SMRABooth: Subject and Motion Representation Alignment for Customized Video Generation

Supplementary Material

1. Overview

The supplementary material includes the following sections:

1. Additional implementation details.
2. A detailed discussion of our methods.
3. Additional qualitative results generated by our DiT-based method.
4. Extended details for our U-Net-based method.
5. More qualitative results generated by our U-Net-based method.
6. A local anonymous project page and a demo video.
7. A folder containing the videos generated by our method.

2. More implementation details

2.1. Further explanation of the ablation experiments for our DiT-Based method

In Sec. 4.4, we propose a subject-motion association decoupling injection strategy that sparsifies LoRAs in injection timing. Specifically, we adjust subject LoRA weights at a critical timestep T_{point} to balance appearance fidelity and video coherence. Lower subject LoRA weights are applied before T_{point} to prioritize motion generation, while higher weights are applied afterward to enhance subject identity preservation and temporal consistency. Here, we conduct two ablation studies to evaluate the choice of T_{point} and the subject LoRA scale before T_{point}.

Effect of Sparse LoRA Injection Timing. To determine T_{point}, we conduct an ablation study alongside the theoretical analysis in Sec. 4.4, as shown in Fig. 1. Setting T_{point} too early causes subject LoRA to interfere with motion LoRA, disrupting motion generation (e.g., Fig. 1, Denoise step=5), as the model focuses on motion in the early denoising stages. Conversely, setting T_{point} too late results in subjects inconsistent with the reference (e.g., Fig. 1, Denoise step=25 and Denoise step=45). At this stage, the model emphasizes fine-grained details, but a low subject LoRA scale allows the text prior to dominate, causing deviations from the reference. Based on our analysis, we set T_{point} to 15 (see Fig. 1, Denoise step=15), striking a balance between maintaining subject appearance and preserving temporal dynamics for coherent motion and accurate subject representation.

Comparison of Different Scales of subject LoRA Injection. In our subject-motion association decoupling injection strategy, we adjust subject LoRA weights at T_{point} to balance appearance fidelity and video coherence. Lower subject LoRA weights are applied before T_{point}, while higher weights are applied afterward to achieve a smooth trade-off between subject identity preservation and temporal consistency. To analyze the effect of the subject LoRA scale, we conduct an ablation study (see Fig. 2). Setting the subject LoRA scale too small (e.g., Fig. 2, Scale=0.1) results in poor preservation of the subject's shape and structure, leading to inaccurate subject generation and loss of criti-

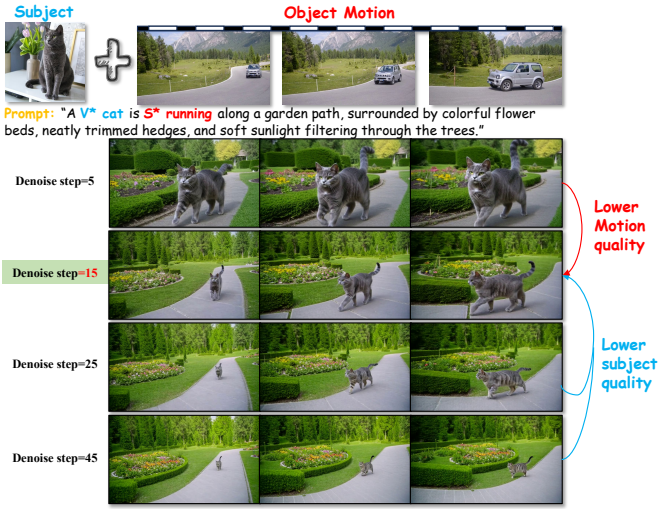

Figure 1. **Ablation study on injection timing** T_{point}: We analyze four choices: 5, 15, 25, 45 and find that applying lower subject LoRA before denoise step 15, as described in Sec. 3.4, achieves the best performance.

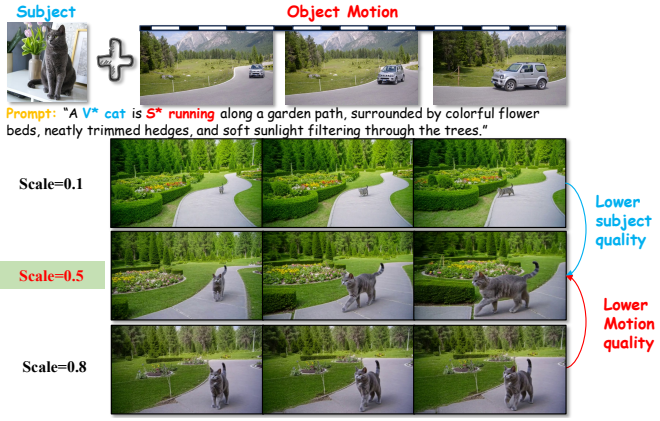

Figure 2. **Ablation study on subject LoRA scale before** T_{point}: By exploring different scales, we find that a subject LoRA scale of 0.5 achieves the best performance before T_{point}.

cal information. On the other hand, setting the scale too large (e.g., Fig. 2, Scale=0.8) interferes with the motion LoRA, making the video appear static and losing dynamic motion consistency. Based on our analysis, we set the subject LoRA scale to 0.5 (see Fig. 2, Scale=0.5), achieving the best balance between subject appearance and temporal dynamics for coherent motion and accurate subject representation.

2.2. Details for Temporal Representation Alignment

In Eq.9, we present an equation to reverse z_t to z_0 with one-step denoising. A more comprehensive explanation of the derivation of Eq.9 is provided in Alg. 1.

Algorithm 1 Prediction Function for z_0 in Flow Matching

Prediction function for z_0 in Flow Matching

In Flow Matching: $z_t = (1-t)\cdot z_0 + t \cdot z_1$

$\Rightarrow z_0 = \frac{z_t - t\cdot z_1}{1-t}$ ...*(1)*

where z_1 is noise, model predicts the velocity: $v_\theta \approx z_1 - z_0$

$\Rightarrow z_1 \approx v_\theta + z_0$ ...*(2)*

Substitute (2) into (1):

$z_0 = \frac{z_t - t\cdot(v_\theta + z_0)}{1-t}$

$z_0 = \frac{z_t - t\cdot v_\theta - t\cdot z_0}{1-t}$

$z_0 \cdot (1-t) = z_t - t\cdot v_\theta - t\cdot z_0$

$z_0 = z_t - t\cdot v_\theta$ ...*(3)*

where in WAN, $v_\theta = u(z_t, c_{txt}, t; \theta)$ ...*(4)*

Substitute (4) into (3):

Therefore: $z_0 = z_t - t\cdot u(z_t, c_{txt}, t; \theta)$

Additionally, Eq.8 requires reversing the latents and using a 3D VAE to decode them on CUDA, which results in significant computational overhead. To address this issue, we adopt a sliding window technique to efficiently select a subset of latents for decoding.

The sliding window is set to a size of 6 in latent space and moves by 2 frames at a time. This means that at each step, the 3D VAE processes $1 + 4 \times (6-1) = 21$ frames simultaneously. Considering the temporal dependency characteristics of the 3D VAE, we discard the first 5 frames and retain the subsequent 16 frames. During experiments, we confirm that the preserved 16 frames effectively retain motion patterns similar to those of the source video frames at corresponding positions, making them suitable for our motion representation.

2.3. Dataset Construction Details

The pipeline for constructing our training dataset is illustrated in Fig. 3. For the subject images, we use LangSAM to generate binarized masks. First, we guide the segmentation model LangSAM by injecting accurate descriptions of the target object to ensure precise segmentation of the subject area. LangSAM can process images of any resolution. After segmentation, we annotate each subject image with the caption "A picture of V* <subject name>" and combine it with the original image and the generated mask to create the training data. For the motion videos, we first sample one frame from the source video and use GPT-4o to generate a caption that provides a detailed description of the visual content and its motion characteristics. After obtaining the caption, we manually refine it by removing redundant information, such as audio-related descriptions. Finally, the corrected caption is paired with the source video to construct the motion dataset. During inference, we use GPT-4o to combine subjects and motions in pairs and generate diverse and creative background descriptions for them. We exclude illogical combinations, such as "a car playing the piano" or "a car playing basketball," to ensure realistic scenarios.

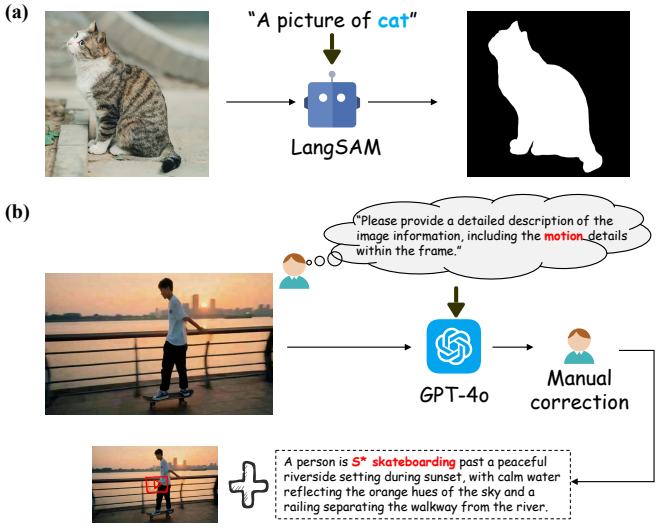

Figure 3. Illustration of our dataset construction. (a) shows the processing details for subject images. We guide the LangSAM model with text prompts to segment the subject area, resulting in a binarized mask. (b) shows the processing details for motion videos. We sample a frame from the videos, use GPT-4o to generate a detailed caption for the frame, manually refine the prompt, add special tokens, and pair the caption with the source video.

2.4. Baseline Details and Training Cost

For DualReal, we adopt CogVideoX-5B as the text-to-video backbone. We run 1,000 training steps for each test case, and each output contains 49 frames at a resolution of 720×480 pixels. For WAN2.1+LoRAs fine-tuning, we follow the official WAN2.1 1.3B training code. LoRA uses a learning rate of 1×10^{-4} with 300 subject LoRA steps and 400 motion LoRA steps, the same as SMRABooth. Each output contains 49 frames at a resolution of 480×832 pixels. For WAN2.1 1.3B, we test the backbone's native reasoning capabilities. Each output contains 49 frames at a

resolution of 480×832 pixels.

For the training cost, our SMRABooth requires approximately 30 minutes to train a single subject LoRA and about one hour to train a single motion LoRA. In contrast, DualReal requires joint training of two hours for the combined LoRAs of subject and motion.

2.5. Detailed introduction of metrics

We establish a comprehensive evaluation framework across three dimensions: Semantic Alignment, Motion Quality and Perceptual Quality, using nine metrics.

- **Semantic Alignment.** (1) **CLIP-T:** This metric evaluates the alignment between text prompts and generated videos by calculating the frame-wise cosine similarity between their embeddings, derived from the CLIP [7] model. (2) **CLIP-I:** This metric assesses the visual-semantic correspondence by comparing the embeddings of reference images and generated video frames. These embeddings are obtained using the image encoder from CLIP [7]. (3) **DINO-I:** Similar to CLIP-I, this metric measures the visual-semantic correspondence; however, it utilizes the DINO [6] vision transformer encoder to compute and compare embeddings of reference images and generated frames. For these three metrics, we evaluate them on our generated video frame by frame and compute their final scores by averaging the results across all frames.
- **Motion Quality.** (1) **Motion Fidelity:** Evaluates the consistency of motion patterns by leveraging CoTracker3 [3], a model designed for diffusion-motion-transfer [11]. (2) **Subject Consistency:** Assesses whether the appearance of the subject (e.g., characters) remains consistent across different frames in the video, as implemented in VBench [2]. (3) **Temporal Consistency:** Quantifies frame-to-frame consistency by calculating the average cosine similarity of CLIP [7] image embeddings across all frame pairs in the generated video.
- **Perceptual Quality.** (1) **PickScore:** Predicts human preference scores using PickScore [5], with results averaged at the frame level. (2) **Aesthetic Quality:** Measures artistic merit using the LAION aesthetic predictor, implemented via VBench [2]. (3) **Imaging Quality:** Evaluates distortions in generated frames, such as overexposure, noise, and blurriness, as assessed via VBench [2].

2.6. User study interface

During the user study, we provide each case video generated by WAN2.1, WAN2.1+LoRAs, DualReal, and our SMRABooth (WAN) for evaluation based on four questions. Each question is rated on a scale from 1 to 5 for the following criteria: (1) The accuracy of generating the video to match the text descriptions (*Prompt Alignment*). (2) Consistency between the generated video and the provided motion mode (*Motion Similarity*). (3) The similarity between the main body of the generated video and the reference image provided (*Appearance Similarity*). (4) The overall quality of the video (*Video Quality*). Fig. 4 shows the format of our questionnaire format.

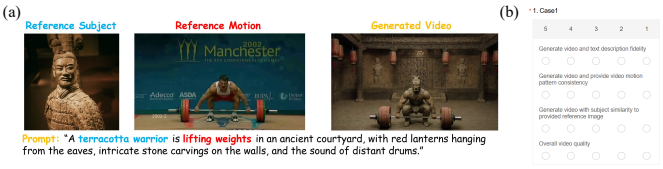

Figure 4. **Human evaluation questionnaire format.** (a) presents the reference subject image, reference video, and the customized video generated by the model. Participants are then asked to complete the evaluation form in (b), rating the quality of the generated video based on Prompt Alignment, Motion Similarity, Appearance Similarity and Video Quality.

3. Discussion

Methods Discussion. While existing methods like VideoJAM [1] leverage optical flow to improve motion quality for T2V models, our Motion Representation Alignment differs significantly: (1) **Task Perspective:** VideoJAM focuses on enhancing the backbone model's motion quality, aiming for generalizable motion across diverse prompts and scenarios. In contrast, SMRABooth specializes in learning fixed motion patterns from reference videos. Additionally, SMRABooth leverages optical flow primarily for motion feature extraction decoupled from appearance, further reducing the coupling with appearance features through sparse LoRAs. (2) **Method Perspective:** VideoJAM relies on extensive training with video and optical flow, whereas SMRABooth adopts a lightweight fine-tuning framework. VideoJAM utilizes in-context learning, which requires a large amount of training data, while SMRABooth's temporal representation alignment enables motion pattern learning from a single video.

Limitations. While SMRABooth excels in customized subject and motion generation, it has some limitations. One key limitation is its inability to handle multi-object customized generation, a related and important area left unexplored in this work. Expanding SMRABooth in this direction could greatly enhance its versatility for more complex scenarios. Another challenge is the lack of open-source datasets for subject and video customization, limiting our ability to conduct more extensive experiments to further validate the model's performance. Developing or accessing such datasets is crucial for advancing research in this field.

4. Additional qualitative results generated by our DiT-based method.

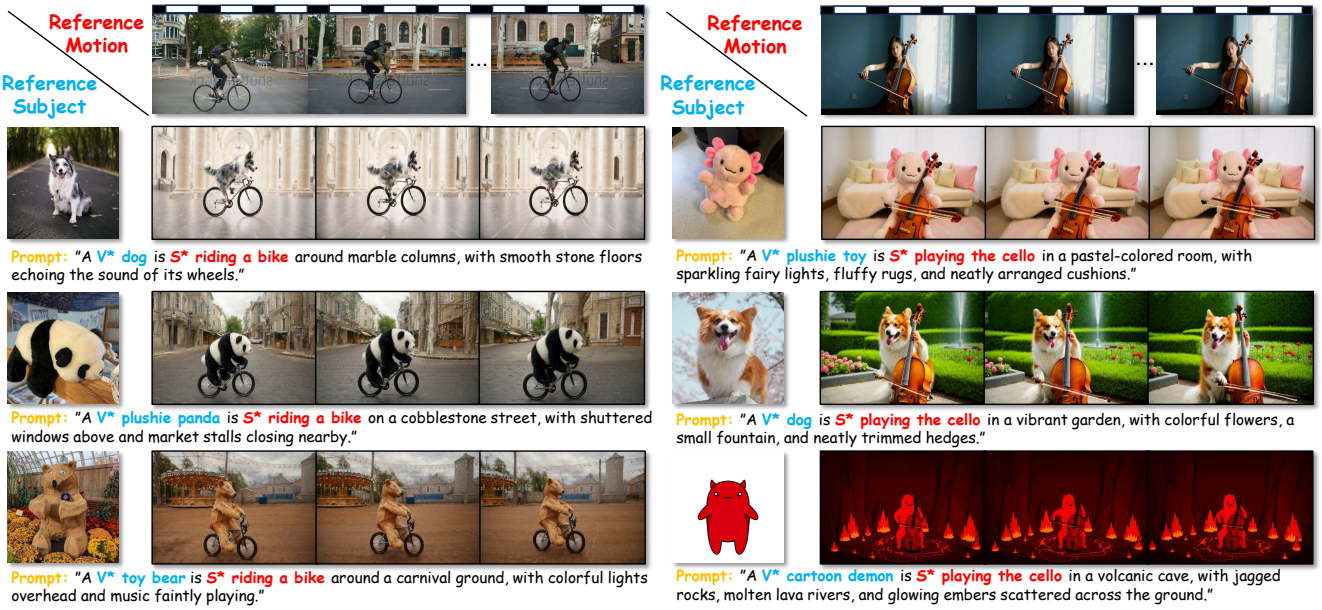

Figure 5. **More qualitative results of our joint customization for subject and motion.** In this case, we use a set of videos to guide our model in learning the motion concept. SMRABooth generates customized videos that accurately preserve subject appearance and motion patterns while remaining faithful to text prompts.

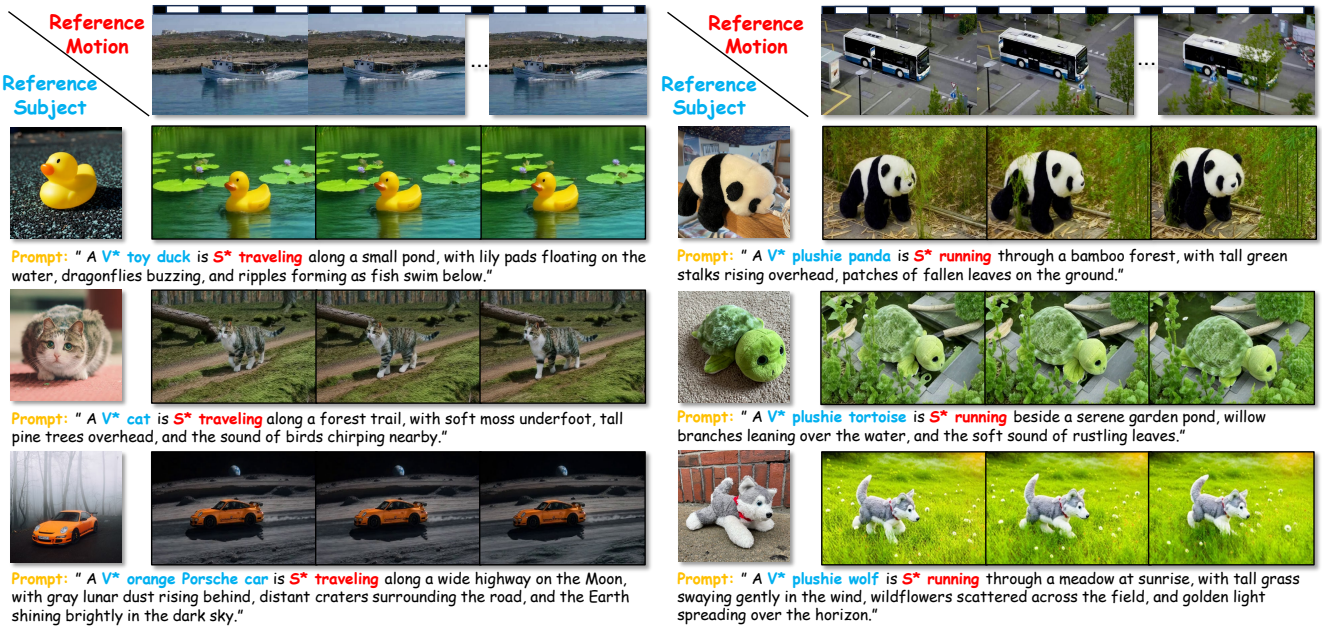

Figure 6. **More qualitative results of our joint customization for subject and motion.** In this case, we use a set of videos to guide our model in learning the motion concept. SMRABooth generates customized videos that accurately preserve subject appearance and motion patterns while remaining faithful to text prompts.

Figure 7. **More qualitative results of our joint customization for subject and motion.** In this case, we use a set of videos to guide our model in learning the motion concept. SMRABooth generates customized videos that accurately preserve subject appearance and motion patterns while remaining faithful to text prompts.

Figure 8. **More qualitative results of our joint customization for subject and motion.** In this case, we use a set of videos to guide our model in learning the motion concept. SMRABooth generates customized videos that accurately preserve subject appearance and motion patterns while remaining faithful to text prompts.

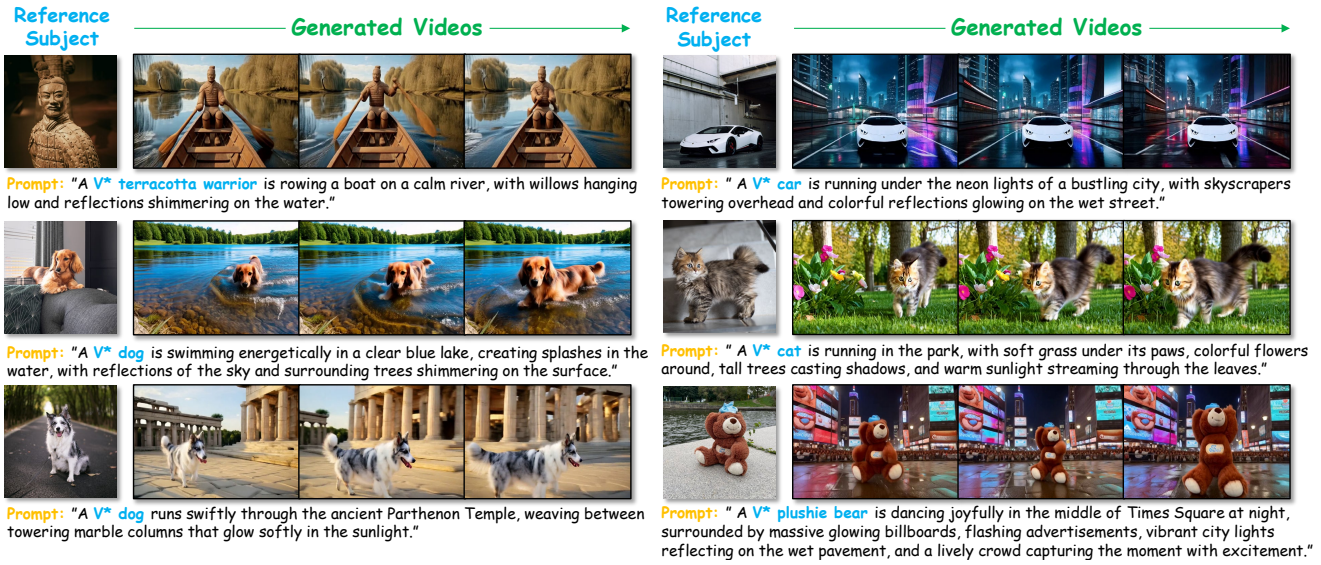

Figure 9. **More qualitative results of our customization for subject.** In this case, We have customized and generated a variety of different subjects, including various world historical sites and natural landscapes. Our cases fully demonstrate the accurate extraction of theme features and the strong generalization capabilities of our model.

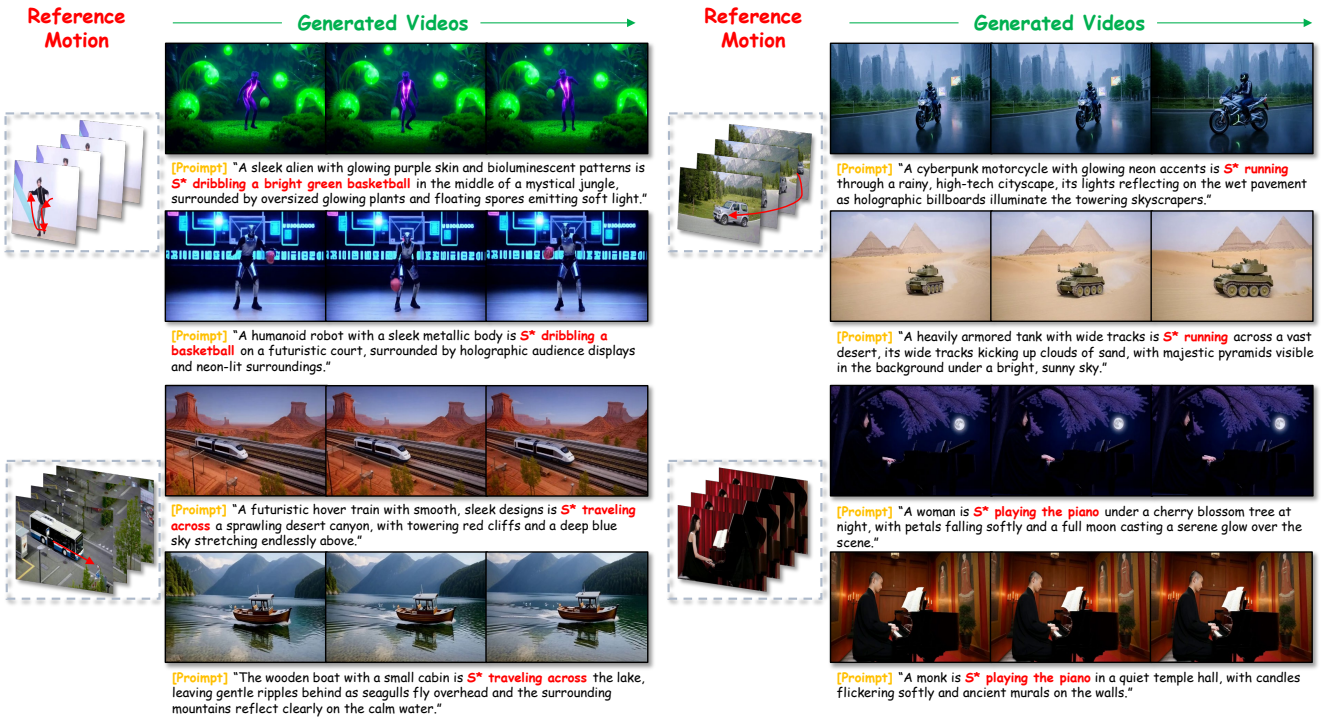

Figure 10. **More qualitative results of our customization for motion.** In this case, We have customized and generated a variety of different motion. Our cases fully demonstrate the accurate extraction of theme features and the strong generalization capabilities of our model.

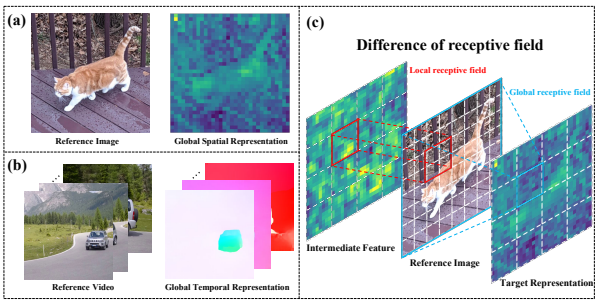

Figure 11. (a) Global spatial representations capture the global spatial structure and semantic information from the reference image. (b) Global temporal representations capture object-level motion trajectories and motion trends from the reference video. (c) Visualization of the receptive field of a patch in the final feature map. The target patch captures global features, while intermediate features focus only on local features.

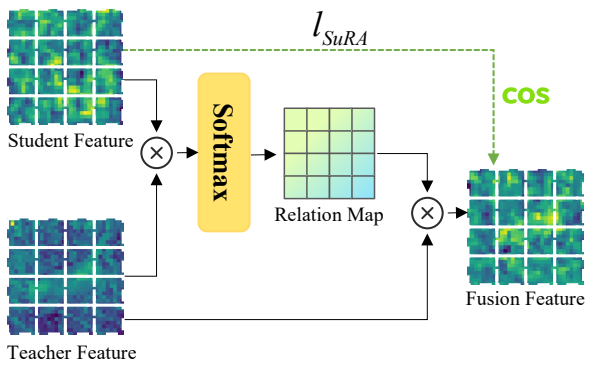

Figure 12. The design of the RAA module: It features a structure similar to cross-attention but effectively bridges the gap between receptive fields of varying sizes.

5. Details for our U-Net-based method

5.1. Technical details

For U-Net-based methods, the main challenge lies in the mismatch of receptive fields between U-Net and ViT. This inconsistency in receptive fields between encoder representations and U-Net features can cause semantic information confusion during direct alignment. Thus, we introduce the **R**epresentation-**A**ware **A**lignment (**RAA**) block (Fig. 11(c) and Fig. 12). This block integrates each patch of local features with the entirety of global representations, significantly augmenting the ability of local features to aware global spatial structure and high-level semantic information. Using these settings to train a spatial low-rank matrix (LoRA) within spatial transformers, we can accurately preserve the subject's appearance from reference images. As shown in Fig. 11(c), the ViT-based target network, with its larger receptive fields and more feature channels, captures richer semantic context per pixel compared to the U-Net-based network, which is limited by smaller receptive fields. Pixel-wise alignment of representations can lead to suboptimal generation quality due to mismatches in receptive field size and semantic richness. While using a homogeneous encoder seems like an intuitive solution, experiments reveal it is less effective than a heterogeneous encoder, as detailed in the supplementary materials. To address this issue, we propose a **R**epresentation-**A**ware **A**lignment (RAA) block, which first fuses the feature distributions of the two architectures instead of directly computing the loss between them. This approach helps bridge the representational gap between the ViT-based vision encoder and the U-Net-based. We first compute the relation map between the intermediate and target features and fuse the two features patch by patch. The relation map is defined as:

$$R = \text{softmax}\left(\frac{h_\phi(z_t)\cdot \mathbf{y}^{*\top}}{\sqrt{d}}\right), \tag{1}$$

The fusion feature $\mathbf{x}^*$ is then calculated as:

$$\mathbf{x}^* = R\cdot \mathbf{y}^*, \tag{2}$$

Finally, feature alignment is realized through the loss function, defined as:

$$\mathcal{L}_{\text{SuRA}}(\theta) = -\mathbb{E}_{\mathbf{x}_*,\boldsymbol{\epsilon},t}\left[\frac{1}{N}\sum_{n=1}^{N} sim(\mathbf{y}^{*[n]}, \mathbf{x}^{*[n]})\right], \tag{3}$$

where n is the patch index, and $sim(\cdot,\cdot)$ is a pre-defined similarity function.

Moreover, to prevent subject LoRA from overfitting to the background of the subject image during training, we introduce masks M generated by SAM [4], which enforce the model to focus only on the subject region. Formally, the masked MSE loss is defined as:

$$\mathcal{L}_{\text{region}} = \mathbb{E}_{\mathbf{z},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t,\mathbf{c}}\left[||(\epsilon - \epsilon_\theta(\mathbf{z}_t,\mathbf{c}_i,t))\cdot\mathbf{M}||_2^2\right], \tag{4}$$

where M represents the mask applied to the subject region. During training, the overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{region}} + \lambda\mathcal{L}_{\text{SuRA}}, \tag{5}$$

where $\lambda > 0$ is a hyperparameter that balances the subject representation alignment loss.

5.2. Experiment evaluation

Quantitative Evaluation. As shown in Table 1, SMRABooth outperforms SOTA methods in text-video alignment, visual similarity to reference images, and temporal consistency. Specifically: (1) Compared to DreamVideo [9], SMRABooth improves CLIP-T from 0.298 to 0.329, CLIP-I

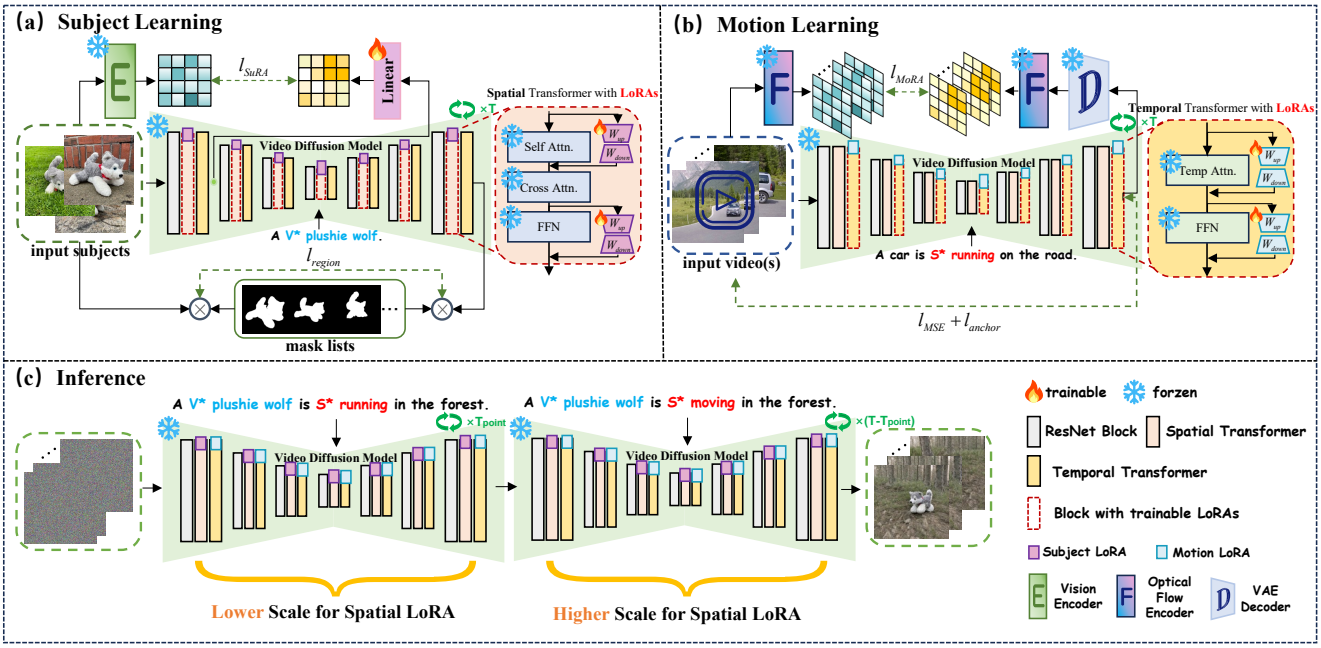

Figure 13. **Overview of SMRABooth for U-Net-based SMRABooth.** The framework divides customized video generation into two stages: subject learning and motion learning.

Table 1. Quantitative experimental results for different methods under the numerical evaluation metrics.

Method	Objective evaluation				User study			
	CLIP-T↑	CLIP-I↑	DINO-I↑	T. Cons.↑	Prompt Alignment	Motion Similarity	Appearance Similarity	Video Quality
DreamVideo	0.298	0.609	0.302	0.970	2.751	2.867	2.676	2.767
MotionBooth	0.301	0.689	0.449	0.954	2.890	2.819	2.915	2.870
MotionDirector	0.295	0.759	0.597	**0.989**	3.023	2.915	3.165	3.078
SMRABooth(Ours)	**0.329**	**0.760**	**0.612**	0.986	**3.488**	**3.501**	**3.543**	**3.499**

Table 2. Quantitative ablation studies on each component in a subset of our training set. We select 15 subject-motion pairs.

Method	CLIP-T↑	CLIP-I↑	DINO-I↑	T. Cons.↑
w/o l_{SuRA}	0.343	0.742	0.523	0.982
w/o RAA	0.338	0.744	0.561	0.985
w/o l_{MoRA}	0.323	0.693	0.489	0.987
Ours	**0.345**	**0.754**	**0.594**	**0.988**

from 0.609 to 0.760, DINO-I from 0.302 to 0.612, and Temporal Consistency from 0.970 to 0.986. (2) Compared to MotionBooth [10], SMRABooth raises CLIP-T from 0.301 to 0.329, CLIP-I from 0.689 to 0.760, and Temporal Consistency from 0.954 to 0.986. (3) Compared to MotionDirector [8], SMRABooth improves CLIP-T from 0.295 to 0.329 while maintaining CLIP-I at 0.760. MotionDirector's higher Temporal Consistency is due to still-image-like outputs lacking realistic motion. In summary, SMRABooth excels in text alignment, temporal coherence, and fidelity to reference images and videos.

Effect of l_{SuRA}**.** As shown in Table 2, l_{SuRA} provides high-level spatial information, helping the model preserve global structure and semantic consistency. This ensures fidelity to reference identities while avoiding over-reliance on low-level details, resulting in more coherent outputs.

Effect of RAA**.** As shown in Table 2, directly using pretrained visual encoder features as alignment targets degrades the model's features, leading to suboptimal results. The RAA block resolves this by fusing heterogeneous features before alignment, allowing the model to better utilize high-level semantic information and improve subject generation quality.

Effect of l_{MoRA}**.** As shown in Table 2, l_{MoRA} provides object-level motion information, enabling the model to capture global motion trends and maintain coherent motion trajectories. This ensures consistent and realistic motion while disentangling appearance from motion dynamics.

6. Additional qualitative results generated by our U-Net-based method.

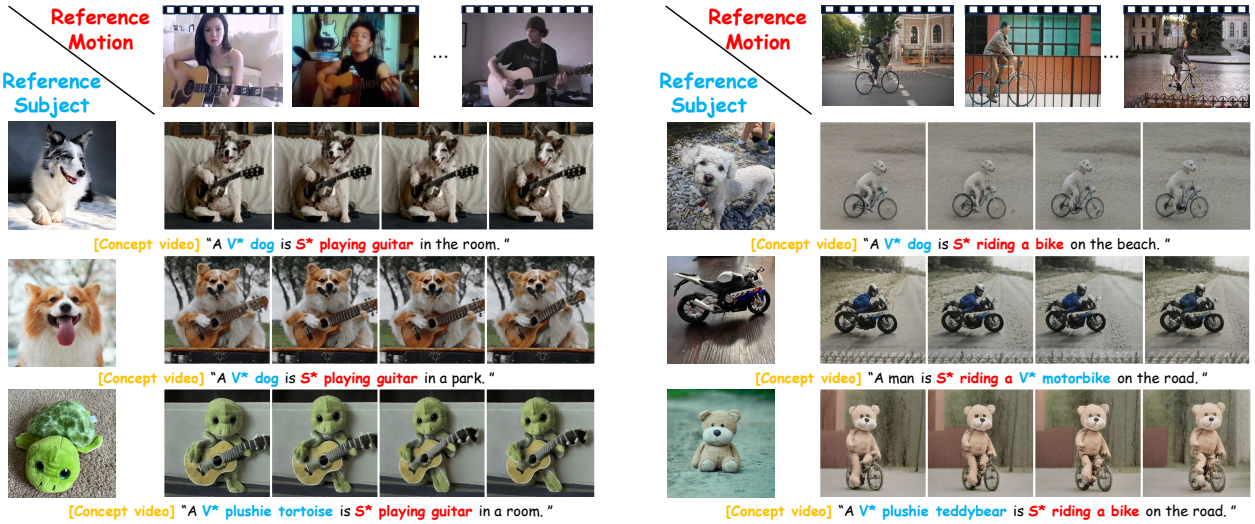

Figure 14. **More qualitative results of our joint customization for subject and motion.** In this case, we use a set of videos to guide our model in learning the motion concept. SMRABooth generates customized videos that accurately preserve subject appearance and motion patterns while remaining faithful to text prompts.

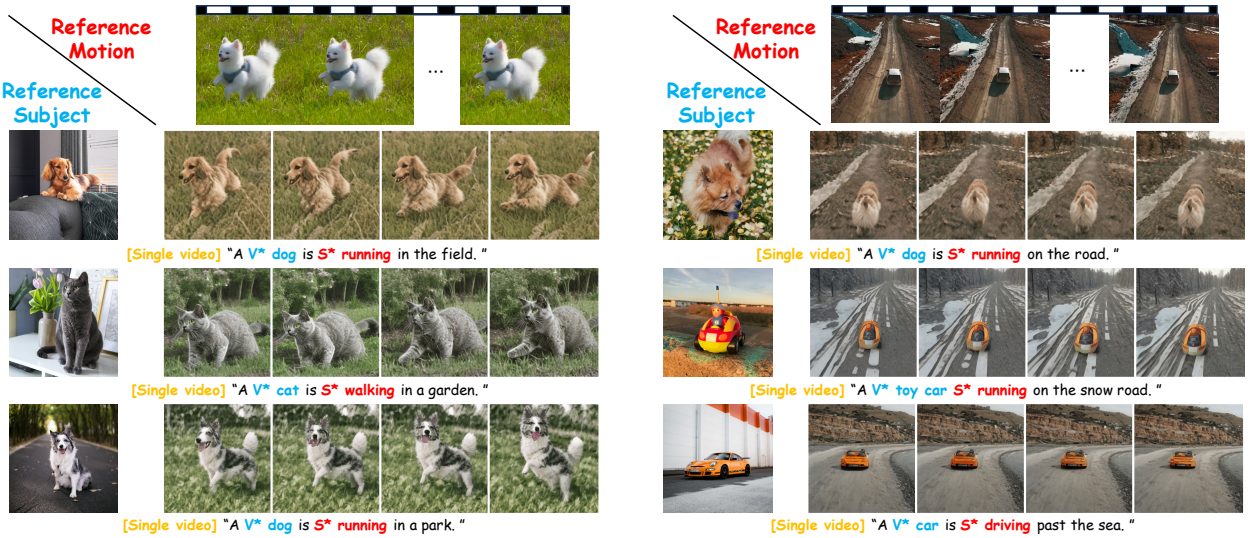

Figure 15. **More qualitative results of our joint customization for subject and motion.** In this case, we use a single video to guide our model in learning the specific object motion. SMRABooth generates customized videos that accurately preserve subject appearance and motion patterns while remaining faithful to text prompts.

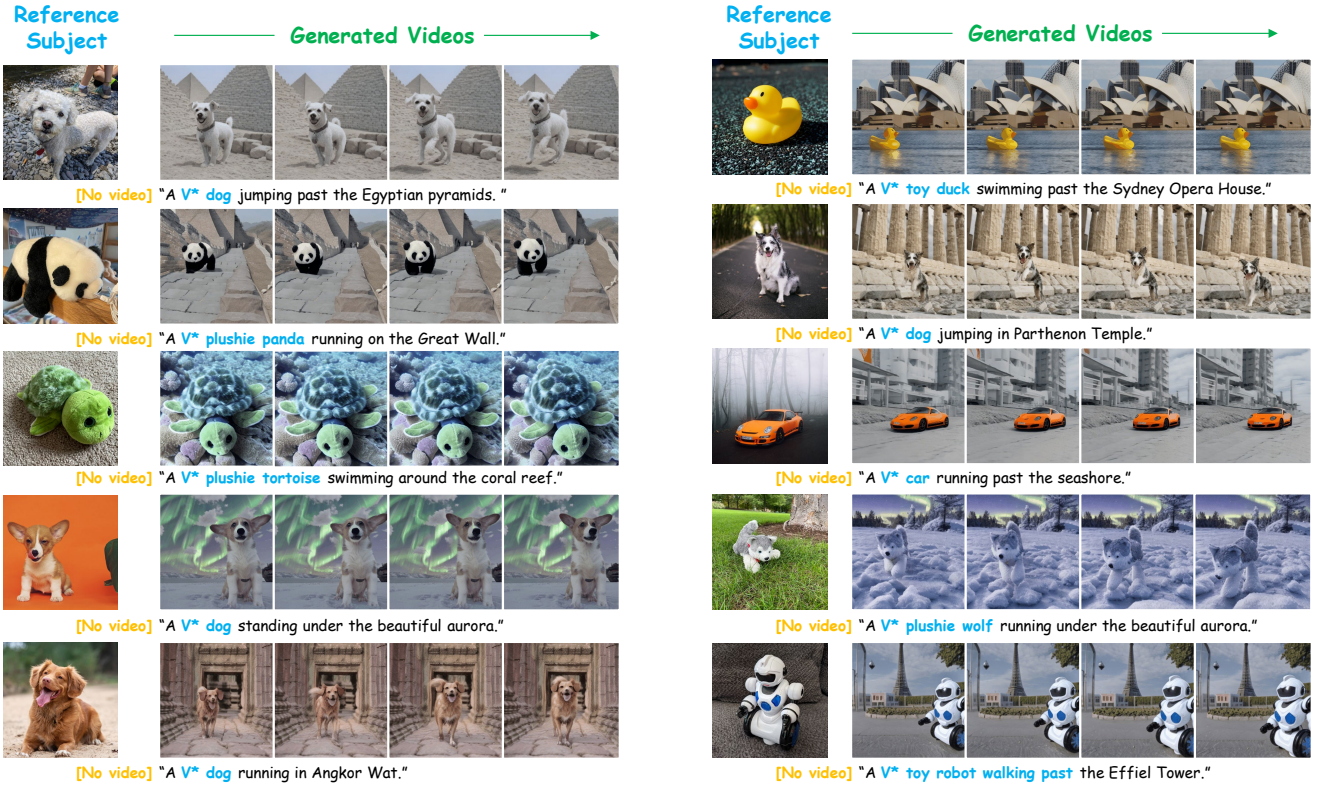

Figure 16. **More qualitative results of our customization for subject.** In this case, We have customized and generated a variety of different subjects, including various world historical sites and natural landscapes. Our cases fully demonstrate the accurate extraction of theme features and the strong generalization capabilities of our model.

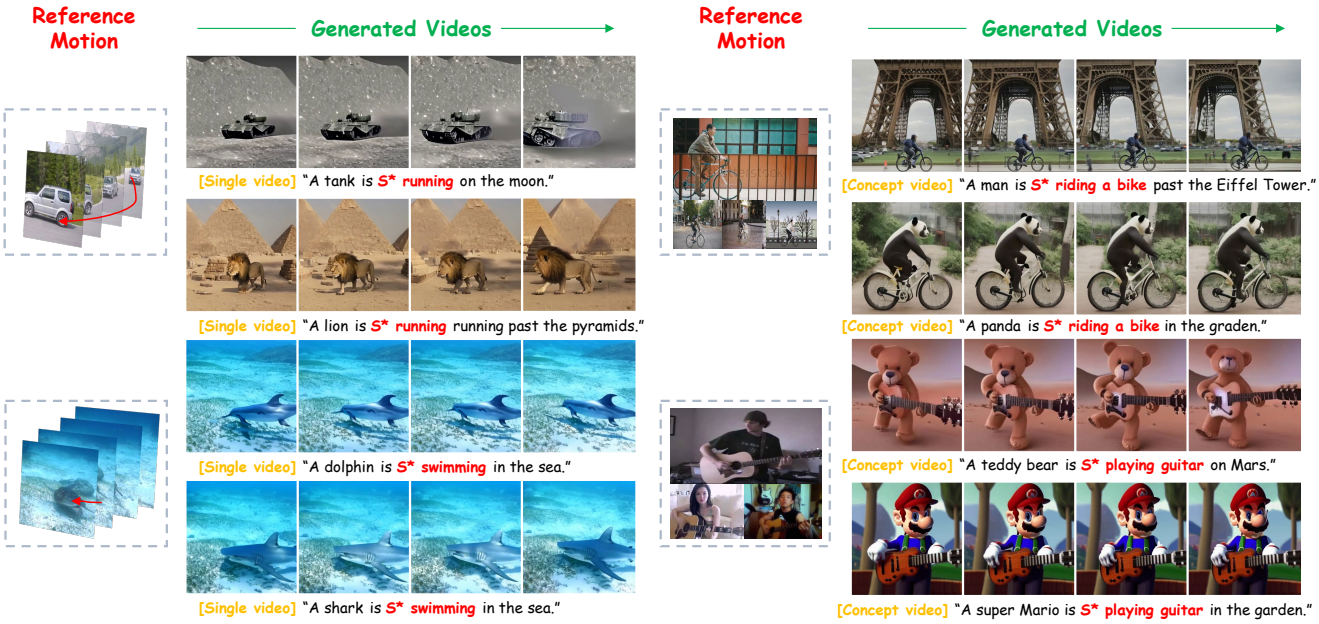

Figure 17. **More qualitative results of our customization for motion.** In this case, We have customized and generated a variety of different motion. Our cases fully demonstrate the accurate extraction of theme features and the strong generalization capabilities of our model.

References

[1] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. In *Forty-second International Conference on Machine Learning*, 2025. 3

[2] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 3

[3] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pages 18–35. Springer, 2024. 3

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 7

[5] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 3

[6] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3

[7] 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 *ICML*, pages 8748–8763. PMLR, 2021. 3

[8] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 8

[9] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 7

[10] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8

[11] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 3