-2 to get question-paired answers. We set the question "*What is the foreground in the picture?*" for BLIP-2 and get the answer $\mathcal{A}$. At the same time, the origin prompt $\mathcal{P}$ is handled with spaCy tool to extract nouns $\mathcal{S} = \{s_i \mid i \in [1, n]\}$. To ensure the accuracy of the subject's label, we maintain a user-given subject-aware list $\mathcal{W}$ to prioritize entities for a certain topic, which contains words of some specific domains, such as "*animal, dog, cat, ...*".

We calculate the distance metric to quantify the disparity between the answer $\mathcal{A}$ and each word within the list $\mathcal{W}$. Subsequently, we assess the similarity between the words in $\mathcal{W}$ and each noun in the subject-aware list $\mathcal{S}$, individually. The selection process is defined in Algorithm 1. Once we get the label of the subject image, we use the SAM-Track [2] tool to segment the subject mask of each frame. It can ensure the consistency of the segmentations among the frames. Furthermore, we append the final subject-image label to the original video caption, generating an augmented caption. Finally, we merge the video-text pairs with segmentation masks for MagDiff training.

Algorithm 1: Subject label selection.

Input: $\mathcal{A}$, $\mathcal{W}$, $\mathcal{S} = \{s_i \mid i \in [1, n]\}$.
Output: e: subject-image label.

```
scores1 = List(), scores2= List()
for i = 0 to length(W) do
    scores1.append(Similarity(A, W_i))
    for j = 0 to length(S) do
        scores2.append(Similarity(S_j, W_i))
if max(scores1) > θ then e = A;
else e = max(scores2) ;
```

Fig. B: The word cloud of the subject image in training data.

Moreover, considering the quality of the existing text-video dataset, we clean up and sample the videos from the Pexel Videos dataset [5]. We show the word cloud of subject images in Fig. B. In addition, to ensure that the videos are suitable for training, we set three filtering rules to clean up the selected videos: **(1)** We filtered out videos with a short side resolution of less than 512. **(2)** We excluded examples with an entity area ratio of less than 5% or more than 60%. **(3)** We filtered out subject-image labels with no clear meaning. After filtering, we get 76K videos for training. We simply divide the label classifications into four kinds: person, animal, objects, and others. We show the statistics of caption length, entity categories, and clip durations in Fig. C. The dataset contains abundant kinds of subjects, such as *woman, man, person, tree, dog*, and so on. Besides, we provide some examples of the training data in Fig. E

D Model and Experiments Details

In this section, we introduce the additional model details and experimental setting details. Note that figure and table references with numerical indices pertain

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

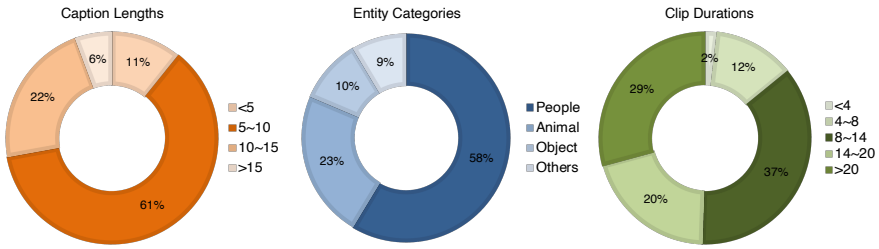

Fig. C: Statistics of caption lengths, entity categories, and clip durations in the training data. Our training data exhibits a diversity of captions and videos with different lengths. The subject regions mainly consist of humans and animals, which offer our training clear boundaries and outstanding dynamic features.

to those within the regular *conf.* paper, while those with alphabetical indices refer to the supplementary materials.

Model details. Our MagDiff is built based on the VidRD [3] model, which is a text-to-video generation framework trained on 5.3M video-text data. Following VidRD model, we use a regularized autoencoder to compress the original pixels into latent space to save computation and memory. The autoencoder contains an encoder $\mathcal{E}$ for encoding pixel features $\mathbf{x}$ into latent features $\mathbf{z}$ and a decoder $\mathcal{D}$ for decoding $\mathbf{z}$ back to $\mathbf{x}$. We employ the autoencoder which is pre-trained by reconstruction loss. In the training stage, the parameters of the autoencoder are frozen. Besides, in the High-Fidelity Alignment module (HFA), each frame is cropped to the size 384×384, 320×320, and 256×256 (width $\times$ height) for training. In the Adaptive Prompts Alignment (APA) module, we merge three kinds of resolutions to 256×256.

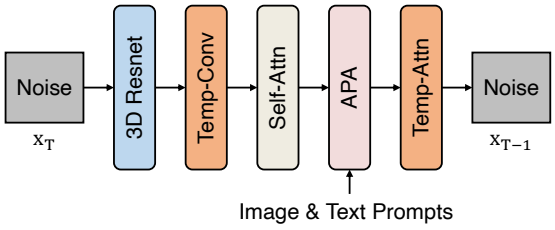

Fig. D: The structure of the U-Net. The noise x_T in T timestep is denoised into x_{T-1}.

We provide the structure of the U-Net in Fig. D. In each denoising step, the noise is processed by 3D Resnet, temporal layers, self-attention layer, and our proposed APA module. The temporal layers include two types of structure: the *Temp-Conv* and the *Temp-Attn*. The *Temp-Conv* represents 3D convolution layers and *Temp-Attn* denotes the temporal attention layers. These temporal

layers are injected into the image U-Net structure to learn the action motion and temporal features from video data. We load the pre-trained parameters of *Temp-Conv* and *Temp-Attn* from VidRD model. The newly added key and value in the APA module and projects in the HFA module are randomly initialized. For efficient training, only part of the U-Net network layers are trainable. In Fig. D, we train the two temporal layers, the self-attention layer, and the APA module, since the Resnet block is frozen.

Additionally, although our MagDiff can achieve both video generation and editing tasks, we only train the MagDiff once. During training, we only use the first image with its mask as the input for the HFA module and as the subject-image prompt for the APA module. It is not necessary to train the model individually for the video editing task. During video editing inference, we can directly replace the first image with the mask to the whole video with masks for each frame. This effect comes from our designed SDA module, which allows the MagDiff to pay attention to each frame of the input during the denoising process. In the generation task, the network generates the frames that the corresponding reference images are masked, thereby achieving content generation by text prompt. In editing tasks, the network focuses on the areas that need to be edited while preserving the subjects of each reference image, thereby achieving content editing. The once-trained and tuning-free inference for two tasks makes our proposed MagDiff more practical.

Experimental setting details. Here, we introduce the details of our experiments, including detailed settings and further analyses. For comparison of FVD and IS metrics on MSR-VTT and UCF-101, we get the first frame of the video in the two datasets and segment the subject's mask as the input of our model. We evaluate all the methods in a tuning-free manner for fair comparison. It is important to note that, due to the misalignment between the image content and the subject's label, about 0.4% of the videos cannot be segmented into subjects based on the label in the two datasets. For these examples, we treat the entire video frame as the subject, with the mask set to null. To maintain the fairness of the compared "text+image-to-video" methods [8,9], we input the whole image as the condition into these methods for evaluation. When evaluating the MagDiff on DreamBooth [6] dataset, we directly use the images to segment the subjects.

For comparison of the video editing task, we sample eight frames for one video as the input of HFA and APA modules. For each frame in one video, we segment the subject with the same label using the SAM-Track tool. We compare two kinds of editing methods, including fine-tuning methods [5,7] and tuning-free method [1]. In a tuning-free manner, our MagDiff has a significant improvement.

For comparison of human evaluations, we carefully select 15 different images and write corresponding text prompts to generate videos, covering diverse scenes, styles, and objects. When doing the user study, 34 volunteers are asked to rank the video quality, text-prompt, and image-prompt alignment from one to five.

In ablation studies, we analyze three alignments in Table 5. We attempt to decouple and individually validate HFA, APA, and SDA modules within MagDiff. "MagDiff w/o SDA" represents we do not use the mask of the subject image

(indicated by × in the table). Because MagDiff needs the reference image as the input in the HFA module and APA module to introduce the image information, we cannot remove the HFA and the APA at the same time. In Fig.5, we test our MagDiff with the image prompt and the text prompt. We employ the image prompts with their corresponding masks (not shown in the figure), *i.e., the Doraemon, the volcano, the panda, and the dog.* In Fig.6, "MagDiff w/o SDA" denotes that the mask is not used and the whole reference image is put into the model. "MagDiff w/o APA" and "MagDiff w/o HFA" are both tested under the condition of having the SDA module (with mask in the image), representing the scenarios where only the HFA module or the APA module exists, respectively.

E More Visualizations

To demonstrate the superiority of our proposed MagDiff, we provide more visualizations of video generation (in Fig. F) and video editing (in Fig. G).

F Limitations

Our proposed MagDiff exhibits remarkable capabilities in preserving the fidelity of the subject image and enough alignment between the image prompt and the text prompt. However, because our model employs the diffusion model as the backbone, it is computationally intensive and time-consuming, especially when dealing with large images. It may also not be suitable for all types of images, such as low-contrast or noisy images. These challenges indicate potential directions for future research, such as efficient inference and model robustness.

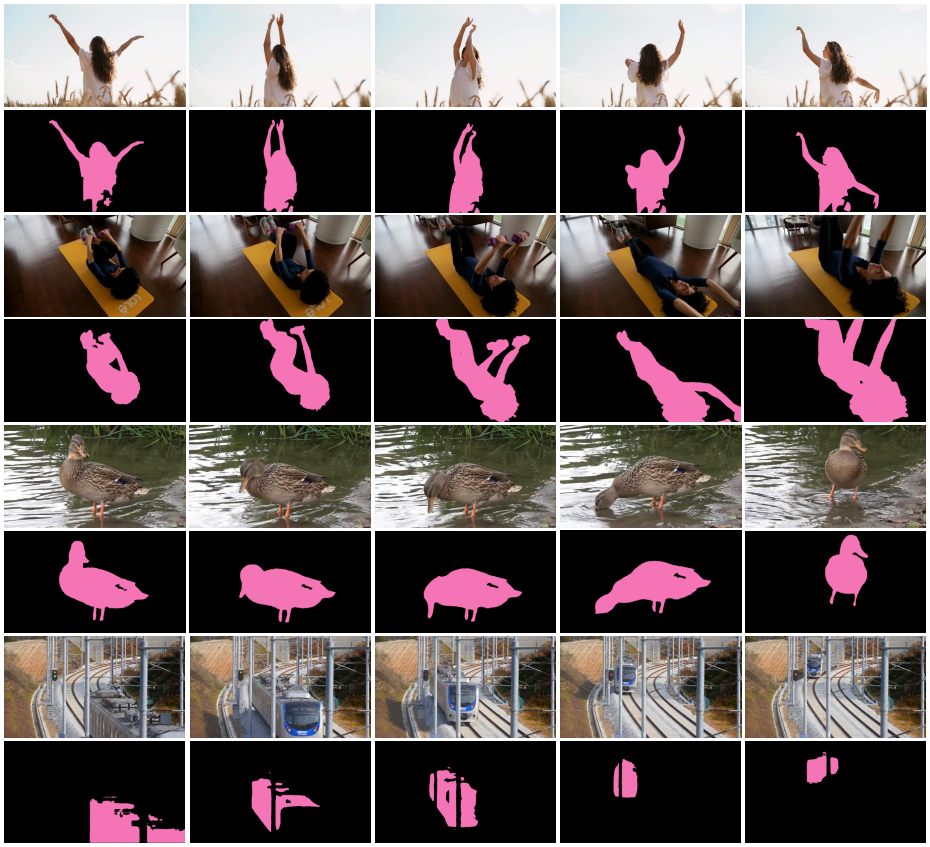

Fig. E: Examples of training data. In each case, the first row is raw video frames and the second row is the subject's mask.

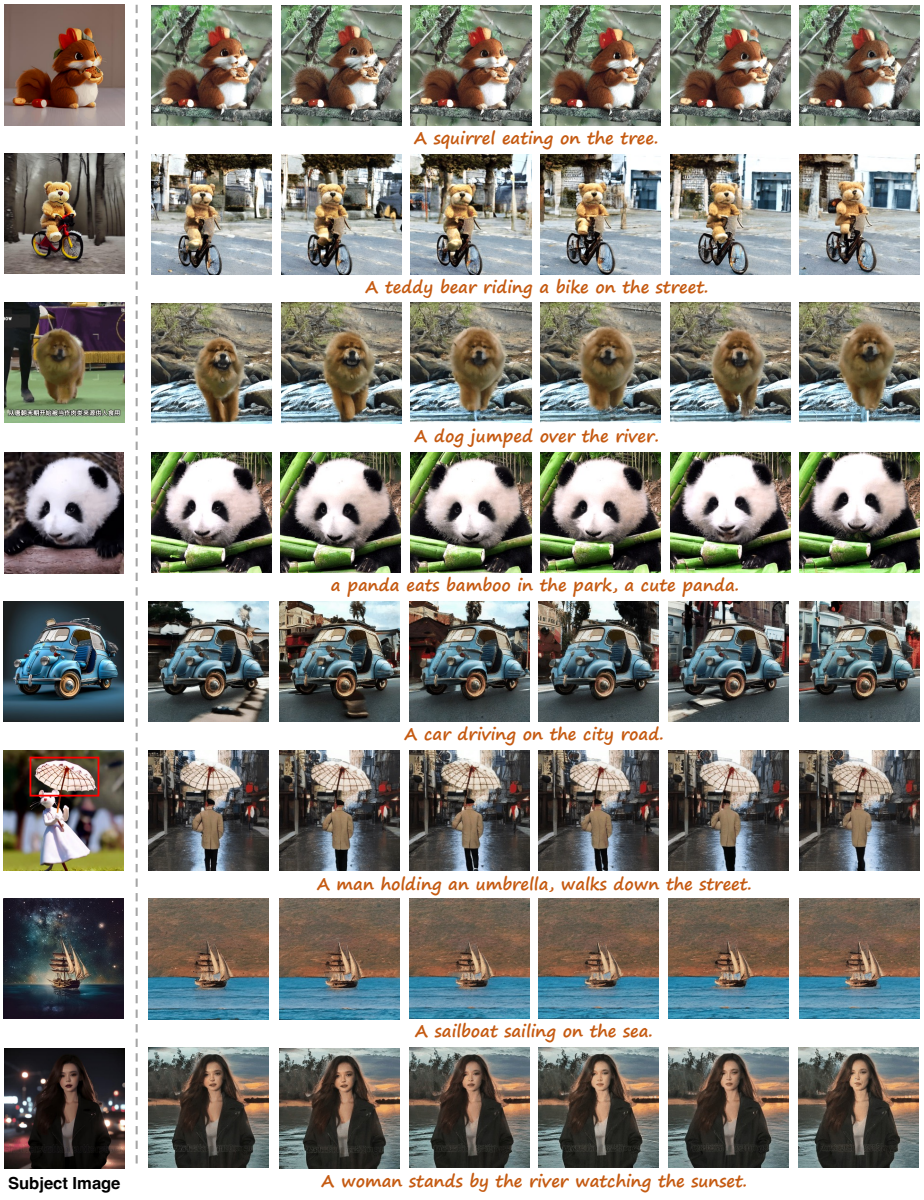

Fig. F: More qualitative generated examples of video generation that conditioned on subject-image prompts. The first column shows the reference subject images.

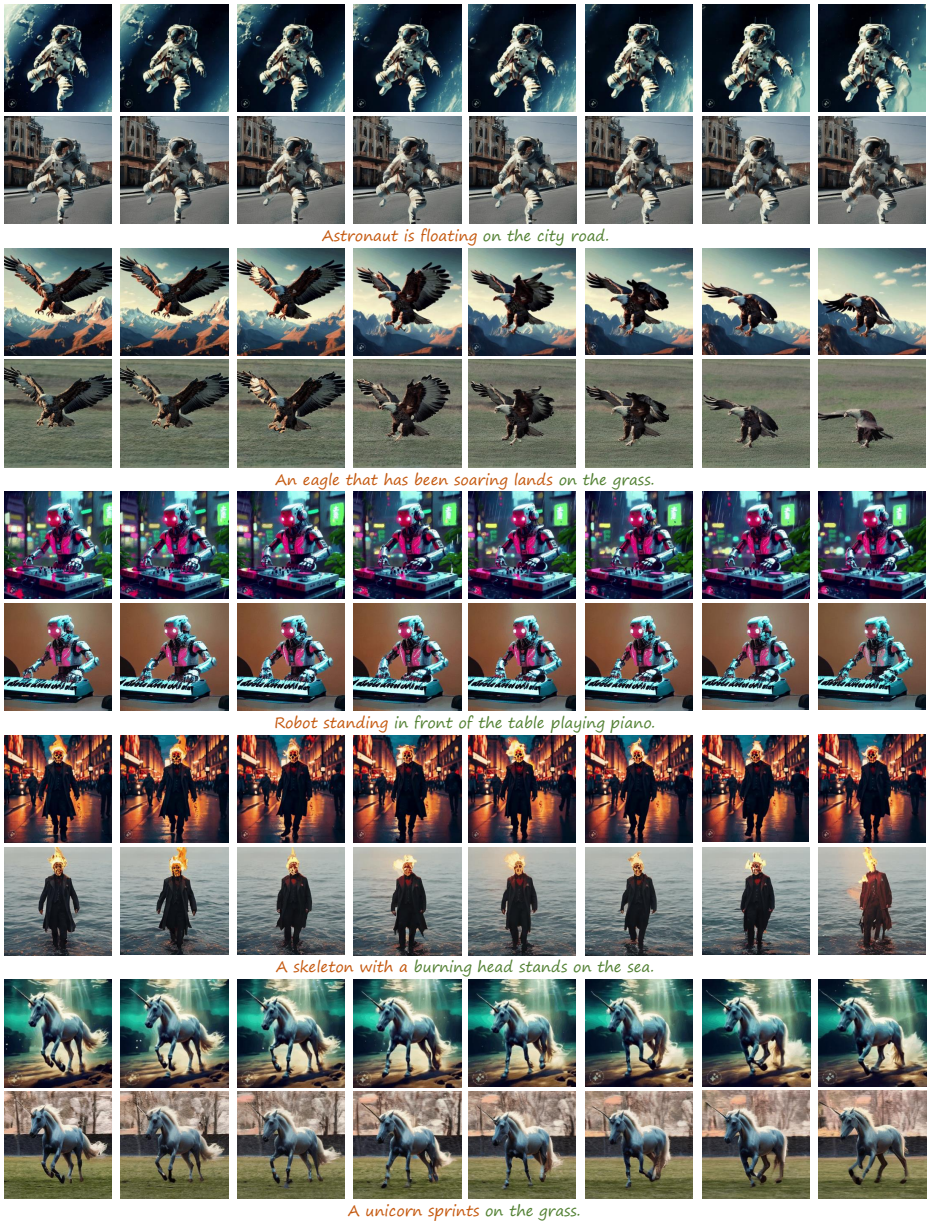

Fig. G: Qualitative results of video editing. In each case, the first row shows the original video and the second row displays the editing result of MagDiff according to the text prompts. We mainly concentrate on editing the green words specified in the prompt.

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR. pp. 18392–18402 (2023)
2. Cheng, Y., Li, L., Xu, Y., Li, X., Yang, Z., Wang, W., Yang, Y.: Segment and track anything. arXiv preprint arXiv:2305.06558 (2023)
3. 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)
4. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
5. 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)
6. 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)
7. 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)
8. 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)
9. 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)