CustomTTT: Motion and Appearance Customized Video Generation via Test-Time Training (*Supplementary Materials*)

Anonymous submission

1 Implementation Details

We conduct all the training, including the spatial LoRA, temporal LoRA, and test-time training, using the Lion optimizer (Chen et al. 2024), with default betas set to 0.9 and 0.99, and the weight decay of 0.1. For training the spatial LoRA, we use a learning rate of $1e-5$ for 500 steps, and we do not add the motion modules in the model to focus the model solely on appearance, thus eliminating motion influence. For training the temporal LoRA, we use a learning rate of $5e-5$ for 500 steps. During the test-time training phase, we learn 30 steps with a learning rate of $1e-6$. Similarly, for the appearance-guided reference UNet, we do not use the motion module for sampling. For the appearance and motion-guided U-Net, we utilize DDIM sampling with 5 steps to obtain the reference latents. Additionally, we set a dropout of 0.1 and a LoRA rank of 32 during LoRA training. To conserve VRAM, we employ mixed precision training with FP16. During inference, we use DDIM (Song, Meng, and Ermon 2021) sampling for 25 steps and a classifier-free guidance (Ho and Salimans 2022) scale of 9. We generate 16-frame videos at a resolution of 256×256 pixels and 8 fps. All experiments are conducted on a single A6000 GPU. Under this experimental setup, training the spatial LoRA takes 4 minutes, training the temporal LoRA takes 4 minutes, and test-time training takes 3 minutes.

2 Additional Results

2.1 More Quantitative Metrics

Motion Fidelity To evaluate the consistency of the motion between the generated videos and the reference videos, we use the Motion Fidelity Score introduced by (Yatim et al. 2024), which is based on the similarity between unaligned long trajectories and accounts for structural deviations between the reference and generated videos. It relies on an off-the-shelf tracking method (Karaev et al. 2023) to estimate $\mathcal{T} = \{\tau_1, \ldots, \tau_n\}, \tilde{\mathcal{T}} = \{\tilde{\tau}_1, \ldots, \tilde{\tau}_m\}$, using the similarity between the trajectories of the generated video and the original video to measure motion consistency. The calculation method is as follows:

$$\frac{1}{m}\sum_{\tilde{\tau}\in\tilde{\mathcal{T}}}\max_{\tau\in\mathcal{T}}\operatorname{corr}(\tau,\tilde{\tau}) + \frac{1}{n}\sum_{\tau\in\mathcal{T}}\max_{\tilde{\tau}\in\tilde{\mathcal{T}}}\operatorname{corr}(\tau,\tilde{\tau}), \quad (1)$$

Table 1: We extend Tab.1 in the main paper with an additional *Motion Fidelity* score to show the motion similarity between each method and the original reference video.

Method	Train-able parameters	CLIP-T	CLIP-I	Temporal consistency	Motion fidelity
Full LoRA	28.26M	<u>0.294</u>	0.687	0.977	**0.746**
DreamVideo	85M	0.271	0.681	<u>0.969</u>	0.563
MotionDirector	<u>21.26M</u>	0.269	<u>0.690</u>	0.965	0.649
DiffDirector	30.46M	0.287	0.685	0.971	0.693
Ours	**12.12M**	**0.301**	**0.712**	**0.978**	<u>0.721</u>

where $\operatorname{corr}(\tau, \tilde{\tau})$ indicates the normalized cross-correlation between tracklets τ from the reference video and $\tilde{\tau}$ from the generated video.

As shown in Table 1, our method achieves better motion consistency across all approaches except for Full LoRA. We argue that the full LoRA training shows obviously artifacts in terms of the CLIP-T, CLIP-I, *etc*., as shown in the visual comparisons.

2.2 More Ablation Results

In this section, we conduct a numerical analysis of the ablations mentioned in the main paper. A total of 50 videos are generated and divided into ten groups based on combinations of 9 different objects and 10 different motions. Each group contains 5 text prompts with varying scenes.

Finetune LoRAs on the Specific Layers we evaluate the impact of training LoRAs at different layers using numerical results. To validate, we choose two layers *i.e*., $\Delta W_s^{0,1}$ and $\Delta W_t^{1,7}$ as the comparison. As shown in Table 3, training LoRAs on these layers fails to achieve appearance or motion customization. Fine-tuning the LoRAs at $\Delta W_t^{2,5}$ or $\Delta W_s^{2,6}$ individually achieves either motion or appearance customization. Moreover, fine-tuning $\Delta W_s^{2,6}$ and $\Delta W_t^{2,5}$ together ensures not only high video quality and visual consistency with the reference image but also shows optimal motion consistency between the generated video and the reference video, thereby achieving comprehensive customization.

TTT Reference Sampling Step We analyze the impact of directly combining the two LoRA components, adding test-time training (TTT), and adjusting the sampling steps of reference latents during TTT, as shown in Table 4, Compared

Table 2: CLIP-T score over 100 prompts to show the significance of different UNet layers in determining appearance (subject) and motion characteristics.

Replace in i-th layer	0	1	2	3	4	5	6	7	8
Subject prompt replacement	0.109	0.109	<u>0.153</u>	0.109	0.112	0.109	**0.200**	0.108	0.108
Motion prompt replacement	0.232	0.231	**0.239**	0.231	<u>0.237</u>	0.231	0.234	0.231	0.232

Table 3: Ablation results for fine-tuning LoRA at different layers.

LoRA Layers	CLIP-T	CLIP-I	Temporal consistency	Motion fidelity
$\Delta W_s^{0,1}\&\Delta W_t^{1,7}$	0.294	0.655	0.978	0.610
$\Delta W_s^{2,6}\&\Delta W_t^{1,7}$	**0.298**	**0.715**	**0.985**	0.638
$\Delta W_s^{0,1}\&\Delta W_t^{2,5}$	0.288	0.645	0.977	<u>0.710</u>
$\Delta W_s^{2,6}\&\Delta W_t^{2,5}$	<u>0.295</u>	<u>0.696</u>	<u>0.979</u>	**0.719**

Table 4: Ablation results for the sampling steps of the reference latent in the test-time training stage.

Method	CLIP-T	CLIP-I	Temporal consistency	Motion fidelity
w/o TTT	0.295	0.696	0.979	0.719
Ref. Step=5	**0.308**	**0.699**	**0.980**	0.739
Ref. Step=10	0.302	0.680	0.978	**0.744**

to directly combining the two LoRA components, adding TTT results in more accurate scene generation in the videos, which increases the CLIP-T score. However, when the number of sampling steps reaches 10, the CLIP-T score begins to decline significantly. Thus, we choose the reference step equal to 5 as the default choice.

3 Prompt Influence Analysis

We discuss the prompt influence in the main paper by an example. Here, we give the numerical support. In detail, we employ 100 text prompts describing appearance and another 100 words describing motion. We represent a pair of text prompts as p and p^*, where p is not equal to p^*. We then inject p^* into one of the 9 layers of the model, represented by layer i, and calculate the CLIP-T score (the metrics between appearance and the text) between the generated video and p^*. The quantitative results are presented in Table 2. We find that layers $i = 2$ and $i = 6$ are the most critical layers that focus on appearance. As for motion, we find that the CLIP score has a related weak relationship between the motion and the motion text, resulting in very similar scores. However, layers $i = 2$ and $i = 4$ are the most important layers for motion. As discussed in the main paper, there is no motion module in i = 4, thus, we train the LoRAs on $W_s^{2,6}$ and $W_t^{2,5}$ for appearance and motion customization, respectively. Additional visualization results are shown in Fig. 1 and Fig. 2 on the next page.On the right side of Fig. 1, although injecting p^* into both $i = 6$ and $i = 4$ produces a horse, the lower body remains unnatural, displaying the compact limb features characteristic of a rabbit. A realistic horse only emerges when p^* is injected simultaneously into $i = 6$ and $i = 2$.

References

Chen, X.; Liang, C.; Huang, D.; Real, E.; Wang, K.; Pham, H.; Dong, X.; Luong, T.; Hsieh, C.-J.; Lu, Y.; et al. 2024. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Karaev, N.; Rocco, I.; Graham, B.; Neverova, N.; Vedaldi, A.; and Rupprecht, C. 2023. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Yatim, D.; Fridman, R.; Bar-Tal, O.; Kasten, Y.; and Dekel, T. 2024. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8466–8476.

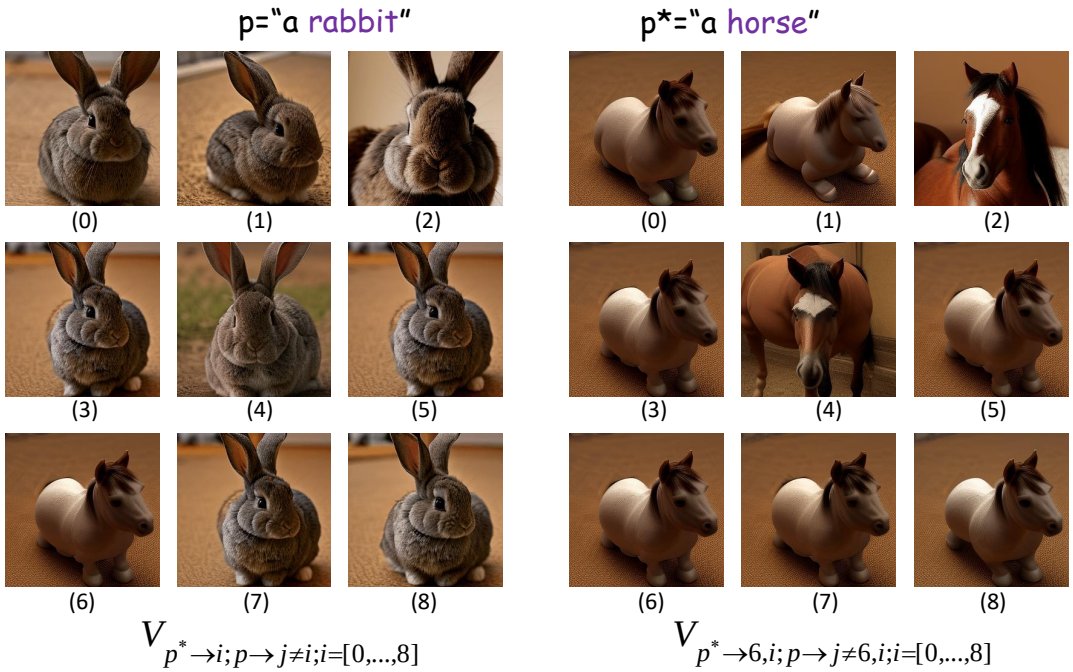

Figure 1: ***Appearance Important Analysis.*** On the left side, p^* is sequentially injected into layers from $i = 0$ to $i = 8$, while the remaining layers are injected with p. To further improve the appearance similarity, on the right side, besides p^* in index 6, we also inject p^* in an additional layer. Thus, $i = 2, 6$ shows the best similarity with the prompt p^*.

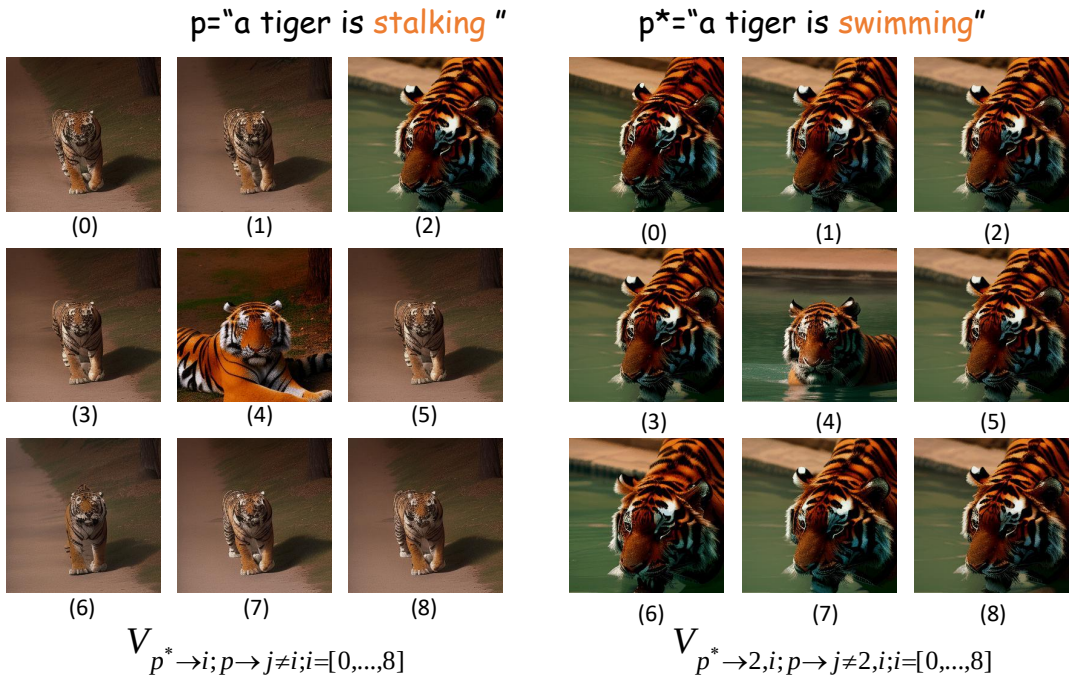

Figure 2: ***Motion Important Analysis.*** On the left side, p^* is sequentially injected into layers from $i = 0$ to $i = 8$, while the remaining layers are injected with p. On the right side, besides injecting p^* in index 2, we add p^* to one of the existing layers. Thus, $i = 2, 4$ are the most necessary layers for motion. We train the LoRA weights on $W_t^{2,5}$ for motion customization.