# PIA: Your Personalized Image Animator via Plug-and-Play Modules in Text-to-Image Models (Supplemental Material)

Yiming Zhang[1,2,*] Zhening Xing[1,*] Yanhong Zeng[1] Youqing Fang[1] Kai Chen[1,†]

[1]Shanghai Artificial Intelligence Laboratory [2]Dalian University of Technology

https://github.com/open-mmlab/PIA

## 1. Implementation Details

### 1.1. Training

We train PIA on WebVid10M [1] with only condition module and temporal alignment layers trainable. Specifically, we use the motion module in AnimateDiff [4] as a pre-trained model for temporal alignment layers. We compute the L1 distance between the condition frame and other frames in HSV space. Subsequently, we utilize this distance to calculate the affinity score. Specifically, we consider the top 2.5th percentile of the samples as the minimum and the 97.5th as the maximum value to linearly scale the affinity score to $[0.2, 1]$. In addition, we set a probability of $20\%$ to zero out the entire input of the condition module. This ensures that PIA retains text-to-video capabilities and promotes the training of the condition module. We train condition module and temporal alignment layers on 16 NVIDIA A100s for 4.5k steps and use a learning rate of $1 \times 10^{-5}$.

### 1.2. Inference

During the inference stage, users can replace the base model with the personalized T2I model to realize the image animation. Besides, we construct the inter-frame affinity according to the affinity score obtained from the training stage. We design three affinity ranges for three different amplitude motions. The maximum value of all three affinity ranges is 1, achieved at the conditional frame. The minimum values are 0.2, 0.4, and 0.8, respectively, with corresponding decreases in motion magnitude. We use classifier-free guidance during the DDIM process [11] and set the classifier-free guidance [6] scale as 7.5. A 512 × 512 image can be animated in around 13.8 seconds (using 25 denoising steps with classifier guidance) on a single A100 GPU.

## 2. AnimateBench

AnimateBench is a comprehensive benchmark, which consists of 105 image and prompt pairs. To cover a wide variety of contents, styles, and concepts, we choose seven base models [9] and LoRA [7]. An example case of AnimateBench is depicted in Fig. 1. We have released AnimateBench in https://huggingface.co/datasets/ymzhang319/AnimateBench.

Base model: MajicMix RealisticVision
Prompts:
1. 1girl is smiling, lowres,watermark
2. 1girl is crying, lowres,watermark
3. 1girl is blinking, lowres,watermark
Negative prompt: wrong white balance, …

Figure 1. **AnimateBench case.** Each curated personalized image corresponds to a personalized text-to-image model and three tailored motion-related text prompts.

### 2.1. Images in AnimateBench

We carefully choose seven of the most popular base models [9] and LoRAs [7] in Cvitai [3]. Each personalized model has very distinct styles and we use them to curate images with impressive high quality by tailored text prompts for image generation. Specifically, these images differ in styles, contents, and concepts and ensure that AnimateBench covers three categories: people, animals, and landscapes.

### 2.2. Prompts in AnimateBench

For each generated image, we design three prompts describing different motions to test the text alignment ability of models. Prompts are mainly composed of three parts: the **subject**, the **motion descriptor**, and the **trigger words**. Subject and motion descriptors specify the content of motion in the generated videos. The trigger word is a well-known technique that is able to activate the DreamBooth or LoRA to generate personalized effects [3]. Only when these prompts are included during inference, DreamBooth or LoRA can achieve optimal performance. Then we can

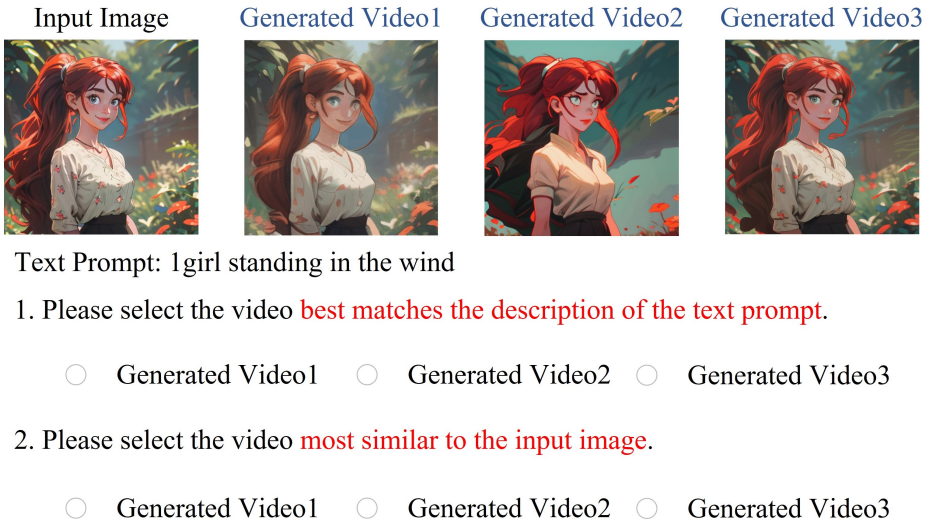


Figure 2. **User Study.** Example of user study questionnaires.

get the complete prompt in AnimateBench. For example, we use '1girl is smiling, white hair by atey ghailan, by greg rutkowski, by greg tocchini.' to generate a personalized image, and then we can get the complete prompt as *'1girl is smiling, white hair by atey ghailan, by greg rutkowski, by greg tocchini'*. In this case, *'1gril'* represents the **subject**, *smiling* represents the **motion descriptor**, and *'white hair by atey ghailan, by greg rutkowski, by greg tocchini'* represents the **trigger word**. We also distinguish motion between different types of subjects. For example, the prompt of people or animals contains more descriptors such as smiling, crying, etc, while the prompt of landscapes or scenes contains more like raining, lightning, etc.

# 3. Evaluation Details

## 3.1. CLIP Score

Following previous work [2, 4, 12], we compute CLIP score to quantitatively evaluate the alignment in generated videos. In addition to calculating text alignment, we measure image alignment by computing the similarity between the embeddings of the generated video frames and the input images. The two average CLIP scores are calculated on AnimateBench which contains 1680 frames. We leverage the code provided by [5] and use ViT-B/32 [8] model to extract the embedding of images and prompts.

## 3.2. User Study

For user study, we randomly select input image and prompt pairs in AnimateBench and then generate videos by using PIA, VideoComposer [12] and AnimateDiff [4, 10] with ControlNet [14] and IP-Adapter [13]. We ask the participants to choose from the three generated videos with the best image alignment or text alignment in each question. We show an illustration of the question cases in Fig. 2. There are 20 questions in total, and the order of options is shuffled. A total of 20 participants were involved in the survey. Following the previous work [2], we calculated the preference rate, and the results are shown in the main paper.

# 4. Ablation

In this section, we introduce more ablation studies to verify the effectiveness of the inter-frame affinity and the fine-tuning of temporal alignment layers.

## 4.1. Inter-frame Affinity

To further verify the effectiveness of inter-frame affinity, we train a model without affinity hints for ablation study. We remove the affinity channel from the input of the condition module and the result is shown in Fig. 3. Compared to our method, videos generated by the model without inter-frame affinity are more incoherent and change suddenly.

## 4.2. Fine-tuning Temporal Alignment Layers

In the training stage, we train both the condition module and temporal alignment layers. We now show the result of only training the condition module with temporal alignment layers frozen in Fig. 4. The result proves that the frozen temporary alignment layers failed to align the condition frame with other frames.

# 5. Visualization Results

## 5.1. Visualization of Attention Map

To demonstrate that the motion alignment of PIA is better than other methods, we visualize the average cross attention

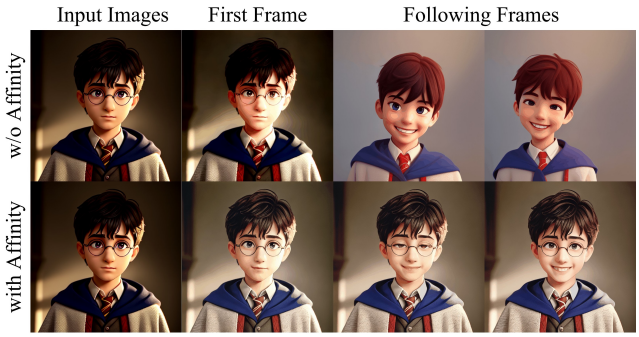


Figure 3. **Ablation study for Inter-frame Affinity.** Without an affinity hint, the generated videos become incoherent and may change significantly after the first given frame. With the inter-frame affinity as inputs, PIA is able to animate images that are faithful to the condition frame.


Figure 4. **Ablation study for fine-tuning the Temporal Alignment Layers.** Pre-trained temporal alignment layers fail to align the condition frame in generated videos. PIA fine-tunes both the condition module and the temporal alignment layers, leading to better preservation of the information in the condition frames.

map of **motion descriptor** token. We use prompt *'the rabbit on is on fire'* as an example and visualize the cross attention map corresponding to token *'fire'*, as shown in Fig. 5. We can observe that in our method, the region attended by the *'fire'* matches the region of flames. In contrast, the motion descriptor token in the baseline method randomly attends to the entire context and cannot form a meaningful structure. This phenomenon demonstrates that our method exhibits better motion alignment performance.

## 5.2. PIA with complex prompts

The temporal layers of PIA focus more on motion-related alignment which leading to improved motion controllability. Therefore, PIA is capable of responding to complex motion descriptions in prompts. We include composite animations in Fig. 6.

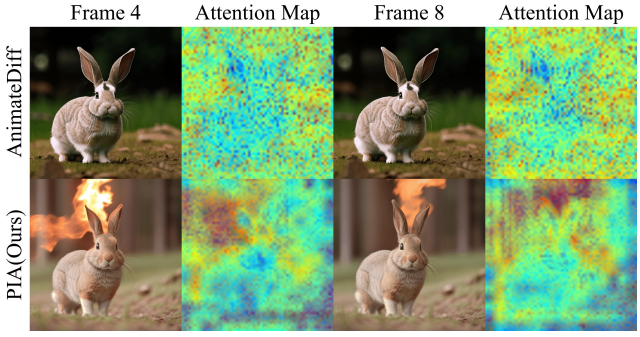


Figure 5. **Visualization of Cross-attention map.** We generate the video using prompt *'a rabbit is on fire'* and visualize the cross-attention map corresponding to the token *'fire'* for both AnimateDiff [4] and our own method. In PIA, token *'fire'* shows more accurate attention to the shape of flames, while in AnimateDiff, the token randomly attends to the entire context. This demonstrates the superior motion alignment performance of our method.


Figure 6. **Animation with complex prompts.** PIA achieves improved motion controllability. Even with the complex text prompt, PIA can correspondingly generate composite animations.

## 5.3. PIA for Open-Domain Images

In this section, we further explore animating open-domain images with PIA without using personalized T2I models. To further enhance the preservation of the information and details of the condition frame, we combine the Image Prompt Adapter (IP-Adapter) [13] with PIA. Specifically, we use a CLIP image encoder to extract features from the input images. Then, these image features are incorporated into each frame through the cross-attention mechanism in the UNet. As shown in Fig. 7, without using personalized models, our model successfully animates an open-domain image with realistic motion by text while preserving the identity of the given image.


Figure 7. **Using PIA to animate open-domain images.**Without providing personalized T2I models, PIA is able to animate the open-domain images with realistic motions by text while preserving the details and identity in condition frame with IP-Adapter[13]

## 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. 1

[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] CIVITAI, 2022. https://civitai.com/. 1

[4] 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. 1, 2, 3

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[7] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 1

[8] 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. 2

[9] 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. 1

[10] s9roll7, 2023. https://github.com/s9roll7/animatediff-cli-prompt-travel. 2

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[12] 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

[13] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 2, 3, 4

[14] 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. 2