LAMP: Learn A Motion Pattern for Few-Shot-Based Video Generation

Ruiqi Wu[1 2 *] Liangyu Chen[2] Tong Yang[2] Chunle Guo[1 †] Chongyi Li[1] Xiangyu Zhang[2]
[1]VCIP, CS, Nankai University [2]MEGVII Technology
wuruiqi@mail.nankai.edu.cn, {chenliangyu, yangtong, zhangxiangyu}@megvii.com
{guochunle, lichongyi}@nankai.edu.cn

Figure 1. **Our text-to-video results**. The motion prompts and video prompts are listed, respectively. Our LAMP works effectively on diverse motions. The generated videos are temporal consistent and close to the video prompts. Moreover, two advantages of LAMP can be reflected in the above results. (1) The proposed first-frame-conditioned training strategy allows us to use powerful SD-XL for first-frame generation, which is beneficial to producing highly detailed following frames. (2) Good semantic generalization properties of the diffusion model are preserved (*e.g*. imposing smile's motion on unseen comic style) since our tuning way.

Abstract

With the impressive progress in diffusion-based text-to-image generation, extending such powerful generative ability to text-to-video raises enormous attention. Existing methods either require large-scale text-video pairs and a large number of training resources or learn motions that are precisely aligned with template videos. It is non-trivial to balance a trade-off between the degree of generation freedom and the resource costs for video generation. In our study, we present a few-shot-based tuning framework, ***LAMP****, which enables text-to-image diffusion model* ***L****earn* ***A*** *specific* ***M****otion* ***P****attern with 8 ∼16 videos on a single GPU. Specifically, we design a first-frame-conditioned pipeline that uses an off-the-shelf text-to-image model for*

* This work is done during Ruiqi Wu's internship at MEGVII Technology.
† Correspondence author.

content generation so that our tuned video diffusion model mainly focuses on motion learning. The well-developed text-to-image techniques can provide visually pleasing and diverse content as generation conditions, which highly improves video quality and generation freedom. To capture the features of temporal dimension, we expand the pre-trained 2D convolution layers of the T2I model to our novel temporal-spatial motion learning layers and modify the attention blocks to the temporal level. Additionally, we develop an effective inference trick, shared-noise sampling, which can improve the stability of videos with computational costs. Our method can also be flexibly applied to other tasks, e.g. real-world image animation and video editing. Extensive experiments demonstrate that LAMP can effectively learn the motion pattern on limited data and generate high-quality videos. The code and models are available at https://rq-wu.github.io/projects/LAMP.

1. Introduction

In recent years, generative models, particularly diffusion-based models [11, 33, 34], have shown remarkable achievements in generating images from textual prompts, *i.e*. text-to-image generation (T2I) [22, 24, 27, 28, 30]. Despite the success made in the T2I field, which has provided substantial technical groundwork, there remain large gaps in the development of text-to-video (T2V) generation: how to generate consistent frames and understand the motion patterns implicit in textural prompts.

Several recent works [2, 9, 12, 31, 36, 42] try to bridge these gaps by directly training a diffusion-based T2V model using millions of text-video pairs. These approaches facilitate a deeper understanding of the relationship between the video and the textural prompt. However, the massive demand for labeled data and the heavy training burden are unaffordable for most researchers, constraining the development of this research line. Another research line [5, 6, 21, 25, 37, 39] involves utilizing a video template and manipulating content using diffusion models while keeping the original motion. Those template-based methods are also known as video editing. Although these methods can prove cost-effective, especially with the proposal of one-shot [37] and even zero-shot [6, 25, 39] algorithms, the use of given video template significantly limits the generation freedom. Besides, recent T2V-Zero [18] modifies the T2I diffusion models to generate consistent videos without training. Nevertheless, it is challenging to transfer the text-image domain knowledge to the text-video domain in a zero-shot manner, resulting in the limitation of T2V-Zero to generate similar-looking frames with random motions.

It is essential to achieve a trade-off between training burden and generation freedom while making models understand the motions. Since the pretrained T2I diffusion model has good semantic comprehension guided by the prompts, it is reasonable that very little data is needed to make it understand the correspondence between prompts and motions and generate diverse videos. Therefore, we attempt to explore a novel few-shot setting for the T2V task. The new setting aims at tuning a T2I diffusion model to **L**earn **A** common **M**otion **P**attern from a small video set.

When tuning a T2I model to a T2V model in a few-shot manner, two issues need to be addressed. (1) Due to the limited data amount, there is a risk of over-fitting the content within the video set. If the generated videos are similar to the video set, it undermines one of our core goals, namely generation freedom. (2) The base operators of T2I diffusion models only work on spatial dimensions, which limits their ability to capture temporal information within videos.

With the two challenging issues, we propose a baseline method for few-shot T2V generation, named ***LAMP***. Our solution to the first issue is the proposed **first-frame-conditioned pipeline**. It decouples the T2V task into two sub-tasks, generating the first frame by a pre-trained T2I model and predicting subsequent frames using our tuned video diffusion model. The proposed pipeline seamlessly integrates the first frame as a condition without involving any additional model modification (*e.g*. changing the data structure of inputs or adding new cross-attention layers). Specifically, during training, we retain the first frame of the input video, adding noise and imposing the loss only on the subsequent frames. Since the first frame provides the majority of the video's content, our model can focus on the relationship between the subsequent frames and the first frame, *i.e*. the motion pattern rather than the contents. During the inference, the first frame is generated by a pre-trained T2I model, such as SD-XL [24]. We observe that a high quality of the first frame can boost the video generation performance through the proposed pipeline. With the reference provided by the first frame, our model, which is based on Stable Diffusion v1.4 (SD v1.4) [28], can preserve the high-quality content generated by SD-XL throughout the video. Facing the second issue, we design **temporal-spatial motion learning layers** to capture the features of temporal and spatial dimensions simultaneously. Since predicting subsequent frames based on the first frame is required in the proposed pipeline, we modified the base operator based on the video prediction tasks [15, 20], which will be introduced in Sec. 3.4. As in previous works [18, 37], we modify the attention layers to build effective communication between frames. Moreover, we adopt a **shared-noise sampling strategy** during inference, which constructs the original noise for each frame from a shared noise. This strategy significantly improves the quality and stability of the generated videos with negligible computational costs.

We evaluate our LAMP on several motion cases. With a simple tuning using 8 $\sim$16 videos on a single GPU, the pro-

posed LAMP can generate videos with the common motion pattern of the video set and generalize well to unseen styles and objects. (See Figure 1). Our key contributions can be summarized as follows:

- We present a new setting of the few-shot tuning for the T2V generation task, aiming to strike a balance between generation freedom and training costs.
- We propose LAMP, a baseline method for few-shot T2V called LAMP, which effectively learns the motion pattern in the given video set following a simple tuning.
- We introduce a first-frame conditioned pipeline that uses the first frame as a condition, effectively decoupling the motion and content, which simplifies the T2V task significantly.
- We introduce the temporal layers and inference tricks, offering key insights into few-shot-based video generation.

2. Related Work

2.1. Text-to-Image Diffusion Models

Recently, diffusion models [11, 33, 34] beat GANs [4, 7, 40], VAEs [19, 32, 35], and flow-based [3, 8] approaches and have been in the limelight for text-to-image generation because of their stable training and outstanding performance. For example, GLIDE [22] uses textural prompts as conditions and adopts classifier-free guidance [10] to improve image quality. DALLE-2 [27] introduces the pre-trained CLIP [26] model to align the features of images and text. Imagen [30] injects the features from a large language model to diffusion models for better prompts understanding and proposes a cascaded pipeline to generate high-resolution images from coarse to fine. To ease the computational burden of the iterative denoising process, Rombach *et al*. propose LDM [28] that uses an autoencoder [4, 19] to reduce the redundancy of images. LDM compresses an image into low-dimension latent space by a pre-trained autoencoder first, then learns to denoise noisy latent data. With the success achieved by LDM, many variants [23, 41] are proposed to improve the performance further. More recently, the SD-XL [24] is presented, which can generate extremely photo-realistic images with high-definition details. In our work, SD-XL is utilized to generate the first frame, and SD-v1.4 is modified for subsequent frame prediction.

2.2. Text-to-Video Diffusion models

The thriving of diffusion-based models in the text-to-image field demonstrates its potential in text-to-video generation. The mainstream works can be divided into two categories: open-domain T2V generation and template-based methods.

Open-domain T2V generation. During the early stage, ImagenVideo [12] and Make-A-Video [31] learn T2V on the pixel level. However, the video length and resolution are significantly limited due to the high computation in the pixel space. MagicVideo [42] is then proposed, which trains a new autoencoder on video data. As the appearance of LDMs [28] to the T2I field, MagicVideo boosts the computational effectiveness for T2V generation. Blattmann *et al*. [2] present an LDMs-based T2V diffusion model, which adds extra 3D convolutional layers on frozen pre-trained layers. VideoComposer [36] adds diverse conditions, *e.g*. sketch and motion vectors, to the T2V model by a novel encoder. AnimateDiff [9] trains a set of motion layers capable of being applied to customized T2I models [14, 29], enabling them to produce videos in a consistent style. The above methods achieve remarkable performance for T2V generation. However, the necessity of training these models on large-scale data like WebVid-10M [1] and HD-VILA-100M [38] poses a significant barrier for most researchers. In addition, some zero-shot methods [13, 16, 18] have been proposed, yet they often suffer from suboptimal frame consistency.

Template-based methods. Template-based T2V generation aims to facilitate video-to-video translation with the guidance of user prompts, which is also known as video editing. Dreamix [21] and GEN-1 [5] are two pioneer works in template-based methods, while their training costs are comparable to open-domain T2V methods. Then, Tune-A-Video [37] proposes a new one-shot setting that uses a T2I model to overfit the origin video, which can be implemented on consumer-grade GPUs. FateZero [25] proposes a training-free method by injecting the cross-attention map of the source video and modifying attention layers. Rerender-A-Video [39] and TokenFlow [6] further improve the consistency of videos with the integration of priors and conditional guidance. Different from the objectives of template-based methods, our few-shot T2V setting aims to achieve a higher degree of freedom in video generation rather than precisely aligning with the motion pattern of a template video.

3. Method

In this section, Sec. 3.1 and Sec. 3.2 first introduce the preliminary knowledge and the new few-shot setting. Next, Sec. 3.3 details our proposed first-frame-conditioned pipeline. Sec. 3.4 is followed to describe how we modify a T2I diffusion model to T2V generation. Finally, Sec. 3.5 introduces our shared-noise sampling strategy and some techniques that can improve performance during inference time.

3.1. Preliminaries

In this section, we introduce the preliminary knowledge of the diffusion-based model. Given data $x_0 \in X$, a Markov chain can be defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \quad (1)$$

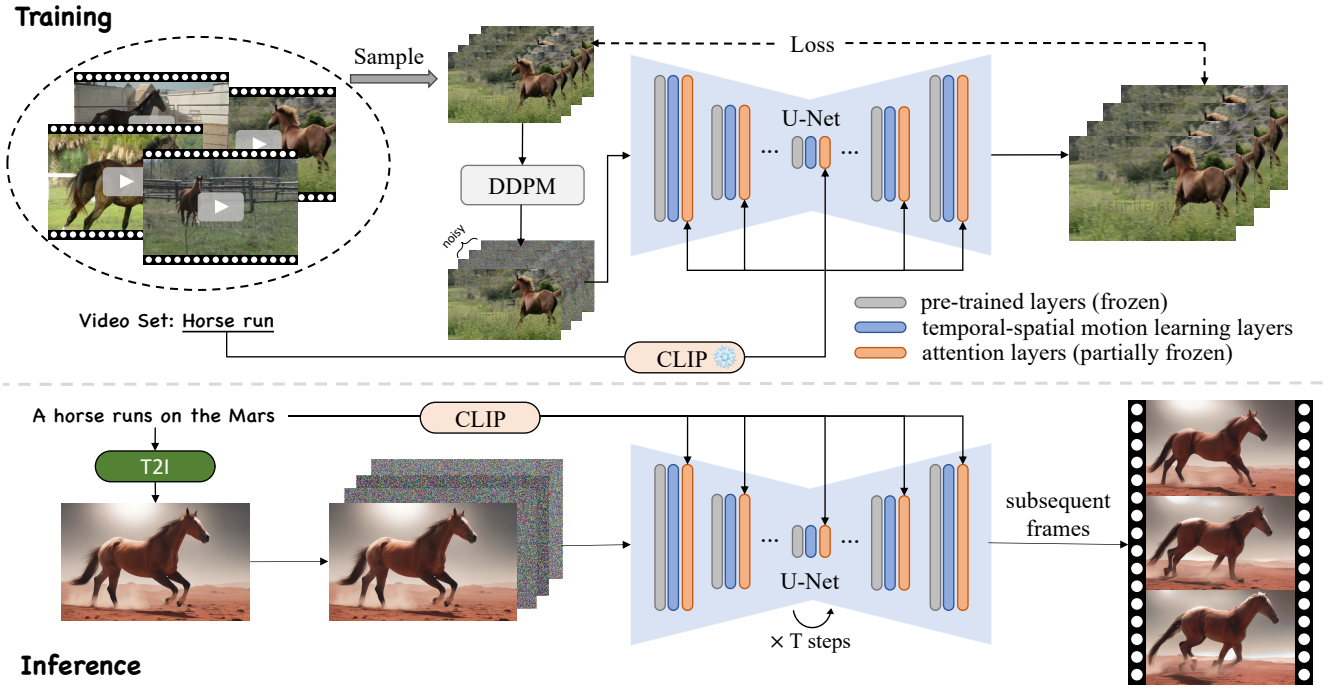


Figure 2. **Framework of LAMP.** LAMP learns a motion pattern from a small video set, enabling the generation of videos imbued with the learned motion patterns. This approach strikes a balance between training resources and generation freedom in video generation. We transfer text-to-video generation to the first-frame generation and subsequent-frame prediction, *i.e.*, decoupling a video's contents and motions. During training, we add noise and compute loss functions for all frames except the first frame. Moreover, only the parameters of newly added layers and the query linear layers of self-attention blocks are tuned. During inference, we use a T2I model to generate the first frame. The tuned model only works on denoising the latent features of subsequent frames with the guidance of user prompts.

where $t = 1, ..., T$, T is the total number of steps. β_t is a coefficient that controls the noise strength in step t. The iterative noise adding can be simplified as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_t)$. Diffusion models learn the distribution of dataset X by minimizing the training objective, which can be written as:

$$\arg\min_{\theta} \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,I), t, c}[||\epsilon - \epsilon_\theta(x_t, t, c)||_2^2], \tag{3}$$

$\epsilon_\theta(\cdot)$ denotes the noise prediction function of diffusion models, c is the conditions like textual prompts. After training, diffusion models can generate data from noise by reversing the noise-adding process.

However, the computational burden becomes substantial when diffusion models are used to generate high-resolution images. To address this challenge, Latent diffusion models (LDMs) for T2I generation have been proposed, adopting an auto-encoder to achieve all operators in the latent space. They can acquire low-redundancy initial data by encoder and reconstruct generated results by decoder. LDMs are also used in our method to generate high-resolution videos.

3.2. Our Few-shot-based T2V Generation Setting

Existing T2V approaches require large-scale data for training or rely on a template video to obtain low-degree-of-freedom generative capabilities. In order to make video generation inexpensive and flexible, we propose a novel setting: few-shot T2V generation. Supposing that there is a video set $\mathbf{V} = \{\mathcal{V}_i | i \in [1, n]\}$ contains n videos and a prompt $\mathcal{P}_m$ to describe the common motion as training data. The proposed new setting is to tune a T2I model on the given video set and the motion prompt. The tuned model can generate a new video $\mathcal{V}'$ with a similar motion pattern to $\mathbf{V}$ from a prompt $\mathcal{P}$ that is related to the motion. We hope to learn the common motion pattern from a small video set while ignoring the contents. Meanwhile, the training cost is affordable because of the small data size. Based on the proposed setting, we modify pre-trained T2I models and present a baseline framework for few-shot T2V generation.

3.3. First-Frame-Conditioned Pipiline

Due to the limited data in the few-shot tunning process, there is a risk of overfitting the content of the small dataset, potentially compromising the degree of generation freedom. To direct our model's focus toward motion, we propose the first-frame-conditioned pipeline to decouple motions and

contents. The proposed pipeline is illustrated in Figure 2. Based on our observation, the first frame contains the majority of the contents of a short video. It is natural to use the first frame as a condition, enabling the model to pay more attention to motions. Therefore, the T2V generation task is translated to first-frame T2I generation and subsequent-frame prediction. There are previous works [5, 36] that have also used the first frame as a condition. They concat it to the input noise or add a specific encoder to inject the features into networks. However, applying these methods in the few-shot setting is challenging, as the limited data makes it nearly impossible to facilitate model training through substantial modifications to T2I models. In contrast, the proposed first-frame-conditioned pipeline can achieve comparable effects with slight parameter changes, as detailed in Sec. 3.4.

Specifically, let $\mathcal{V} = \{f^i | i = 1, ..., l\}$ be a video contains l frames and encode them into latent space: $\mathcal{Z} = \{z^i | i = 1, ..., l\}$. When training the model, we preserve the original signal of z^1 and add noise to $\{z^2, ..., z^l\}$. The loss functions can be written as:

$$\mathcal{L} = \mathbb{E}_{\mathcal{Z}, \epsilon \sim \mathcal{N}(0,I), t, c}[||\epsilon^{2:l} - \epsilon_\theta^{2:l}(\mathcal{Z}_t, t, c)||_2^2], \quad (4)$$

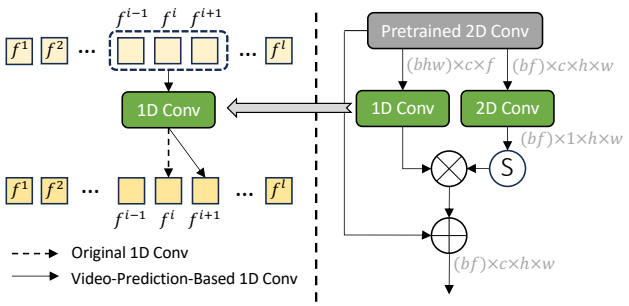
where $\epsilon^{2:l}$ is the added noise from 2nd to l-th frame, respectively. Other notations are consistent with Eq. (3). After training, the model gains the capability to generate a video with the motion pattern of the video set according to the first frame. During inference, the powerful SD-XL [24] is employed to provide the first frame $\hat{f}^1$, which is decoded to $\hat{z}^1$ Then, a sequence, $[\hat{z}^1, \epsilon^2, ...\epsilon^l]$, where ϵ is a random noise, is fed to the model for the whole video generation. At each step, we preserve the latent of the first frame and denoise the subsequent frames.

The proposed pipeline effectively avoids learning the contents of the video set so that it can train a model on limited data. Another advantage lies in the quality of content generated by SD-XL, providing a good reference for video generation. This approach enables us to leverage the advantages of well-established T2I techniques. The first-frame-conditioned pipeline significantly benefits both prompt alignment performance and generation diversity. Moreover, this pipeline is also appealing in its flexibility in applications *e.g.* real-world image animation and video editing, as detailed in Sec. 5.

However, the original T2I models treat frames as independent samples. Thus, the features of the first frame cannot be used to establish temporal relationships between frames and generate videos. The following section introduces how we enable the model to work at the temporal level.

3.4. Adapt T2I Models to Video

Temporal-spatial motion learning layers. To empower the T2I model to extract temporal features, we inflate



(a) The proposed 1D conv (b) Temporal-spatial motion learning layer

Figure 3. The details of the proposed temporal-spatial motion layers. (b) illustrates that 1D convolutions are added on pre-trained layers to capture information along the temporal dimension. 2D convolution layers with an output channel number of 1 control the spatial level's motion strength. The 1D convolutional layers utilize the former two frames to generate the current frame, as shown in (a). f^i denotes the i-th frame.

the pre-trained 2D convolutional layers into the proposed temporal-spatial motion learning layers. As illustrated in Figure 3(b), the proposed layer consists of two branches. Suppose the latent features of the input video are represented as as a 5D tensor with a shape of $b \times c \times f \times h \times w$. In the temporal branch, the tensor is reshaped into $bhw \times c \times f$ and fed to a 1D convolutional layer. However, since the 1D convolution kernel can only work on a spatial coordinate at a time, it fails to take the essential spatial features into account. Consequently, a 2D convolution with an output channel of 1 along with a Sigmoid function is added as compensation for spatial features. The input features are reshaped into $bf \times c \times h \times w$ in the spatial branch.

Considering that our first-frame-conditioned pipeline needs to predict subsequent frames based on the given first frame, which is similar to video prediction [15, 20], we design our 1D convolutional layers in a video prediction manner, as shown in Figure 3(a). When the kernel slides through the features of frames $\{f^{i-1}, f^i, f^{i+1}\}$, our **video-prediction-based 1D convolution** produces the features of f^{i+1} instead of f^i as in the original version. Thus, we can utilize the former two frames to predict the subsequent frame, *i.e.* effectively achieving video prediction in the base operators. Moreover, to avoid our newly added layers polluting the generation capability of the original T2I model, all parameters are zero-initialized, as done in ControlNet [41].

Attention layers. We also modify the attention layers to ensure consistency. For self-attention layers, all key and value features are obtained from the first frame, which can be written as:

$$\text{Attention}(Q^i, K^1, V^1) = \text{Softmax}(\frac{Q^i (K^1)^T}{\sqrt{d}}) V^1, \quad (5)$$

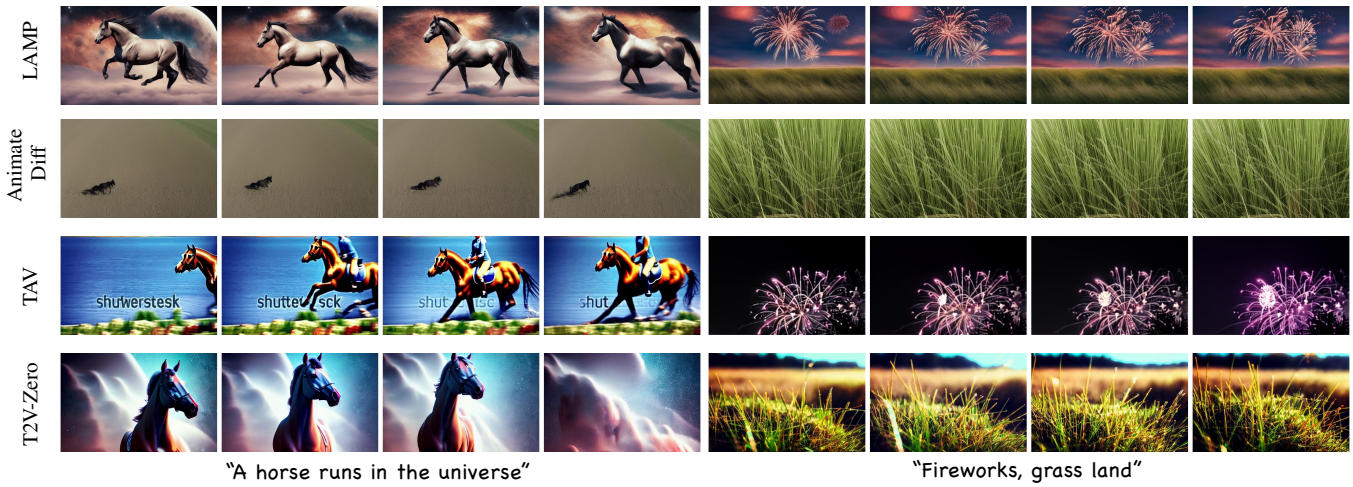


Figure 4. Qualitative comparison between the proposed LAMP and three baselines. **Zoom in for the best view.**

the superscript $i \in \{1, ..., l\}$ indicates the features are from the i-th frame. Combined with the proposed pipeline, the reformulated self-attention layers facilitate subsequent frames to refer back to the conditions established by the first frame. Besides, it has been demonstrated that such a modification can effectively preserve the main object even without tuning [18]. Moreover, following the modification in [37], we have incorporated temporal attention layers, which are self-attention layers working on the temporal dimension.

3.5. Shared-noise Sampling During Inference

During inference, we propose a simple yet effective shared-noise sampling strategy further to improve the quality of the generated videos quality. Specifically, we first sample a shared noise $\epsilon^s \sim \mathcal{N}(0, I)$. Then, a noise sequence $[\epsilon^2, ..., \epsilon^l]$ with the same distribution as the base noise is sampled. In our sampling strategy, the original noise ϵ^i for the i-th frame generation is updated as:

$$\epsilon^i = \alpha\epsilon^s + (1 - \alpha)\epsilon^i, \tag{6}$$

where α is a coefficient to control the sharing degree. We empirically set $\alpha = 0.2$ in our experiments. This approach ensures a consistent noise level across each frame, ultimately manifesting as consistency in the generated videos. Intuitively, this approach is in accordance with the prior knowledge that every frame of a video has certain similarities. Mathematically, the reduced noise variance can contract the dynamic range of the latent space, contributing to a more stable generation process. Besides, the AdaIN [17] technique on latent space and histogram matching at the pixel level are used for post-processing. The efficacy of our free-lunch inference strategies will be demonstrated in Sec. 4.3.

4. Experiments

4.1. Implementations

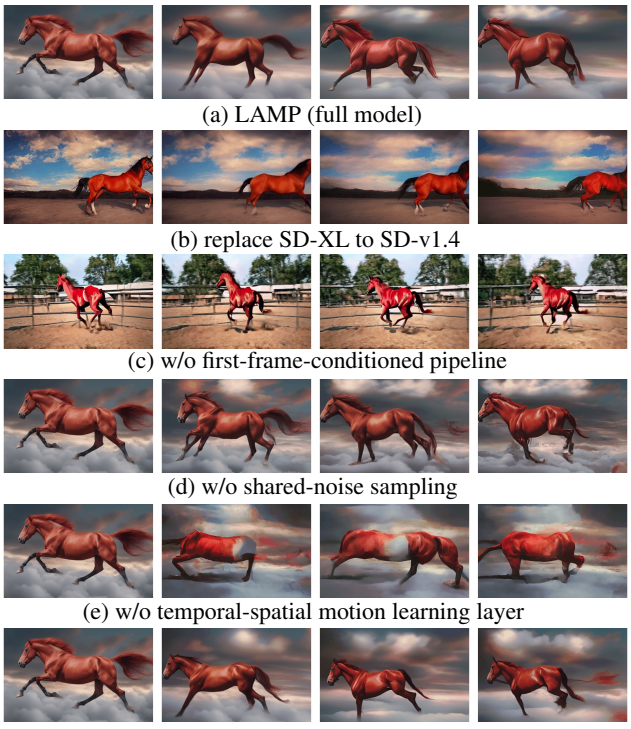
In our experiments, we generate videos with resolutions of 320×512 and 16 frames. We use SD-XL [24] for the less computationally intensive first frame generation and the relatively more lightweight SD-v1.4 [28] for the more computationally demanding prediction of subsequent frames, thereby balancing the inference cost of the two stages. For training, we use a set of self-collected videos ranging from $8 \sim 16$, randomly sampling a 16-frame clip during each iteration All frames are resized to a resolution of 320×512. Only the parameters of new-added layers and the query linear layer in self-attention blocks are tuned, and the learning rate is set to 3.0×10^{-5}. All experiments are implemented on a single A100 GPU and only need $\sim$ 15 GB vRAM for training and $\sim$ 6 GB vRAM for inference.

4.2. Comparisons

We train our LAMP for 8 motions, including helicopter (rigid motion), waterfall (fluid motion), rain & firework (particle motion), horse running (animal motion), birds flying (multi-body motion), turn to smile (human emotion) and play the guitar (human motion). We design 6 prompts for each motion to build an evaluation set containing 48 videos. Three publicly available methods, which are large-scale pre-trained AnimateDiff [9], one-shot-based video editing method Tune-A-Video [37], and zero-shot-based Text2Video-Zero [18], are selected as our comparison baselines. We consider representative work under a variety of mainstream settings, thus effectively reflecting the advantages of our few-shot learning setting. Notably, for each motion pattern, we randomly select a video from the corresponding video set as the template to train Tune-A-Video [37]. The comparisons are constructed in the view of objective and subjective.

Table 1. **Quantitative comparisons with the evaluated text-to-video methods.**

Method	Alignment ↑	Consistency ↑	Diversity ↓
Tune-A-Video [37]	27.2227	94.8742	84.7186
T2V-Zero [18]	26.9424	91.4713	73.0136
AnimateDiff [9]	28.8779	97.8131	73.4723
LAMP (Ours)	**31.3547**	**98.3085**	**71.6535**



(a) LAMP (full model)

(b) replace SD-XL to SD-v1.4

(c) w/o first-frame-conditioned pipeline

(d) w/o shared-noise sampling

(e) w/o temporal-spatial motion learning layer

(f) w/o video-prediction-based 1D Conv

Figure 5. **Ablation results.** The given prompt is 'A red horse runs in the sky'.

Quantitative results. We evaluate our LAMP against baselines in terms of textural alignment, frame consistency, and generation diversity. The objective metrics and user study are used for a comprehensive evaluation.

Objective metrics. To measure the textual alignment of a video, we average each frame's CLIP score [26]. Following [37], we also represent the frame consistency by the mean cosine similarity of CLIP image embedding across all frame pairs. Since generation freedom is one of our core goals, we also include generation diversity in quantitative evaluation. We use the average CLIP image embedding of all frames to represent a video, subsequently computing and averaging the cosine distance across all video pairs. A lower score denotes lower similarity, *i.e.*, better diversity. Table 1 presents the quantitative results of LAMP and baselines. Across all three evaluation criteria, our method achieves state-of-the-art performance against the other baselines.

User study. We further conduct a user study to evaluate our approach and three baselines subjectively. We randomly select 24 cases from our evaluation set. In each case, we ask the participant "Which video do you think has better visual quality and better matches the scene and motion of the prompt '...'?" The user study garnered a total of 70 responses from a diverse group of participants, including both experts in the field and individuals with no specific background knowledge. Statistically, 46.84% of respondents favor our method, with AnimateDiff [9] achieving 19.11% and Tune-A-Video [37] achieving 22.15%. However, it is worth noting that Tune-A-Video polarizes choices in different situations. When there are similarities between the layout of the given video template and the scene described by the prompt, combined with its own good frame consistency, it can be approved by most volunteers. Conversely, the textural alignment of its generated video is poor, *e.g.* "Fireworks, grass land" shown in Figure 4. Besides, 11.90% of the participants select Text2Video-Zero [18] as their preference. As a result, our LAMP obtains the highest approval rate among the participants.

Qualitative results. We present several visual examples of our method and three baselines in Figure 4. AnimateDiff [9] learns motion layers on large-scale data and inserts them into personalized T2I models to generate videos with specific styles and better visual quality. However, this approach cannot be combined with the better-performing but heterogeneous T2I model, resulting in a limitation in textural alignment capabilities even though the consistency and diversity are satisfying. This limitation is apparent in cases 'A horse runs in the universe' and 'Fireworks, grass land'. Tune-A-Video (TAV) can only generate videos with the same motion, with the prompts sometimes unable to effectively control the generated videos due to overfitting on the given video. While T2V-Zero produces visually pleasing frames, it falls short in generating videos with meaningful motion patterns. In contrast, our LAMP achieves good consistency and generates videos with proper motion patterns, benefiting from the proposed motion learning layers. Besides, using the advantage of our first-frame-conditioned pipeline, the proposed method achieves visual quality on par with state-of-the-art T2I models, even with the modifications based on SD-v1.4. Figure 1 and supplementary materials provide more visual results. Our method understands the learning motions well and can generalize to diverse, even unseen, scenes and styles.

4.3. Ablation Study

We conduct ablation experiments to demonstrate the effectiveness of each proposed component. The visual results are shown in Figure 5. As we can see in Figure 5(b), using SD-v1.4 to generate the first frame will decrease the performance compared to the full model. Upon comparing Figure 5(c) with the video generated by the full model,

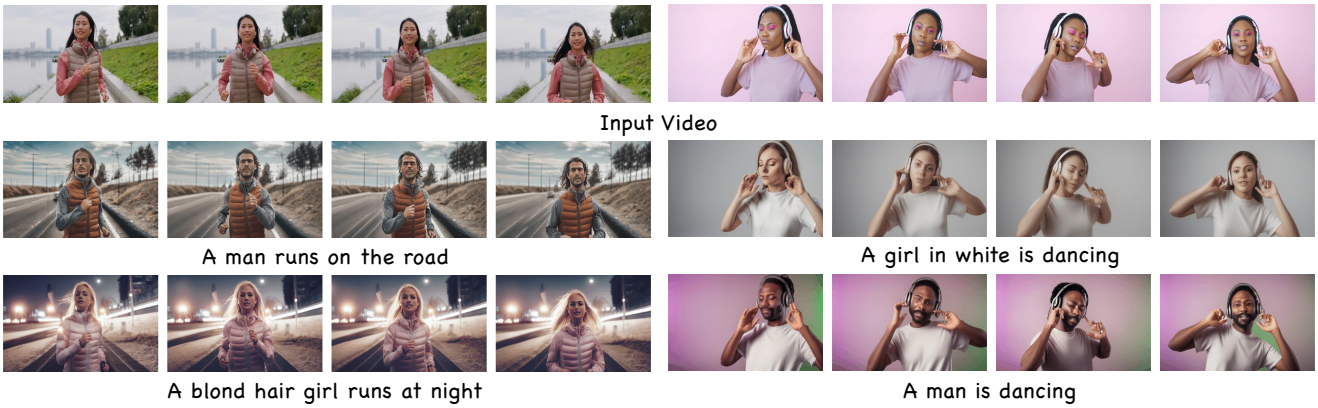


Figure 6. Visual results of our video editing application. **Zoom in for the best view.**

Figure 7. Visual results of LAMP animates the real-world images.

the model without the first-frame-conditioned pipeline produces low-quality results. Notably, the presence of unrelated objects, such as fences and dirt, in the video indicates an overfitting of the content of the video set. In addition, the model w/o shared-noise sampling can generate relatively consistent frames but lacks smoothness in the result. When the temporal-spatial motion learning layers are removed, the model cannot effectively capture the complex motion pattern, leading to failed results. Finally, when we turn video-prediction-based 1D convolution into the original version of 1D convolution, the main object of the video becomes inconsistent. The proposed layer can effectively preserve and propagate the features of the first frame to the subsequent frames, as depicted in the results. These results verify the significant contributions of each key module to the final full model.

5. More Applications

In this section, we provide more applications of the proposed LAMP. While primarily designed for text-to-video generation, our framework can also be used for real image animation and video editing.

5.1. Real Image Animation

Through the training of the proposed first-frame-conditioned pipeline, our LAMP contains a network that predicts the subsequent frames based on the given first frame. This enables the animation of real-world images generated by T2I models. Thus, our method naturally gains the capability to animate real-world images based on the learned motion patterns if these images are placed in the first frame. Figure 7 shows several representative cases in which the 'horse run' model animates a famous horse painting created by Beihong Xu and the 'waterfall' model makes the wonderful Niagara waterfall flow. This application further demonstrates our generalization performance, even when dealing with complex real-world scenes.

5.2. Video Editing

In cases where the given training set contains only a single video clip, our method can only learn a specific motion rather than a motion pattern. In this special case, our method effectively turns into a video editing algorithm. The training process remains similar to that in the few-shot setting. During inference, we adopt the ControlNet [41] based on SD-XL [24] and condition it on canny edges to edit the first frame. DDIM inversion [37] is also used to provide a base motion, thereby ensuring better subsequent-frame prediction. Similarly to video generation, our approach can also take full advantage of image-editing technologies when applied to video editing. As visual examples shown in Figure 6, our LAMP generates photo-realistic videos while maintaining good frame consistency.

6. Limitation and Future Works

In our experiments, we observed that the occurrence of failure cases increased as our method attempted to learn complex motions. More effective modules for motion learning are potential solutions to this issue. Besides, we found that the motion of the foreground object sometimes influences the background's stability. We believe that learning the foreground and background movements independently might be an effective solution. We leave these improvements in our future work.

7. Conclusion

This paper proposes a novel setting, few-shot tuning for T2V generation, which learns a common motion pattern from a small video set to achieve a trade-off between training burden and generation freedom. The proposed LAMP serves as a baseline for this new setting. In our method, we transfer the T2V task into T2I generation for the first frame and predict the subsequent frames. This avoids overfitting the content of the dataset during few-shot tuning while leveraging the advantages of text-to-image techniques. Moreover, our novel design in network architecture and inference strategy further boosts the performance of T2V generation. Extensive experiments demonstrate the effectiveness and generalization capability of our method. We believe that the few-shot tuning setting offers superior trade-offs and will aid the broader T2V field in exploring the lower bounds on the data required for video diffusion training.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 3

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3

[3] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2, 3, 5

[6] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2, 3

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[8] Matej Grcić, Ivan Grubišić, and Siniša Šegvić. Densely connected normalizing flows. *Advances in Neural Information Processing Systems*, 34:23968–23982, 2021. 3

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3, 6, 7

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3

[13] Susung Hong, Junyoung Seo, Sunghwan Hong, Heeseong Shin, and Seungryong Kim. Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[15] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2023. 2, 5

[16] Hanzhuo Huang, Yufan Feng, and ChengShi LanXu JingyiYu SibeiYang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494*, 2023. 3

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 6

[18] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2, 3, 6, 7

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[20] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016. 2, 5

[21] 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. 2, 3

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[24] 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. 2, 3, 5, 6, 8

[25] 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. *arXiv preprint arXiv:2303.09535*, 2023. 2, 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 7

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 3

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 6

[29] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[31] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3

[32] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 3

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 3

[35] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[36] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2, 3, 5

[37] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3, 6, 7, 8

[38] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 3

[39] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 2, 3

[40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 5, 8

[42] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 3