VideoMage: Multi-Subject and Motion Customization of Text-to-Video Diffusion Models

Supplementary Material

6. Limitation and Future Work

While our method effectively customizes multiple subjects and their motions in videos, it currently lacks the capability to customize long motions and generate corresponding extended videos (e.g., minute-long videos). This limitation is common across all existing methods, as customizing longer videos requires significant computational resources, either during training or inference.

To address this, future work will explore integrating long video generation techniques or training-free customization methods to enable longer customized video generation. By leveraging advancements in efficient video synthesis capable of handling long video sequences, we aim to improve the generation of longer and more intricate customized video content.

7. Additional Experimental Setup

7.1. Additional Implementation Details

Appearance-Agnostic Motion Learning. As described in Sec. 3.2, we employ Textual Inversion [11] to obtain the special tokens representing subject appearances from the reference motion video for our proposed *appearance-agnostic motion learning*. Specifically, we extract a single frame from the reference video and use Grounded-SAM [31] to obtain segmentation masks for each subject. We then crop each subject based on its corresponding mask and learn a special token (i.e., embedding) for each subject using Eq. (4). This approach ensures that the learned tokens accurately reflect the visual identities of the subjects without incorporating any motion information, which is crucial for the appearance-agnostic motion learning phase.

Spatial-Temporal Collaborative Composition. As mentioned in Sec. 3.3, we sample and preprocess two single-subject training videos by combining them into a CutMix-style [15, 50] video for regularizing the LoRA fusion. Specifically, for each video, we use Grounded-SAM2 [30, 31] to generate segmentation masks for the subjects. We then crop

Algorithm 1 Spatial-Temporal Collaborative Sampling (SCS)

Model: Pre-trained video diffusion model θ, fused multi-subject LoRA $\Delta\hat{\theta}_s$, motion LoRA $\Delta\theta_m$
Input: Target text prompt c_{tgt} (w/ subjects' special tokens) and $\tilde{c}_{tgt}$ (w/o special tokens), initial noise map x_T
Output: Sampled video x_0

for $t = T, T-1, ..., 1$ **do**
  Duplicate x_t to create x_t^{sub} and x_t^{mot};
  $\epsilon_t^{sub} = \epsilon_{\hat{\theta}_s}(x_t^{sub}, c_{tgt}, t)$; {Subject branch noise}
  $\epsilon_t^{mot} = \epsilon_{\theta_m}(x_t^{mot}, \tilde{c}_{tgt}, t)$; {Motion branch noise}
  if $T - t < \tau$ **then**
    / Collaborative Guidance */*
    $\mathcal{L}_{s\to m} = \|\mathcal{M}_{\text{SCA},s} - \mathcal{M}_{\text{SCA},m}\|_2^2$;
    $\mathcal{L}_{m\to s} = \|\mathcal{M}_{\text{TSA},s} - \mathcal{M}_{\text{TSA},m}\|_2^2$;
    $x_t^{sub} := x_t^{sub} - \alpha_t \nabla_{x_t^{sub}} \mathcal{L}_{m\to s}$;
    $x_t^{mot} := x_t^{mot} - \alpha_t \nabla_{x_t^{mot}} \mathcal{L}_{s\to m}$;
    Execute lines 3 and 4 to get updated ϵ_t^{sub} and ϵ_t^{mot};
  end if
  $\epsilon_t = \beta_s \epsilon_t^{sub} + \beta_m \epsilon_t^{mot}$ and obtain x_{t-1};
end for
Return x_0;

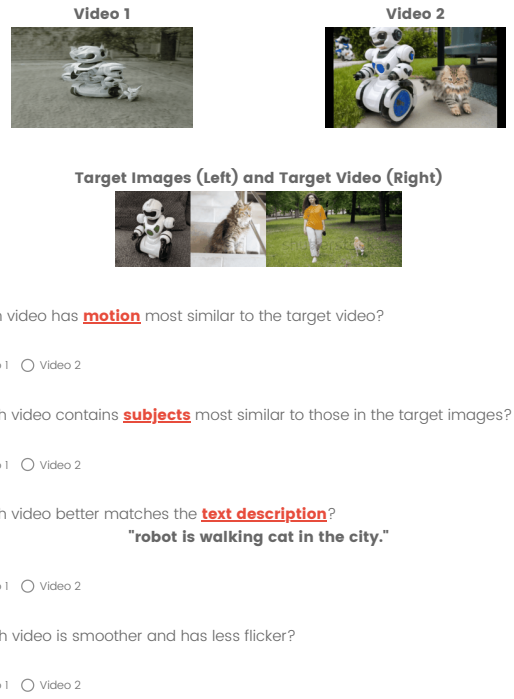

Figure 8. **User study interface.** Given two generated videos, reference subject images, and a reference motion video, participants compare the generated videos based on *Motion Fidelity*, *Subject Fidelity*, *Text Alignment*, and *Video Quality*.

λ_1	CLIP-T	CLIP-I	DINO-I	T. Cons.
0.1	0.658	0.667	0.398	0.981
0.25	**0.662**	**0.670**	**0.407**	**0.983**
1.0	0.660	0.667	0.401	0.980

(a) Weight for video preservation loss λ_1

λ_2	CLIP-T	CLIP-I	DINO-I	T. Cons.
0.1	0.641	0.659	0.362	0.980
0.6	**0.662**	**0.670**	**0.407**	0.983
1.0	0.656	0.665	0.402	**0.984**

(b) Weight for attention regularization loss λ_2

Template for c_{ap}	**CLIP-I**	**DINO-I**
"A static video of <sub1> and <sub2>."	**0.670**	**0.407**
"A video of <sub1> and <sub2> being still."	0.664	0.405
"A video of <sub1> and <sub2>."	0.659	0.395

(c) Template for appearance prompt c_{ap}

ω	CLIP-T	CLIP-I	DINO-I	T. Cons.
0.1	0.646	0.658	0.372	0.981
0.5	**0.662**	**0.670**	**0.407**	**0.983**
1.0	0.657	0.667	0.403	0.980

(d) Scale factor of negative guidance ω

α_t	CLIP-T	CLIP-I	DINO-I	T. Cons.
10^3	0.657	0.665	0.401	**0.988**
10^4	**0.662**	**0.670**	**0.407**	0.983
10^5	0.634	0.658	0.379	0.976

(e) Scale factor of collaborative guidance α_t

τ	CLIP-T	CLIP-I	DINO-I	T. Cons.
5	0.658	0.664	0.401	0.978
15	**0.662**	**0.670**	**0.407**	**0.983**
30	0.657	0.664	0.399	0.975

(f) Steps of collaborative guidance τ

Table 3. Ablation studies on various hyperparameters, including the weights for video preservation loss (λ_1) and attention regularization loss (λ_2), the template for appearance prompt (c_{ap}), the negative guidance scale factor (ω), the collaborative guidance scale (α_t) and steps (τ).

the subjects from the original frames and place them onto a clean background video. To encourage potential interactions between the subjects, we allow some degree of overlap in their placements. We initialize the fused LoRA $\hat{\theta}_s$ with the average of the subject LoRAs. The training steps range from 250 to 450, depending on the subject combination.

For *spatial-temporal collaborative sampling* (SCS), we provide the details in Algorithm 1. Following prior works [44, 51], we initialize the noise map as $x_T = \sqrt{\beta}\epsilon_m + \sqrt{1-\beta}\epsilon$, where $\beta = 0.3$, ϵ_m is the DDIM [37] inverted noise of the motion video, and ϵ is Gaussian noise sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This initialization is consistently applied to all comparison methods in all experiments.

7.2. Human User Study

In Fig. 8, we present the interface for our human preference study. In this study, participants are provided with reference subject images, a reference motion video, and two customized videos: one from our *VideoMage* method and one from a comparison method (i.e., DreamVideo [44] or MotionDirector [51]). They are asked to choose their preferred video based on four questions, each evaluating: *Motion Fidelity*, *Subject Fidelity*, *Text Alignment*, and *Video Quality*. A total of 360 videos were generated for each method, and 25 participants participated in the study.

8. Additional Results

8.1. Ablation Studies on Hyperparameter Choices

Effect of λ_1 in subject learning. As illustrated in Tab. 3(a), a video preservation loss weight of $\lambda_1 = 0.25$ achieves the best performance, while both smaller (0.1) and larger (1.0) values lead to declines. Thus, we set $\lambda_1 = 0.25$ for all experiments.

Method	**CLIP-T**	**CLIP-I**	**DINO-I**	**T. Cons.**
DisenStudio [8]	0.661	0.658	0.381	0.842
CustomVideo [43]	**0.676**	0.679	0.402	0.849
***VideoMage* (ours)**	0.674	**0.681**	**0.403**	**0.851**

Table 4. **Quantitative comparison on multi-subject customization.** Following [8, 43], we evaluate using CLIP-Text Alignment (*CLIP-T*), CLIP-Image Alignment (*CLIP-I*), DINO-Image Alignment (*DINO-I*), and Temporal Consistency (T. Cons.).

Effect of λ_2 in multi-subject fusion. As shown in Tab. 3(b), the optimal performance is achieved with the attention regularization loss weight set to $\lambda_2 = 0.6$, whereas smaller (0.1) or larger (1.0) values lead to reduced performance. Thus, we use $\lambda_2 = 0.6$ in our experiments.

Effect of c_{ap} and ω in appearance-agnostic motion learning. As shown in Tab. 3(c), we experiment with three different templates for the subject appearance prompt c_{ap} used in our appearance-agnostic motion learning. The template, "A static video of <sub1> and <sub2>," achieves the best performance and is therefore used in our experiments. Notably, all three templates outperform the second-best result achieved by MotionDirector [51], as presented in Tab. 1. Similarly, in Tab. 3(d), we present the ablation study on the scale factor of negative guidance ω. We observe that setting ω to 0.5 yields the best results; thus, $\omega = 0.5$ is adopted for all experiments.

Effect of τ and α_t in spatial-temporal collaborative sampling. In Tab. 3(e) and Tab. 3(f), we ablate the scale factor α_t and steps τ in our proposed spatial-temporal collaborative sampling, respectively. As shown in Tab. 3(e), increasing α_t from 10^3 to 10^4 improves performance, but further increasing it to 10^5 results in a decline. Consequently, we set $\alpha_t = 10^4$ for our experiments. Similarly, in Tab. 3(f), increasing τ to 15 improves performance, while any further increase leads to a drop. Therefore, we set $\tau = 15$.

8.2. Multi-Subject Customization

To validate the effectiveness of our proposed *test-time multi-subject fusion*, we compare *VideoMage* with state-of-the-art methods on the multi-subject customization task. Using the subject sets and prompts described in Sec. 4.1, we generate 720 videos and evaluate performance using *CLIP-T*, *CLIP-I*, *DINO-I*, and *T. Cons.*, following [8, 43]. For fair comparison, we omit the additional bounding boxes required by DisenStudio. As shown in Tab. 4, *VideoMage* outperforms the second-best method in *CLIP-I*, *DINO-I*, and *T. Cons.*, and is comparable to CustomVideo in *CLIP-T*.

8.3. More Qualitative Results.

In Fig. 9, we present additional qualitative results of *VideoMage* customizing videos with multiple subjects and motion, successfully demonstrating diverse subject-motion combinations across various scenes, including cases with more than two subjects.

Figure 9. **Additional qualitative results.** Each row presents the subject images, the motion video, the corresponding customized video results, and the input prompt.