Prompt Image to Life: Training-Free Text-Driven Image-to-Video Generation

Jinxiu Liu[1] Yuan Yao[2,3] Bingwen Zhu[1] Fanyi Wang[1] Weijian Luo[1] Jingwen Su[1]
Yanhao Zhang[1] Yuxiao Wang[1] Liyuan Ma[1] Qi Liu[1] Jiebo Luo[3] Guo-Jun Qi[1,2]
[1] OPPO Research Institute [2] Innopeak Technology [3] University of Rochester

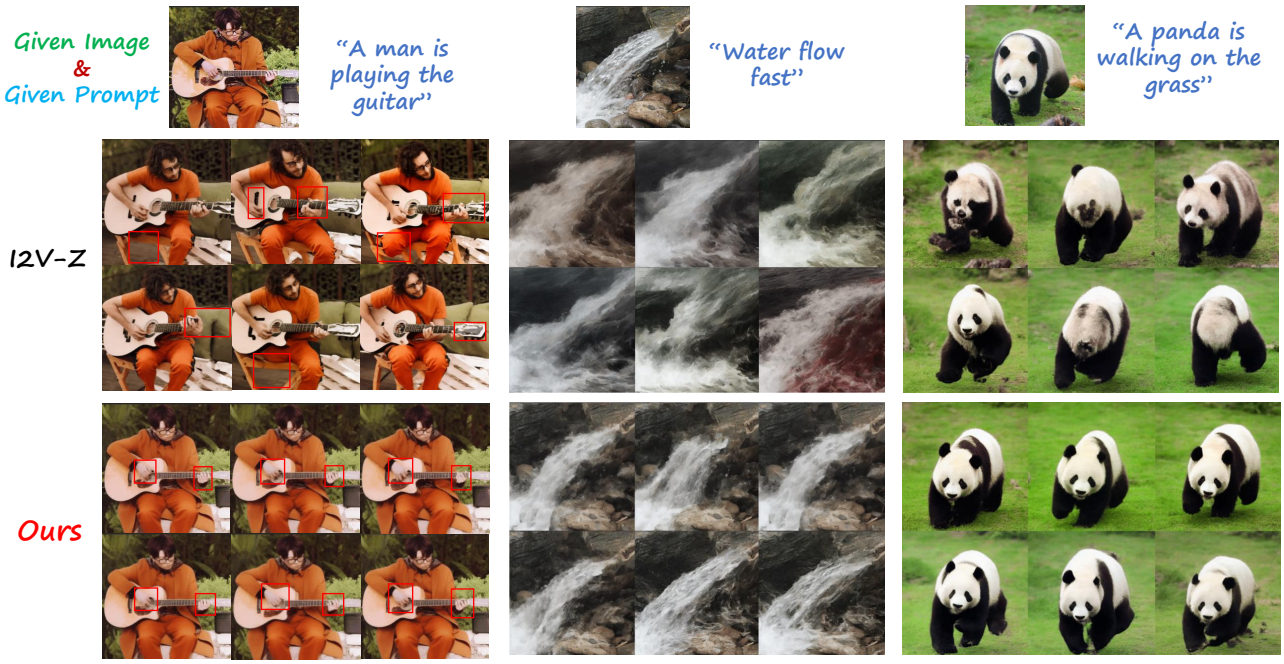

Figure 1. A comparison of PiLife with baseline I2V-Zero method given the same text and image inputs. I2V-Zero is a direct extension of T2V-Zero [16] that accepts both image and text inputs. I2V-Zero suffers from visual collapse and image inconsistency. PiLife solves these issues and outperforms I2V-Zero significantly.

Abstract

*Image-to-video (I2V) generation is a challenging task that requires transforming a static image into a dynamic video according to a text prompt. For a long time, it has been a challenging task that demands both subject consistency and text semantic alignment. Moreover, existing I2V generators require expensive training on large video datasets. To address this issue, we propose PiLife (*Prompt image to Life*), a novel training-free I2V framework that leverages a pre-trained text-to-image diffusion model. PiLife can generate videos that are coherent with a given image and aligned with the semantics of a given text, which mainly consists of three components: (i) A motion-aware diffusion inversion module that embeds motion semantics into the inverted images as the initial frames; (ii) A motion-aware noise initialization module that employs a motion text attention map to modulate the diffusion process and adjust the motion intensity of different regions with spatial noise; (iii) A probabilistic cross-frame attention module that leverages a geometric distribution to randomly sample a frame and compute attention with it, thereby enhancing the motion diversity. Experiments show that PiLife significantly outperforms the training-free baselines, and is comparable or even superior to some training-based I2V methods. Our code will be publicly available.*

1. Introduction

The success of text-to-image generative models, particularly stable diffusion (SD) [28–30], has led to remarkable progress

in text-to-video generation [2, 13, 27, 31, 42]. Despite their achievements, the limited controllability of textual input has spurred a growing trend in the field of image-to-video (I2V) generation, aiming to produce a video sequence given both an image and a textual description [27, 38, 43]. Recent studies on I2V generation [35, 38, 43] attempt to leverage the power of pre-trained SD model by incorporating temporal layers into existing SD models and training these larger models on video and image datasets. While these approaches have displayed promising results, a significant drawback remains their heavy reliance on extensive training with large-scale labeled datasets [9, 39]. This can be prohibitively expensive, limiting the accessibility and development potential of these methods.

In this paper, our objective is to address this challenge through a "training-free" approach for I2V generation. Inspired by Text2Video-Zero [15], our approach utilizes well-trained text-to-image models without the need for fine-tuning image or video data. To achieve this, we aim to modify the Text-to-Image (T2I) diffusion models, injecting image prior into the diffusion process and implementing adjustments to ensure temporal consistency. However, we have found this process not trivial. First, adding image prior can cause the problem of visual collapse. As shown in Fig. 1 (right), directly adding image prior to Text2Video may produce a lot of artifacts. Second, it is hard to align the generated videos with the given image, even with the trained I2V generation models. This can be observed in Fig. 4, where existing I2V generation models generate videos with different styles or subjects from the image. The reasons behind these shortcomings are twofold, illustrated in Fig. 2. Firstly, the image input may not follow the distribution of the diffusion model for generating images, leading to sub-optimal image-wise quality. Secondly, the noise variance across frames is evenly distributed across the entire image, making it challenging for the model to grasp motion semantics.

To this end, we introduce **PiLife**, a novel training-free framework based on a pre-trained T2I diffusion model. PiLife consists of three novel components: (1) A motion-aware diffusion inversion module, which injects motion semantics into the inversion process and takes the inverted images that better reflect the motion semantics as the initial frames; (2) A motion-aware noise initialization module, which harnesses a motion text attention map to guide the DDPM, and obtains a spatial noise intensity distribution that is consistent with the attention map distribution. The spatial noise reflects the motion regions and intensities according to the input text description and it modulates the motion amplitude of different regions. Thus, our model can effectively capture and animate the corresponding parts with appropriate intensity to generate videos that match the text semantics while mitigating visual collapse. (3) A probabilistic cross-frame attention module, which controls T2I diffusion model not only the consistency with the first frame and temporal coherence but also the motion diversity of the generated frames. Overall, by combining these components, our model can generate realistic and diverse videos from texts that capture the motion semantics and dynamics.

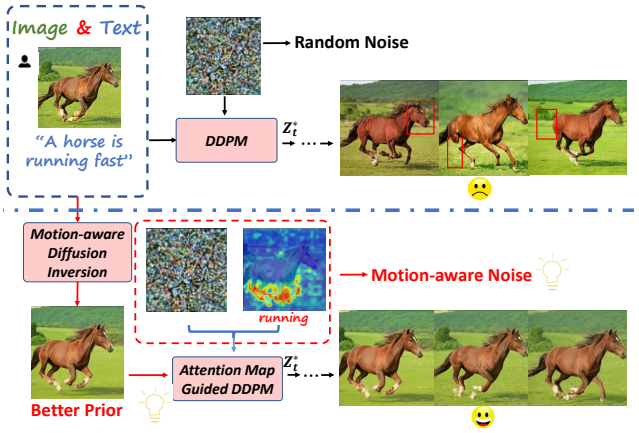

Figure 2. We conclude two factors that affect image-to-video generation qualities. First, image input misalignment with the diffusion model's distribution. Second, the noise variance across frames uniformly distributed across the entire image.

Experiments show that our model can generate high-quality videos that aligns with the given text and image. We evaluate our approach by comparing it with both training-free and training-based I2V methods. Our proposed PiLife outperforms the training-free baseline method significantly, and achieves comparable results or sometimes surpasses the training-based methods in terms of subject accuracy, temporal consistency, and motion diversity.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to introduce training-free framework for I2V generation based on the T2I generation model.
- We introduce a motion-aware diffusion inversion and noise initialization module that improves the DDIM latent code of the diffusion model by incorporating the motion semantics into the diffusion process. This yields more semantically coherent motion and mitigates the visual collapse.
- We devise a probabilistic cross-frame attention module, which balances temporal consistency, subject fidelity, and motion diversity of generated video.
- Experiments demonstrate that our method significantly surpasses the training-free I2V generation baseline and matches or even exceeds the training-based methods in some aspects.

2. Related work

2.1. Image to Video Generation

Image-to-video (I2V) generation is a task of animating a static image based on a text prompt. Many existing methods [4, 19, 21, 25, 26, 36] use a reference driving signal (*e.g.* videos, images) to extract motion, appearance or posture information and guide the generation process. Some GAN-based methods [8, 17] use additional guidance (*e.g.* keypoint [32, 40], optical flow) that is pre-computed or predicted from the original image to perform image translation and generate a video. However, these methods require extra signal guidance, which limits their flexibility and applicability. In the realm of diffusion models, VideoComposer [35] is a multimodality video synthesis approach that can also perform open domain I2V generation, but it requires predefined motion vectors to guide the specific actions in the generated video. I2VGen-XL [43] and VideoCrafter1 [38] are some of the few works that address the text-driven I2V generation challenge. However, these methods need additional large-scale training data, and bring a huge computational burden. Moreover, they have difficulty in generating specific or large movements without prior guidance. In contrast to these methods, we focus on a training-free text-driven image-to-video model, which can generate temporally coherent and large-scale motion videos with high quality.

2.2. Zero-shot Video Synthesis

Zero-shot video synthesis is the task of generating videos from text without any training or optimization. Several previous works have tried to tackle this task using different techniques. One common technique is to use pre-trained text-to-image (T2I) diffusion models [23], which can generate high-resolution images from text prompts using a diffusion process. These models can be extended to generate videos by animating the text prompts with different motion dynamics, such as walking, running, jumping, or flying. For example, Text2Video-Zero [16] introduced a cross-frame attention mechanism to ensure temporal consistency across frames, while Free-Bloom [14] utilized large language models (LLMs) [20] as the director and latent diffusion models (LDMs) [23] as the animator. Duan *et al*. [9] proposed a latent in-iteration deflickering framework and a video deflickering algorithm to reduce the flickering effect. Other approaches [3, 5, 6, 11, 22, 33, 34, 41, 44] explored text-guided zero-shot video editing tasks, which only modify some parts of existing videos. However, these methods cannot animate static objects in an image, which is the focus of our work.

3. Method

In this section, we provide detailed descriptions on the proposed modules, motion-aware diffusion inversion module in Sec 3.2, motion-aware noise initialization module in Sec. 3.3 and probabilistic cross-frame attention module in Sec. 3.4, the overall pipeline are shown in Fig. 3.

3.1. Preliminaries on Diffusion Models

Diffusion models are probabilistic generative models that can produce realistic images from random noise by reversing a stochastic process. We use Stable Diffusion (SD) [23], a powerful latent diffusion model, for text-to-image generation. SD uses an image encoder $\mathcal{E}$ and decoder $\mathcal{D}$ to map an input image $\mathcal{I}$ to a low-dimensional latent code $x_0 = \mathcal{E}(\mathcal{I})$, and then adds Gaussian noise to x_0 gradually through the diffusion forward process: ////

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $t = 1, \ldots, T$, denotes the timesteps, and $\beta_t \in (0, 1)$ is a predefined noise schedule. We can sample x_t at any timestep from x_0 directly using a parameterization trick:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\alpha_t = 1 - \beta_t$. Therefore, noisy data can be obtained through $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The diffusion model uses a neural network ϵ_θ to learn to predict the added noise ϵ by minimizing the mean square error of the predicted noise which writes:

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

For SD, the network takes both a conditional index t and a text-prompt $\mathbf{c}$ to predict the noise $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})$, therefore the SD can generate images that have contents whose semantic meaning aligns with the input text-prompt $\mathbf{c}$. Once the model is trained, we can adopt a deterministic sampling process, called DDIM [29], to iteratively recover $x_0 \sim \mathcal{P}_{data}(x)$ from random noise x_T:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\hat{x}_{t\to 0}}_{\text{predicted } x_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t, \mathbf{c})}_{\text{direction pointing to } x_{t-1}}, \tag{4}$$

where $\hat{x}_{t\to 0}$ is the predicted x_0 at timestep t,

$$\hat{x}_{t\to 0} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t, \mathbf{c})}{\sqrt{\bar{\alpha}_t}}. \tag{5}$$

During the inference phase, we can exploit DDIM sampling [29] to synthesize a denoised representation x_0 from the standard Gaussian noise $x_T = z_T, z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and then decode x_0 into a generated image $\mathcal{I}' = \mathcal{D}(x_0)$ using the pre-trained decoder $\mathcal{D}$. DDIM inversion [7] can perform a deterministic forward diffusion process to recover the latent code $x_t, t = 1, \ldots, T$ from the encoded image x_0.

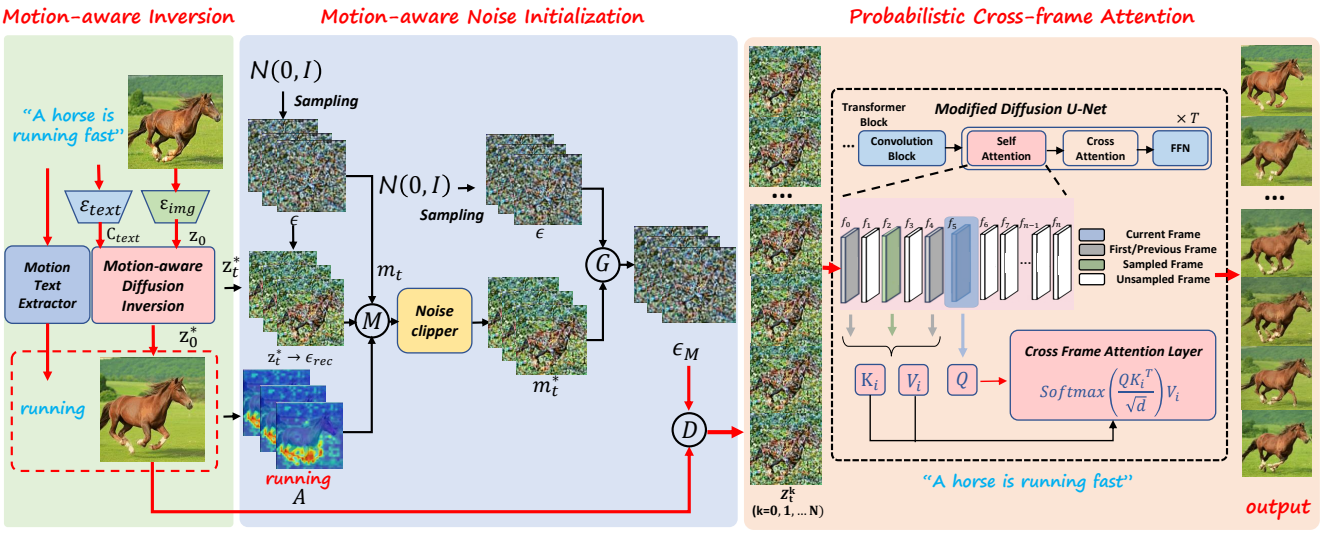

Figure 3. Our PiLife framework consists of three main modules. (1) The motion-aware diffusion inversion module processes the input text and image to obtain the initial frame with motion information. (2) The motion-aware noise initialization module generates the noise with motion features and initial latent code. $\mathbb{Z}_0^* \rightarrow \epsilon_{rec}$ is formulated in Eq. 8, 9. M, G, D are formulated in Eq. 10, 12, 7, respectively. (3) The probabilistic cross-frame attention layer replaces the self-attention layer to enhance motion features.

3.2. Motion-aware Diffusion Inversion

In the context of the I2V task, we have observed that the quality of the image prior plays a vital role in the overall quality of the generated videos. If the input image is of low quality or falls out of the typical SD image distribution, it can directly lead to videos that suffer from reduced image fidelity. Furthermore, the image prior may not be aligned with the motion semantics conveyed in the text prompt. This misalignment poses a challenge for the Diffusion model in producing videos that correspond to the motion semantics outlined in the text prompts. Consequently, our objective is to reconstruct the input image that not only maintains the visual characteristics of the original image but also conveys the motion-related semantics specified in the accompanying text description.

Inspired by the null-text inversion [18], we propose a novel motion-aware diffusion inversion module that incorporates motion features into the denoising network. We optimize the conditional embedding of the target text by minimizing the diffusion loss. Formally, we have the optimized object

$$\psi_t^* = \text{argmin}_\psi \, \mathcal{E}(\epsilon_\theta(Z_t^*, t, \psi_t), Z_{t-1})), \tag{6}$$

where ψ_t is the input embedding for DDIM sampling at tth timestamp, and ψ_t^* is the optimized embedding at the t-th timestamp, $\mathcal{E}(\cdot)$ is the reconstruction error, Z_{t-1} is the diffusion trajectory of DDIM inversion [12], Z_{t-1}^* is the intermediate latent code of the DDIM sampling process with Z_t^* and optimized embedding ψ_t^* as the input.

By using the optimized embedding from the motion text embedding as the unconditional embedding input for the denoising step, we can obtain a reconstructed image prior that better preserves the appearance details and reflects the motion semantics of the text. This image prior serves as the first frame in the generated video. As shown in Fig. 3, the image after motion-aware diffusion inversion provides a motion-aware prior for both the latent code and the cross-frame attention of the subsequent frames.

3.3. Motion-aware Noise Initialization

To prevent static image regions from distortion by motion, we propose the motion-aware noise initialization module. It has two main components: (1) Attention-guided DDPM, which can disentangle dynamic and static parts and adjust the motion by incorporating motion semantic attention scores into DDPM [12] to get different noise intensity. (2) Noise clipper, which confine the noise within the scope of the attention map and constrains the L2 norm of the noise, thereby reducing the noise discrepancy between frames that leads to visual collapse.

Attention-guided DDPM To address flickering artifacts and visual collapse in static regions caused by motion, we disentangle static and dynamic regions. This is achieved by introducing a motion attention score (denoted as A) that reflects the significance of each pixel to the dynamic semantic correlation, into the diffusion process of DDPM [12]. Attention-guided DDPM models the spatially dynamic trans-

formation of the initial latent code by adjusting the noise intensity of the static and dynamic regions. Specifically, as explained in Sec. 3.1, the initial latent code z_t is obtained by reparameterizing the diffusion process as follows:

$$z_t^* = \sqrt{\bar{\alpha}_t} \cdot z_0^* + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_{\mathbf{M}}, \epsilon_M \sim N(0, I) \quad (7)$$

where z_0^* and z_t^* denote the original image latent codes and initial latent codes inverted to the t-th step, respectively. We consider the inverted noise ϵ_{rec} to represent the random noise that needs to be input when adding noise to the original image during the DDPM reparameterization process. Adding this noise can obtain the current step's initial latent code, which is the inverse operation of reparameterization.

$$\epsilon_{rec} = \frac{z_t^* - \sqrt{\bar{\alpha}_t} \cdot z_0^*}{\sqrt{1-\bar{\alpha}_t}} \quad (8)$$

ϵ_{rec} may not follow $\epsilon_{rec} \sim N(0, I)$. To ensure the relevance of the noise predicted by the UNet and the structural information of the original image, we adjust it proportionally according to the variance and map it to the $N \sim (0, 1)$ distribution:

$$\epsilon_{rec}^* = \frac{\epsilon_{rec} - \mu}{\sigma} \quad (9)$$

The attention-guided noise m_t is then defined as:

$$m_t = A \odot \epsilon + \sqrt{1-A^2} \odot \epsilon_{rec}^* \quad (10)$$

where $\odot$ denotes element-wise multiplication. The motion attention score A is computed by,

$$\mathbf{A} = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}_m}{\sqrt{d}}) \quad (11)$$

The deep spatial features of the noisy image $\phi(\mathbf{z}_t)$ are projected to a query matrix $\mathbf{Q} = \mathcal{L}_Q(\phi(\mathbf{z}_t))$. To extract the motion text from the given sentence, we use a pre-trained text classifier to identify the words that describe the motion semantics. We constructed a small dataset and finetuned a bert to recognize the tokens that indicate motion text in the sentence. The motion text embedding $\psi(\mathbf{P})$ is projected to a key matrix $\mathbf{K} = \mathcal{L}_K(\psi(\mathbf{P}))$ via learned linear projections $\mathcal{L}_Q$ and $\mathcal{L}_K$.

The attention-guided noise m_t can maintain the variance of the original noise ϵ in the regions where the attention map A is high, and decrease the variance of the noise in the regions where the attention map A is low. In this way, the regions of the initial latent code that are relevant to the dynamics of the video have higher noise variance, which can generate more diverse motions, while the regions that are irrelevant to the dynamics have lower noise variance, which can preserve the features of the original image and reduce the impact of the classifier guidance on these regions.

To control the influence intensity of the attention-guided noise, we use a Gaussian mixture to fuse the attention-related noise and the random noise and introduce a parameter π to control the relative weights of the two components. The Gaussian mixture is as follows:

$$\epsilon_M = \frac{1}{\sqrt{1+\eta^2}} \cdot \epsilon + \frac{\pi}{\sqrt{1+\eta^2}} \cdot m_t \quad (12)$$

where $\eta > 0$. When $\eta \to 0$, $\epsilon_M \approx \epsilon$, when $\eta \to \infty$, $\epsilon_M \approx m_t$.

Algorithm 1 Clipping technique for noise generation

Require: Initial latent code ϵ_{rec}

1: Sample a noise signal ϵ_M with the same length and parameters as ϵ_{rec}^*
2: Define $\|\epsilon\|_2 = \sqrt{\sum_{i=1}^{N} \epsilon_i^2}$ as the L2 norm of ϵ
3: Compute $\Delta\|\epsilon\|_2 = \|\epsilon_M - \epsilon_{rec}\|_2$ and $\|\epsilon_M\|_2$
4: Set $\gamma \leftarrow \epsilon_M$
5: **while** $\Delta\|\epsilon\|_2 > \|\epsilon_M\|_2$ **do**
6: Sample a random parameter t from a uniform distribution on the interval $[\eta, 1-\eta]$, where η is a small positive constant
7: Update $\gamma := (1-t) \cdot \epsilon_{rec} + t \cdot \gamma$
8: Recalculate $\Delta\|\epsilon\|_2 = \|\gamma - \epsilon_{rec}\|_2$
9: Return γ

Noise Clipper To reduce the excessive deviation of the initial latent code from the original distribution and prevent visual degradation, we propose a clipping technique to limit the L2 norm difference between the noise and the inverted noise. We linearly interpolate the noise signals until the difference is below a threshold if it exceeds the noise norm. The clipping technique is shown in Algorithm. 1, which ensures that the noise signals stay close to the original distribution and retain the features of the initial latent code. This improves the quality and diversity of the video generation and prevents visual collapse.

3.4. Probabilistic Cross Frame Attention

We introduce probabilistic cross-frame attention (PCFA) to enhance the temporal consistency of our model. Specifically, we replace the self-attention layers in the original SD model by calculating the cross-attention score of the current frame with the first frame, the previous frame, and a randomly sampled frame based on a geometric distribution. PCFA balances the strength of cross-frame attention between different frames based on their relative distance from the current frame t. This way, the PCFA achieves three key objectives: (1) it maintains consistency with the given first frame; (2) it keeps continuity with the previous frame; (3) it brings in motion diversity by sampling frames according to a probabilistic distribution. The details are as follows.

We use geometric distribution to measure attention strength, defined as:

$$P(k) = (1-p)^{k-1} \cdot p \tag{13}$$

where k is the distance from the current frame, and p is the probability of sampling a frame. The distribution has the property that the sampling probability decreases with the distance, which is desired for applying stronger attention to frames near the source frame. PCFA adjusts the self-attention value according to the diffusion timestep. This prevents visual collapsing and ensures temporal consistency. The distribution also introduces a tunable parameter p, which controls the temporal smoothness and diversity of the video generation. Intuitively, larger p leads to more diverse motions, while smaller p leads to smoother motions. The PCFA maintains the motion continuity by probabilistically assigning attention to previous frames.

Besides the probabilistic frame sampling, we also sample the first frame for cross-frame attention to maintain appearance consistency with the original image. The set of sampled frame indices for the current frame t is:

$$S_t = \{1, t-1\} \cup \{t-k | k \sim P(p)\} \tag{14}$$

where 1 is the first frame index, $t-1$ is the previous frame index, and $t-k$ is sampled from the distribution (13). To avoid future information leakage, we only sample frames that are causally before the current frame. Overall, the Probabilistic Cross Frame Attention (PCFA) is then defined as:

$$\begin{aligned} \text{PCFA}(Q_t, S_t) &= \text{Attention}(Q_t, [K_j | j \in S_t], [V_j | j \in S_t]) \\ &= \text{Softmax}(\frac{Q_t \cdot [K_j | j \in S_t]^T}{\sqrt{d_k}}) \cdot ([V_j | j \in S_t]) \end{aligned} \tag{15}$$

where Q_t, K_t, and V_t are the query, key, and value vectors of the current frame t, respectively, K_1 and V_1 are the key and value vectors of the first frame, K_{t-1} and V_{t-1} are the key and value vectors of the previous frame. d_k is the dimension of the key vector.

4. Experiments

4.1. Experiment Settings

We take the pre-trained Dreamlike Photoreal v2.0 [1] as the basis diffusion model, which is a photorealistic model based on Stable Diffusion 1.5, specializing in generating real-world images. In our experiments, we generate $m = 6$ frames with 512×512 resolution for each video. However, our framework allows generating any number of frames by increasing m. The initial time step of our generated latent code is 761 for all instances. All experiments are performed on one Tesla V100 (32GB).

4.2. Comparison with Other Methods

We compare the performance of PiLife with three existing methods. One of them is the baseline method Text2Video-Zero [16], which is a well-known training-free video generation method. We adapted this method to the I2V task and named it I2V-Zero, which we implemented ourselves by closely following the official one since the official implementation of T2V-Zero does not support the I2V task. We also compare our PiLife with the state-of-the-art training-based methods, namely I2VGen-XL [43] and VideoCrafter1 [38], to provide a comprehensive comparison of PiLife with both training-based and training-free methods.

4.2.1 Qualitative Results

We show some results of our method in Fig. 1 and Fig. 4, and compare them qualitatively with the baseline methods. Our method aims to (i) address the issue of visual collapse and (ii) retain more details of the source image.

Comparing our results with I2V-Zero [16], as shown in Fig. 1, we demonstrate that our methods solved the problem of visual collapse. For instance, in the case of *walking panda*, our method has less distortion and more stability on the frame changes. We also observe that our method disentangles the motion semantics of the dynamic part from the static part while preserving the static features from being influenced by redundant motion, *i.e.*, the panda's feet move while the rest of its appearance remains undistorted. Such advantage can be attributed to the motion-aware diffusion inversion design, which provides a better understanding of moving and static parts. This result illustrates the accuracy and consistency of our method in capturing the motion semantics of the dynamic and static parts.

Furthermore, we compared our method with the latest training-based image-to-video models, I2VGen-XL [43] and VideoCrafter1 [38]. It is noteworthy that I2VGen-XL fails to generate high-quality videos that were consistent with the given image in all the three cases shown in Fig. 4, supporting the claim that image-to-video generation is a challenging task, even with a training-based model. Although VideoCrafter1 [38] and our method could both generate videos with good temporal consistency and image quality, our method achieved better alignment with the input image. Such advantage can be easily observed in the case of *a girl playing the piano*, where the images generated by I2VGen-XL [43] are completely inconsistent with the given image. The frames generated by VideoCrafter1 [38] are closer to the original image, but still far from being as similar as ours. We think such an unsatisfactory performance is caused by the limitation of the training-based I2V model's insensitivity to data that deviates from the video generation model distribution. On the contrary, our model accurately captures the motion semantics from the given text by motion-aware

Figure 4. Qualitative comparison of our method and the training-based methods I2VGen-XL [43] and VideoCrafter1 [38].

noise initialization. For the piano-playing action, the arm movement amplitude and the finger movement amplitude are more obvious than those of the other two training-based methods without visual collapse, which demonstrates that our model can accurately capture the motion semantic and make a balance between the subject fidelity, the temporal consistency and the motion diversity of the generated frames, rather than simple translation.

4.2.2 Quantitative Results

In this part, we numerically evaluate I2V methods. We use both objective metrics and user studies to measure the quality of our generated videos. We consider the following three aspects: **(i) Subject fidelity**: the generated videos should contain the given subjects. We use the pre-trained FasterRCNN-MobileNet-V3-large model to detect the subjects in each frame and calculate the DINO score [24] between the detected subjects and the given subjects. **(ii) Textual fidelity**: the generated videos should be consistent with the given textual prompt. We use the average CLIP-T [10, 24] score between each frame and the prompt to measure the textual fidelity. **(iii) Temporal consistency**: the generated videos should be smooth and coherent. We use the average CLIP image cosine similarity between all pairs of frames to measure the temporal consistency [37]. We also present a user study, where the expert participants are asked to rate the fidelity, temporal coherence, and semantic coherence on a scale of 1 to 5. The users are also asked to rank the videos based on their personal preferences.

The results are shown in Table. 1. PiLife significantly outperforms the baseline method I2V-Zero [16] on all the metrics, which demonstrates the effectiveness of our model. Our training-free method even surpasses the other two training-based methods on subject fidelity, which demonstrates the superiority of our model in preserving the original subjects in video generation. Moreover, our method achieves comparable performance with VideoCrafter1 [38] on CLIP-T

Method	Training-Free	Objective Metric			User Study			
		DINO ↑	CLIP-T ↑	Temporal Consistency↑	Fidelity ↑	Semantic ↑	Temporal ↑	Rank ↓
VideoCrafter1	×	0.513	0.329	0.937	4.6	3.624	3.771	**1.875**
I2VGen-XL	×	0.426	0.301	0.927	4.1	3.247	3.584	2.031
I2V-Zero	✓	0.451	0.287	0.781	2.7	2.958	2.238	3.762
Ours	✓	**0.556**	**0.321**	**0.923**	**3.9**	**3.598**	**3.385**	**2.332**

Table 1. Quantitative results of training-based or training-free methods on image to video generation task. All metrics are average evaluations.

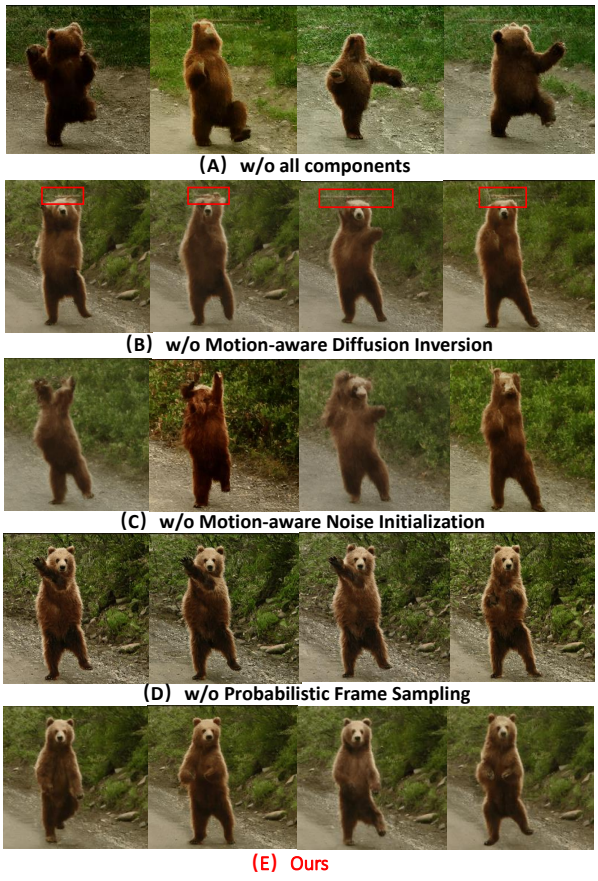

Figure 5. Ablation study. The text prompt is "A bear is dancing happily", and the given image can be seen in Fig. 4.

and temporal consistency, though the latter is trained on large-scale video datasets.

4.3. Ablation Study

We conducted an ablation study to evaluate the effectiveness of each component of our framework. We choose the case "A bear is dancing happily" as the text prompt for the image-to-video generation task. The results are shown in Fig. 5.

w/o all components For the baseline model without all our proposed components, the generated videos suffered from severe visual collapse, indicating that controlling the frame consistency with the original image is challenging for diffusion-based video generation methods.

w/o motion-aware diffusion inversion Without motion-aware diffusion inversion, the generated videos produce more artifacts and are less aligned to the text prompt. In this example, the bear's head deviates from the original image details and hardly moves across the frames. This demonstrates the importance of injecting motion information in the inversion process, which provides a better prior first frame for appearance preservation and motion amplification of the subsequent frames.

w/o Motion-aware Noise Initialization We observed that the generated frames suffered from visual collapse and discontinuity when lacking motion-aware noise initialization. This demonstrates that the motion-aware noise initialization module can capture the parts with large motion magnitudes, preserve the non-motion parts and enhance cross-frame stability.

w/o Probabilistic Frame Sampling We found that when not performing probabilistic frame sampling, the generated frames had monotonous and small motions, and lacked diversity. Probabilistic frame sampling can increase the temporal diversity of the generated videos.

5. Conclusion

We propose PiLife in this paper, a novel training-free framework for image-to-video generation with text guidance. PiLife leverages the power of diffusion models and introduces three key innovations: motion-aware diffusion inversion, motion-aware noise initialization, and probabilistic cross-frame attention. These innovations enable PiLife to generate high-quality videos that are consistent with the input image and text, while diverse in motion patterns. We demonstrate the effectiveness of PiLife through extensive experiments and comparisons with existing methods. Experimental results also show that PiLife can generate videos for various scenarios, such as human actions, animal movements, and natural phenomena. PiLife opens up new possibilities for image-to-video generation and paves the way for future research in this direction.

References

[1] https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0. 6

[2] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 2

[3] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 3

[4] Chia-Chi Cheng, Hung-Yu Chen, and Wei-Chen Chiu. Time flies: Animating a still image with time-lapse video as reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5641–5650, 2020. 3

[5] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023. 3

[6] Paul Couairon, Clément Rambour, Jean-Emmanuel Haugeard, and Nicolas Thome. Videdit: Zero-shot and spatially aware text-driven video editing. *arXiv preprint arXiv:2306.08707*, 2023. 3

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[8] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1161–1170, 2019. 3

[9] Zhongjie Duan, Lizhou You, Chengyu Wang, Cen Chen, Ziheng Wu, Weining Qian, Jun Huang, Fei Chao, and Rongrong Ji. Diffsynth: Latent in-iteration deflickering for realistic video synthesis. *arXiv preprint arXiv:2308.03463*, 2023. 2, 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 7

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

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[13] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[14] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494*, 2023. 3

[15] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690, 2023. 2

[16] 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. 1, 3, 6, 7

[17] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 600–615, 2018. 3

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 4

[19] Makoto Okabe, Ken Anjyo, Takeo Igarashi, and Hans-Peter Seidel. Animating pictures of fluid using video examples. In *Computer Graphics Forum*, pages 677–686. Wiley Online Library, 2009. 3

[20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 3

[21] Ekta Prashnani, Maneli Noorkami, Daniel Vaquero, and Pradeep Sen. A phase-based approach for animating images using video examples. In *Computer Graphics Forum*, pages 303–311. Wiley Online Library, 2017. 3

[22] 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. 3

[23] 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. 3

[24] 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. 7

[25] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 3

[26] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 3

[27] 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

[28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

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

[30] 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. 1

[31] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[32] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*, pages 3332–3341, 2017. 3

[33] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 3

[34] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3

[35] 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

[36] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 3

[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. 7

[38] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 2, 3, 6, 7

[39] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023. 2

[40] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 3

[41] 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. 3

[42] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 2

[43] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2, 3, 6, 7

[44] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3