SynMotion: Semantic-Visual Adaptation for Motion Customized Video Generation

Supplementary Material

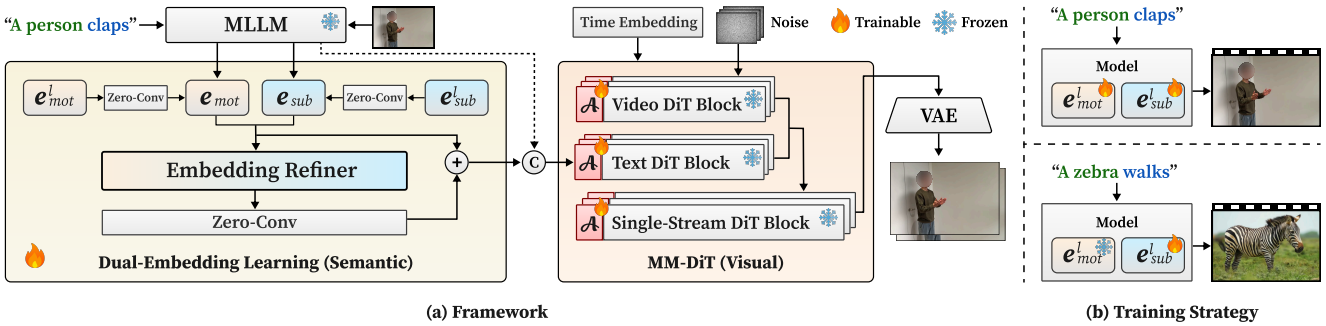


Figure 1. Pipeline of our method generalized on Image-to-Video task. The pipeline of `SynMotion` in the I2V setting. Given an image input and a prompt in the form of $<$ subject, motion $>$, we first extract the image embedding using a Multi-modal Large Language Model (MLLM). Meanwhile, the prompt is processed as in the T2V setting to obtain text embeddings, which are decomposed into a subject embedding e_{sub} and a motion embedding e_{mot}. Each component is augmented via a zero-initialized convolutional residual (Zero-Conv) $\mathcal{Z}$ with learnable embeddings e^l_{sub} and e^l_{mot}. The image embedding and the refined prompt embeddings are then concatenated along the token length dimension. This fused representation is passed through the Embedding Refiner $\mathcal{R}$ to integrate subject and motion semantics. The refined embeddings are reinserted via Zero-Conv $\mathcal{Z}$ and injected into the video generation backbone. As in the T2V pipeline, a dedicated motion Adapter $\mathcal{A}$ enhances motion-specific features, enabling the model to generate temporally coherent and motion-aligned videos guided by both visual and textual inputs.

1. I2V Motion Customization

1.1. Pipeline.

As shown in Fig. 1, given both an image input and a prompt of the form $<$ subject, motion $>$, we begin by extracting two types of embeddings: (1) the image embedding, obtained using a Multimodal Large Language Model (MLLM) applied to the visual input, and (2) the text embedding, derived from the input prompt using the same MLLM. We then perform prompt-aware decomposition on the text embedding to obtain two components: the subject embedding e_{sub} and the motion embedding e_{mot}, based on token roles within the prompt. As in the T2V setting, we attach learnable residual embeddings (i.e., e^l_{mot} and e^l_{sub}) via a Zero-Conv layer $\mathcal{Z}$ to endow these components with adaptability. The initialization of e^l_{mot} and e^l_{sub} follows the same strategy: e^l_{mot} is initialized using the embedding of the phrase '*a person claps*' to preserve motion-specific semantics, while e^l_{sub} is randomly initialized to allow subject flexibility. Finally, the image embedding and the dual-embedding learning outputs are concatenated along the token dimension and jointly injected into the downstream modules for generation.

The Embedding Refiner $\mathcal{R}$ remains consistent with the design in the T2V setting, serving to facilitate interaction between subject and motion semantics in the latent space. This ensures better alignment between visual appearance and motion intent, and the refined embedding is reintegrated through the same residual injection process.

1.2. Comparison.

Comparison Setting. Since no existing method is specifically designed for motion customization in the I2V setting, we compare against three state-of-the-art I2V base models: CogVideoX-I2V [14], HunyuanVideo-I2V [5], and Wan2.1-I2V [12]. We assess each method using a combination of automated and perceptual metrics covering motion fidelity, subject identity, temporal coherence, and visual quality:

- **Motion Accuracy.** To evaluate whether the generated video accurately reflects the intended motion, we follow the protocol used in the T2V setting. Each video is input into QwenVL [2, 3, 10], a state-of-the-art Vision-Language Model (VLM), and prompted with a yes/no question asking whether the specified motion is depicted. The binary answers are converted into accuracy scores. This process is repeated 10 times for each motion category, and the average is reported.
- **Motion Smoothness.** Measures the temporal continuity and smooth transitions of the subject's motion across frames, reflecting kinematic realism.
- **Temporal Flickering.** Quantifies frame-wise incon-

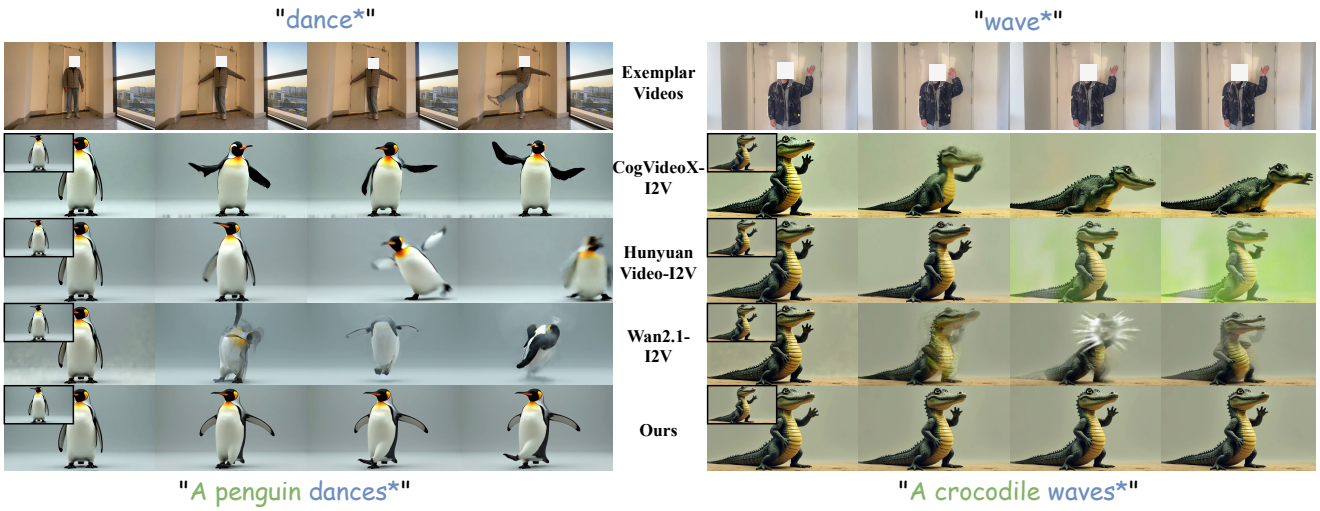


Figure 2. The results of our method generalized on Image-to-Video task. For each example, the input image is shown in the top-left corner.

Method	Motion Accuracy	Motion Smoothness	Temporal Flickering	I2V Subject	Subject Consistency	Background Consistency	Aesthetic Quality
CogvideoX-I2v [14]	44.29%	98.92%	98.33%	97.65%	93.22%	95.60%	60.59%
HunyuanVideo-I2V [15]	27.81%	99.13%	98.53%	97.32%	93.38%	95.72%	60.54%
Wan2.1-I2V [17]	53.25%	98.44%	98.01%	94.02%	90.36%	95.28%	59.01%
SynMotion	**78.30%**	**99.68%**	**99.69%**	**99.83%**	**98.28%**	**98.40%**	**65.52%**

Table 1. Quantitative comparison results in image-to-video setting. The best results for each column are **bold**.

sistencies in appearance, which indicate instability and degrade the overall coherence of the video.

- **I2V Subject Accuracy.** Evaluates whether the identity and visual appearance of the generated subject match the input reference image.
- **Subject Consistency.** Assesses whether the generated subject remains visually coherent throughout the video, without noticeable changes in style, shape, or texture.
- **Background Consistency.** Measures the stability of the background across frames to ensure that motion generation does not introduce environmental flicker.
- **Aesthetic Quality.** Captures the overall visual appeal of the video, including color fidelity, sharpness, and artistic coherence.

Qualitative Comparison. As shown in Fig. 2, we compare the visual outputs of different I2V baselines on motion customization tasks. CogVideoX-I2V is able to preserve the subject identity throughout the video but fails to generate the correct motion, often resulting in static or incorrect behaviors. HunyuanVideo-I2V and Wan2.1-I2V, while occasionally producing partial motion patterns, suffer from subject instability, where the subject either gradually fades, deforms, or drifts in the later frames. This inconsistency severely impacts the realism and coherence of the generated video. In contrast, our method maintains both the structural integrity of the subject and the temporal consistency of the motion. The generated videos demonstrate clear execution of the intended motions while preserving the subject's appearance across all frames. This highlights the strength of our dual-level design in achieving fine-grained motion customization without sacrificing subject fidelity.

Furthermore, we conduct a comparison with FlexiACT [16] under the same settings. Fig. 8 shows the results on our MotionBench, where FlexiACT fails to learn the correct motion in the 'wave' case. For instance, instead of the panda waving its hand, its entire body appears to sink, and severe artifacts emerge on the arm of cat. In contrast, our method not only learns the motion from the reference video but also correctly transfers it to the target image. Fig.9 displays the results on the FlexiACT benchmark. Although FlexiACT learns the exercise motion, it also overfits to the content information from the reference video, such as incorporating human legs. In the sea lion example, it even transforms the sea lion into a human figure. By comparison, our method effectively preserves the identity from the target image while successfully animating it with the correct motion.

Quantitative Comparison. Tab. 1 presents the quantitative comparison between our method and several state-of-the-art I2V baselines. While CogVideoX-I2V maintains

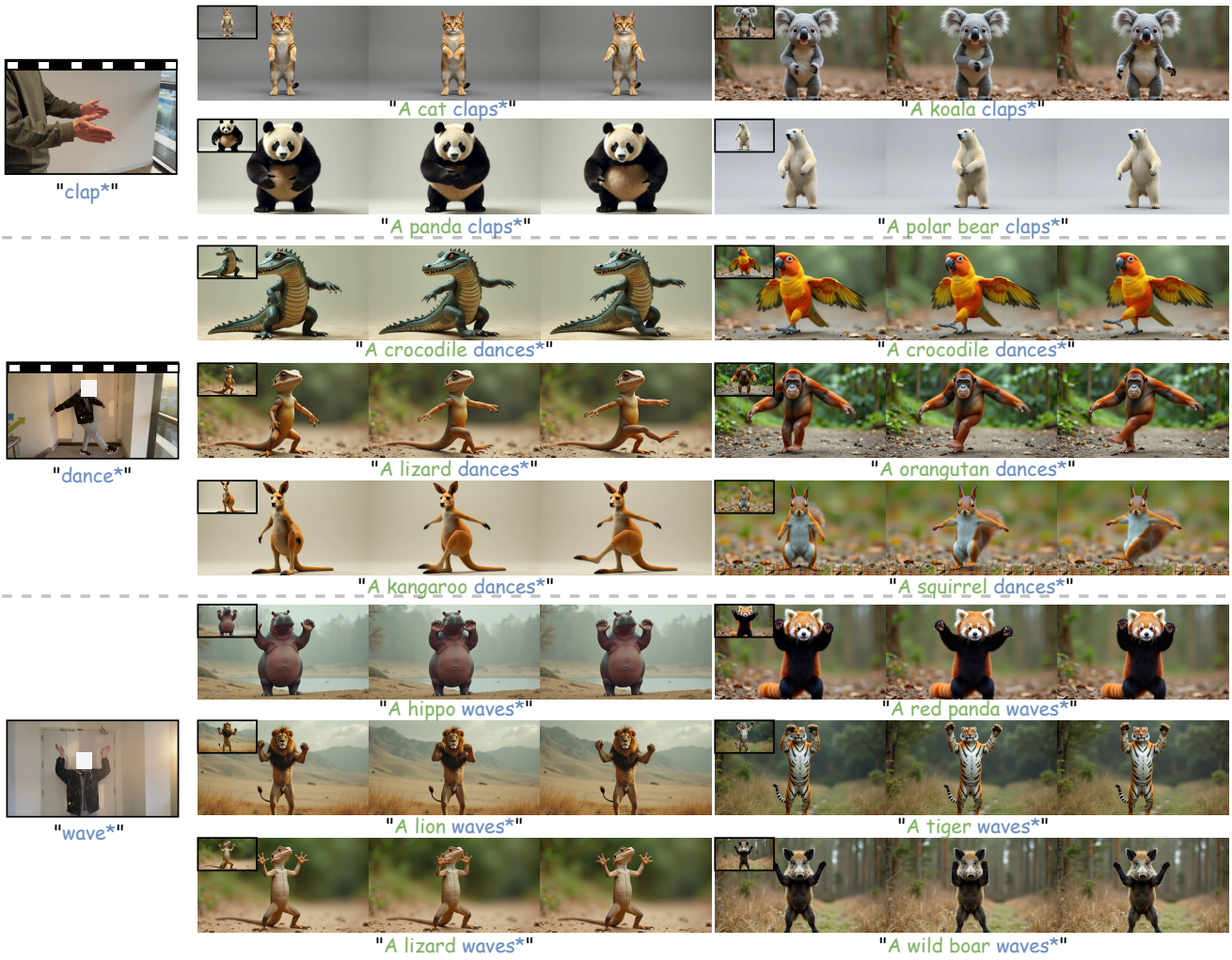


Figure 3. More results generated by our method.

Configuration	Motion Accuracy (↑)	Subject Accuracy (↑)	CLIP-T (↑)	FVD-3DRN50 (↓)	FVD-3DInception (↓)	Flow Score (↓)	Temporal Consistency (↑)
FlexiACT-M	69%	88%	0.23 ± 0.04	395.00	4327.56	0.34 ± 0.01	0.99 ± 0.00
Ours-M	**76%**	**97%**	**0.36** ± 0.04	**342.21**	**3407.80**	**0.31** ± **0.00**	**0.99** ± **0.00**
FlexiACT-F	62%	86%	0.228 ± 0.01	436.87	3330.52	5.64 ± 0.14	0.97 ± 0.00
Ours-F	**72%**	**98 %**	**0.31** ± **0.02**	**329.47**	**3105.14**	**3.29** ± **0.02**	**0.99** ± **0.00**

Table 2. Comparison with FlexiACT in MotionBench (M) and FlexiACT (F) dataset.

subject identity well, it fails to produce accurate motions. HunyuanVideo-I2V and Wan2.1-I2V generate partial motions but exhibit issues like temporal flickering and subject degradation. In contrast, our method achieves significantly better motion accuracy and smoothness, while also improving subject consistency and background stability. Furthermore, it delivers superior aesthetic quality, highlighting the overall effectiveness of our dual-level design in motion customization.

As shown in Tab. 2, we present a quantitative comparison of our method against FlexiACT on both the MotionBench and FlexiACT benchmarks. Our method achieves superior results over FlexiACT on multiple metrics. This aligns with the qualitative comparisons in Fig. 8 and Fig. 9, demonstrating the effectiveness of our approach.

More subjects generated by `SynMotion`. As shown in Fig. 10, we present the results of motion customization applied to a broader set of subjects. In addition to the primary animal subjects demonstrated in Figs. 3–5 (*e.g.*,

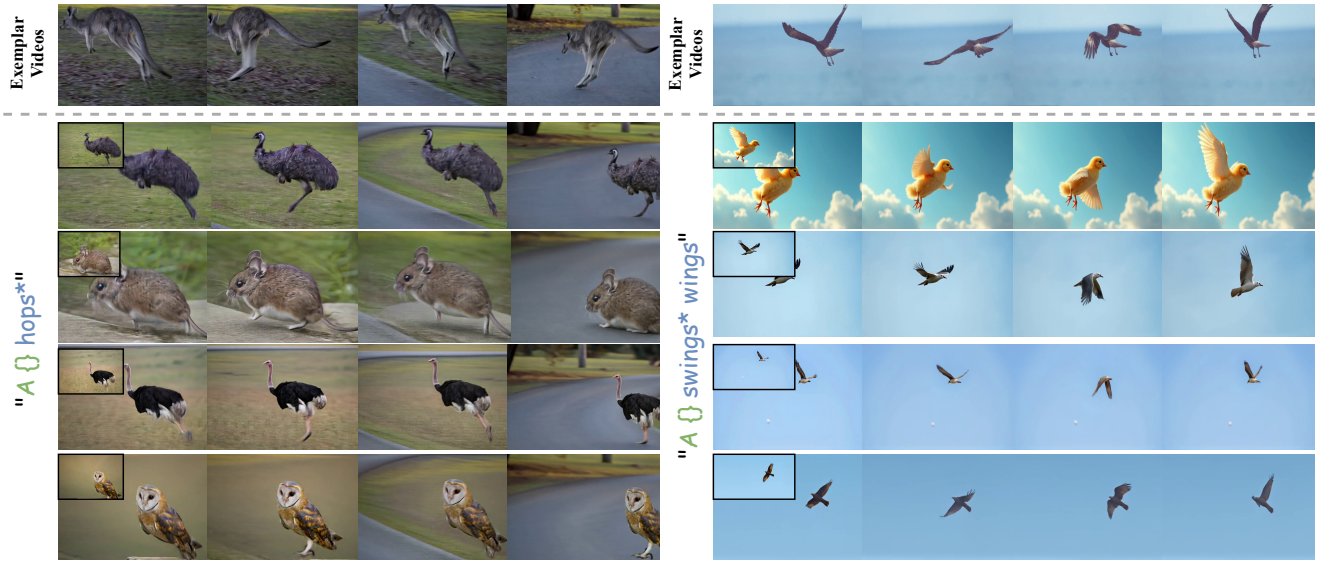


Figure 4. More results generated by our method in animal-to-animal setting.

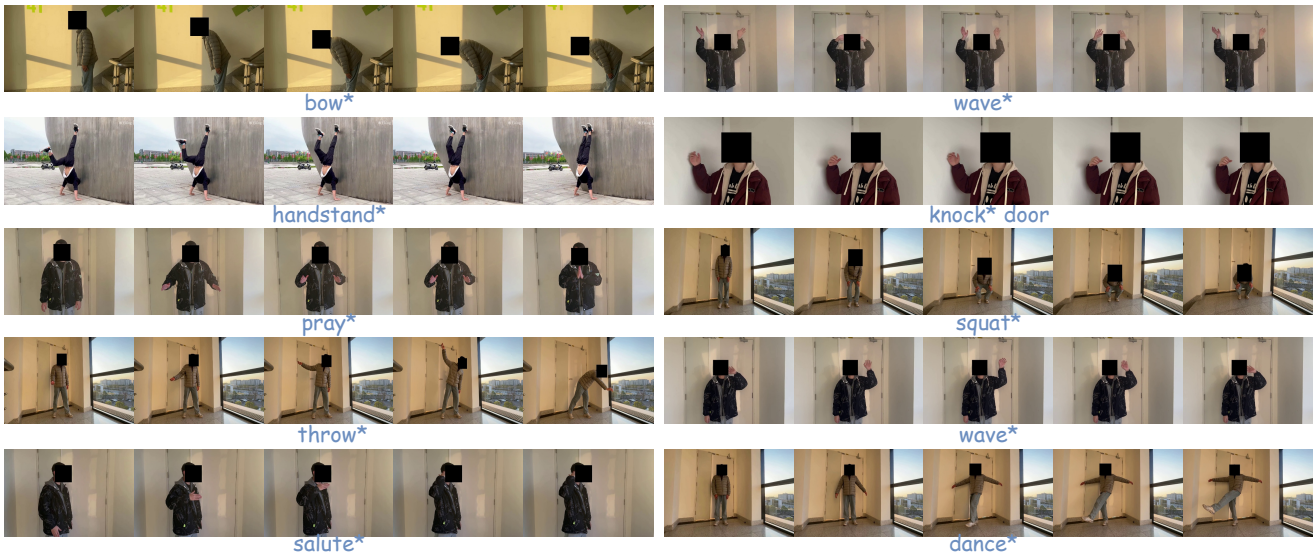


Figure 5. Examples from our MotionBench.

'rabbit', *'sea lion'*), we also include celebrities (*e.g.*, *'Marilyn Monroe'*, *'Barack Obama'*) and fictional characters (*e.g.*, *'Captain America'*, *'Iron Man'*). Our method preserves the identity features of these diverse subjects, and successfully transfers the motion dynamics demonstrated in the exemplar videos. These results highlight the strong generalization ability of our framework across a wide range of subject identities, as well as its high motion fidelity.

1.3. More I2V results.

To further demonstrate the effectiveness and generalization ability of our method in the I2V setting, we present additional qualitative results in Fig. 3. Specifically, we select three representative motions and apply them to six diverse subjects. Despite variations in appearance and anatomy, our method successfully generates motion-consistent and identity-preserving videos for each subject-motion pair. These results highlight the robustness of our framework in adapting customized motions to a wide range of subject

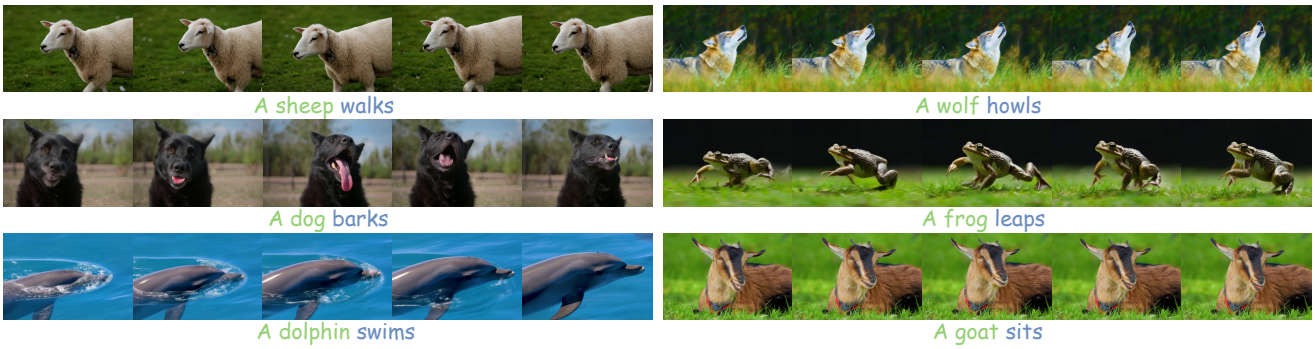


Figure 6. Examples from our SPV dataset.

types, validating its strong generalization capability under the I2V paradigm.

In addition, we extend our exploration to an intriguing application aimed at further assessing the generalization ability of our method: transferring motion patterns from one animal to another. As shown in Fig. 4, we provide two representative cases: one on the left utilizing data from the available dataset published by FlexiAct [16], and one on the right using videos collected by ourselves. The results demonstrate that our method can effectively learn motion representations from animal videos and generalize them across different animal species. This not only validates the generalization strength of our approach but also showcases its potential for broader applications in cross-species motion customization.

2. Other Results

2.1. Ablation Study

In the main paper, we present the qualitative results from our ablation studies and provide detailed analyses. To more intuitively illustrate the contribution of each component, we further include quantitative ablation results, as shown in Tab. 3. The baseline achieves relatively high subject accuracy but suffers from low motion accuracy, indicating that while it generalizes well across different subjects, it lacks sufficient expressiveness in motion representation. After introducing e^l_{mot}, motion accuracy improves significantly; however, subject accuracy drops sharply. This suggests that relying solely on e^l_{mot} enhances motion customization but compromises subject generalization. The addition of e^l_{sub} mitigates this issue, yet motion customization performance declines again, mainly due to the lack of interaction between e^l_{mot} and e^l_{sub}. To address this, our proposed refiner module $\mathcal{R}$ balances motion expressiveness and subject consistency, effectively resolving the conflict. Finally, the motion adapter further enhances motion controllability, achieving the highest motion accuracy.

In addition, we also conducted the following experiments to examine the individual contributions of LoRA and our strategy. (1) Baseline + LoRA only: As shown in Table 4, this setup fails to address the key challenges of motion-customized video generation. Although video fidelity improves slightly, there is no noticeable improvement in motion alignment or subject consistency. These results confirm that LoRA alone is insufficient for our task. (2) SynMotion without the new training strategy: We further ablate our embedding-specific training strategy. As shown in Table 4, subject and motion accuracy drop significantly when removing our embedding-specific training strategy. Although the model tends to overfit to seen subjects, it struggles to adapt motion to unseen entities because standard training objectives cause incorrect updates that entangle subject and motion semantics. These results highlight the essential role of our training strategy in achieving proper disentanglement and generalization.

2.2. More T2V Results

To further demonstrate the generalization ability of our method across diverse motions and subjects, we supplement the main experiments with additional text-to-video (T2V) results. As shown in Fig. 7 and Fig. 11, these examples cover a broader spectrum of motion types, including both simple (*e.g.*, bowing, waving) and complex (*e.g.*, dancing, opening door) actions, as well as a wide variety of subjects. Our method consistently generates motion-consistent and identity-preserving videos across different combinations of prompts. These results not only highlight the flexibility of our dual-level modeling approach but also validate its robustness in handling motion customization under diverse semantic conditions.

2.3. Failure cases

Despite the strong generalization capability of our method across various motions and subjects, we also observe several failure cases. As illustrated in the left part of Fig. 12, when handling rare and complex motions such as "*backflip*" that involve leaping backward off the ground, rotating mid-air with the back facing downward, and landing upright,

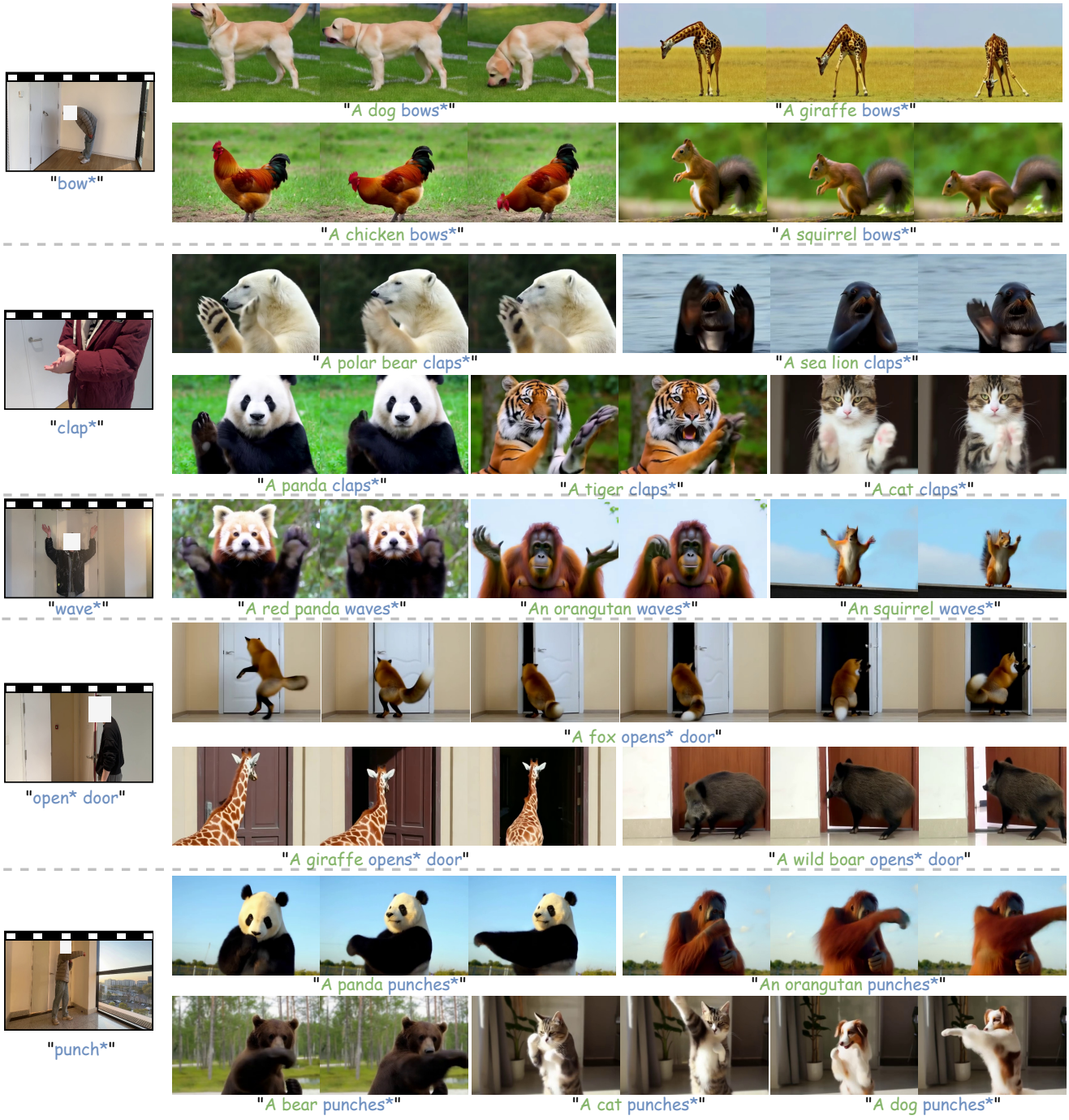


Figure 7. More results generated by our method.

the model struggles to maintain consistent kinematic coherence. This results in noticeable motion artifacts, particularly during the airborne phase. The inherent difficulty lies in the high-level coordination required for such dynamic motion, making it challenging for the model to generate temporally consistent and realistic frames. On the other hand, as shown in the right part of Fig. 12, in the I2V setting, the model may fail to perform the desired customized motion if the input reference image lacks essential body parts needed for the motion (*e.g.*, hands or paws for clapping). In such cases, the subject often remains static, as the model lacks sufficient visual cues to synthesize the corresponding action.

3. Discussion

3.1. Theories and method improvement

We position our work as a significant empirical contribution that advances the field through novel task formulation,

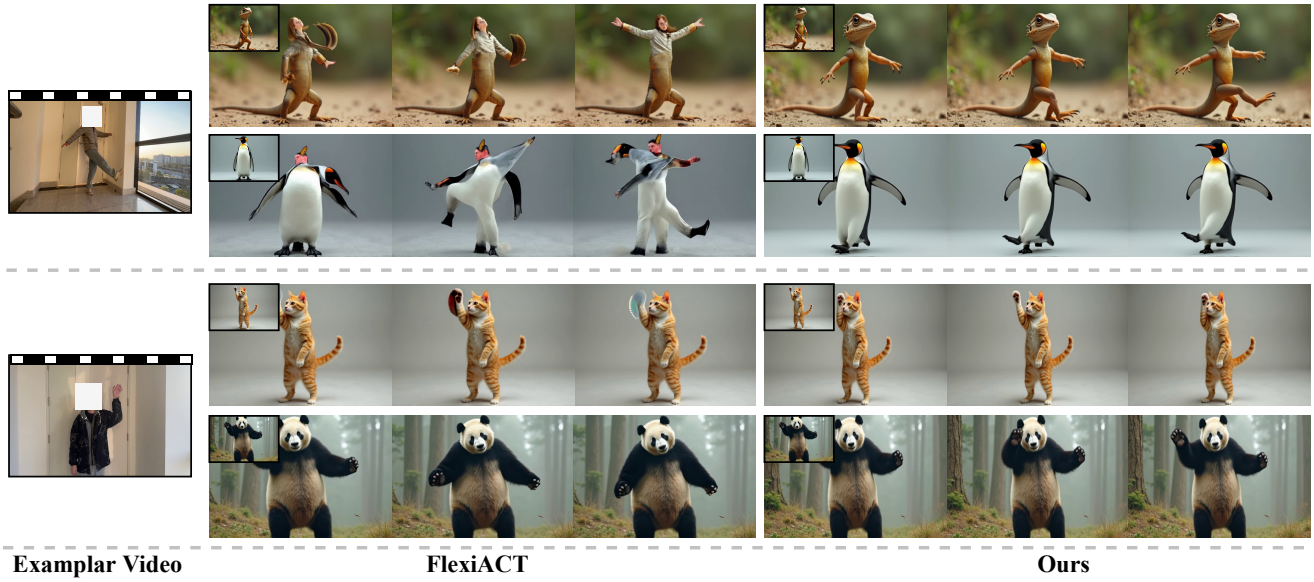


Figure 8. Comparison with FlexiACT on our MotionBench.

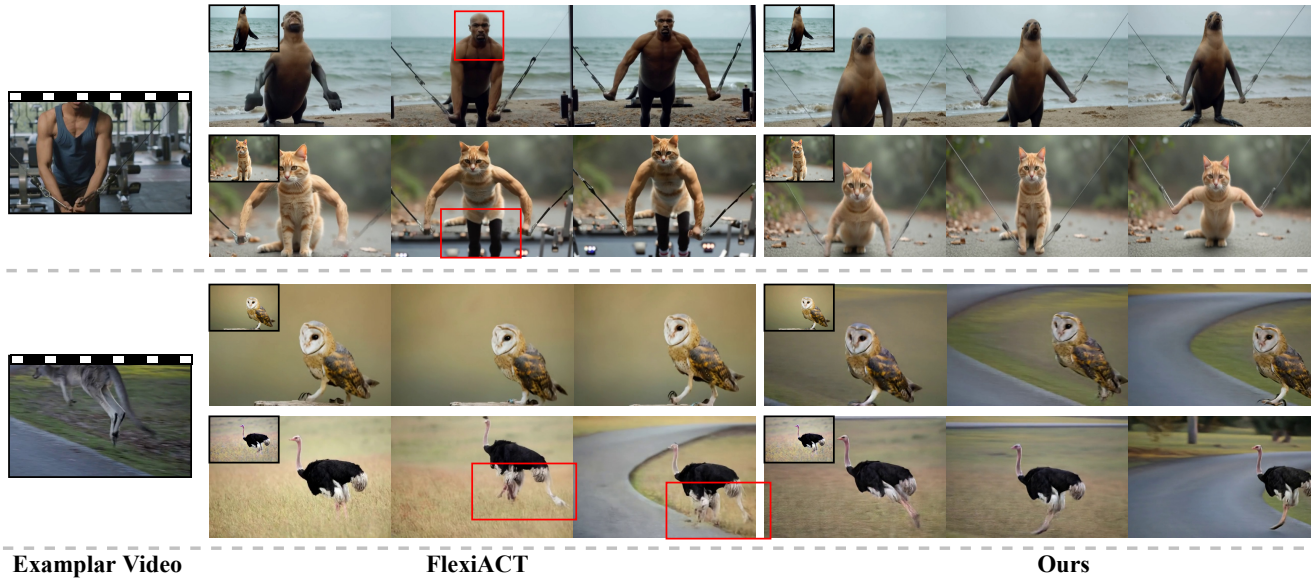


Figure 9. Comparison with FlexiACT on our FlexiACT dataset.

Method	Motion Accuracy	Motion Consistency	Subject Accuracy	Subject Consistency	Imaging Quality	Dynamic Degree	Background Consistency
Baseline	22.22%	<u>98.82%</u>	81.34%	92.86%	47.46%	85.00%	95.01%
B+e^l_{mot}	61.43%	97.97%	57.77%	96.30%	60.36%	**91.11%**	96.39%
B+e^l_{mot} +e^l_{sub}	43.14%	98.34%	91.11%	96.84%	59.99%	71.11%	<u>97.69%</u>
B+e^l_{mot} +e^l_{sub} +$\mathcal{R}$	<u>65.53%</u>	98.87%	<u>95.55%</u>	<u>96.89%</u>	<u>61.62%</u>	62.22%	**98.29%**
SynMotion	**68.60%**	**99.50%**	**97.67%**	**98.26%**	**69.47%**	<u>88.24%</u>	97.59%

Table 3. Ablation study results. The best and second results for each column are **bold** and <u>underlined</u>, respectively.

Figure 10. More interesting results.

Method	Motion Accuracy (↑)	Subject Accuracy (↑)	CLIP-T (↑)	FVD-3DRN50 (↓)	FVD-3DInception (↓)	Flow Score (↓)	Temporal Consistency (↑)
B + LoRA Only	52%	76%	0.287 ± 0.00	389.59	3393.93	1.27 ± 0.13	0.98 ± 0.00
Ours wo training strategy	48%	82%	0.291 ± 0.02	365.28	3179.36	0.78 ± 0.07	0.99 ± 0.00
Ours (Full Model)	**68%**	**97%**	**0.322 ± 0.02**	**212.05**	**3129.19**	**0.41 ± 0.02**	**0.99 ± 0.00**

Table 4. Ablation study results.

model architecture, complete experiments, and state-of-the-art results. We provide the details about the mathematical connections between our approach, the loss function, and the underlying diffusion model theory. These connections demonstrate that our framework is grounded in established principles, making it both empirically strong and theoretically supported. The specifics are detailed below:

Our Dual-Embedding mechanism is a principled method to "disentangle subject and motion representations", which is a core weakness in prior approaches. Formally, given a prompt decomposed into subject p_{sub} and motion p_{mot}, the MLLM produces initial embeddings e_{sub} and e_{mot}. Our method introduces learnable residuals e^l_{sub} and $e^l_m ot$ and a refiner $R(\cdot)$. The final embedding e' is formulated as:

$$e' = [e_{mot} + \mathcal{Z}(e^l_{mot}), \quad e_{sub} + \mathcal{Z}(e^l_{sub})] + \mathcal{Z}(\mathcal{R}([e_{mot} + \mathcal{Z}(e^l_{mot}), e_{sub} + \mathcal{Z}(e^l_{sub})])) \tag{1}$$

To prevent semantic interference, we introduced a new training strategy to regularize the embedding space. Given the user-provided custom data $\mathcal{D}_{custom}$ and our constructed SPV data $\mathcal{D}_{SPV}$, the objective $\mathcal{L}_{total}$ is:

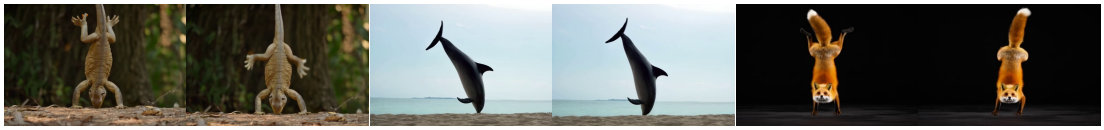
$$\mathcal{L}_{total} = \alpha \cdot \mathbb{E}_{(x,p)\sim\mathcal{D}_{custom}}[\mathcal{L}_{diff}(x, p; \theta_{mot}, \theta_{sub})] + (1-\alpha) \cdot \mathbb{E}_{(x',p')\sim\mathcal{D}_{SPV}}\big[\mathcal{L}_{diff}(x', p'; \text{stop_grad}(\theta_{mot}), \theta_{sub})\big] \tag{2}$$

Here, $\mathcal{L}_{diff}$ is the standard diffusion loss, α is the sampling probability, and θ_{mot} and θ_{sub} represent the parameters for motion and subject embeddings, respectively.

3.2. Action selection

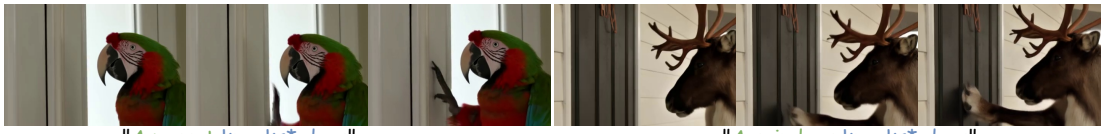
The actions in this work are deliberately selected because they are recognized challenges for video generation, involving intricate physical dynamics [7, 9]. For example, motions like "*spinning*" and "*handstands*" are notoriously difficult to synthesize realistically. The failures of state-of-the-art models. We provide visual demonstrations of these complex movements in our supplementary materials (*e.g.*, videos/group2/spin/*.mp4). Beyond the inherent complexity of these actions, our framework introduces a

"handstand*"



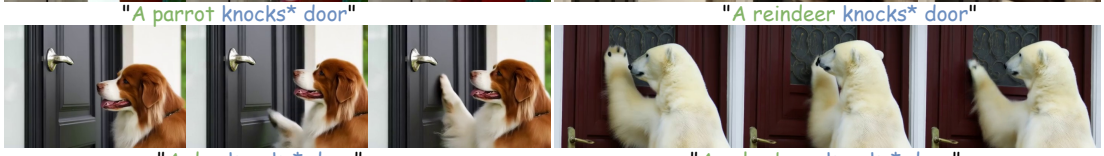
"A lizard handstands*" "A dolphin handstands*" "A fox handstands*"

"A leopard handstands*" "A zebra handstands*" "A dog handstands*"

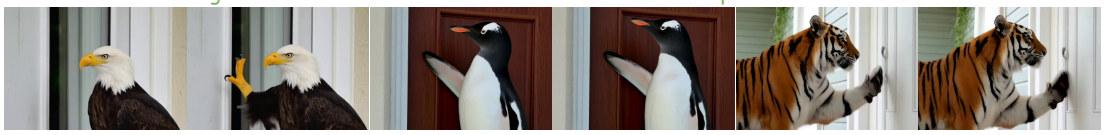


"A parrot knocks* door" "A reindeer knocks* door"

"knock* door*"



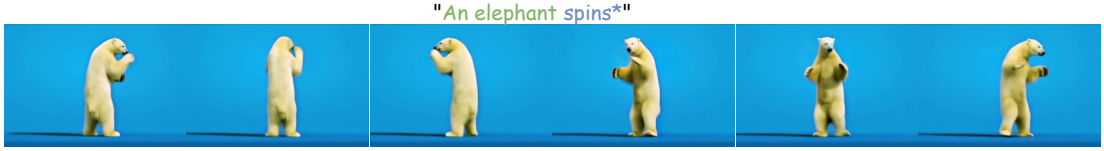
"A dog knocks* door" "A polar bear knocks* door"



"A bald eagle knocks* door" "A penguin knocks* door" "A tiger knocks* door"

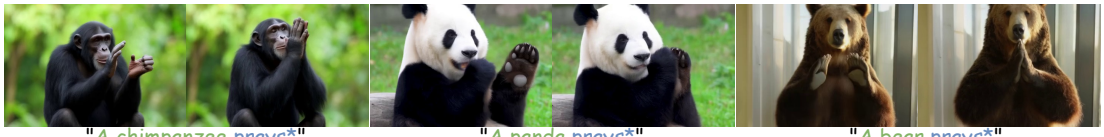
"An elephant spins*"

"spin*"



"A polar bear spins*"

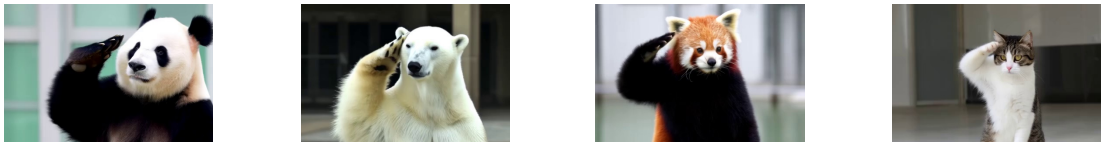
"A cat spins*" "A donkey spins*"



"A chimpanzee prays*" "A panda prays*" "A bear prays*"

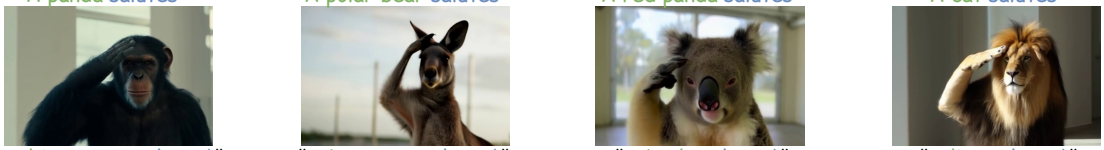
"pray*"

"A cat prays*" "A dog prays*" "A tiger prays*"



"A panda salutes*" "A polar bear salutes*" "A red panda salutes*" "A cat salutes*"

"salute*"



"A chimpanzee salutes*" "A kangaroo salutes*" "A koala salutes*" "A lion salutes*"

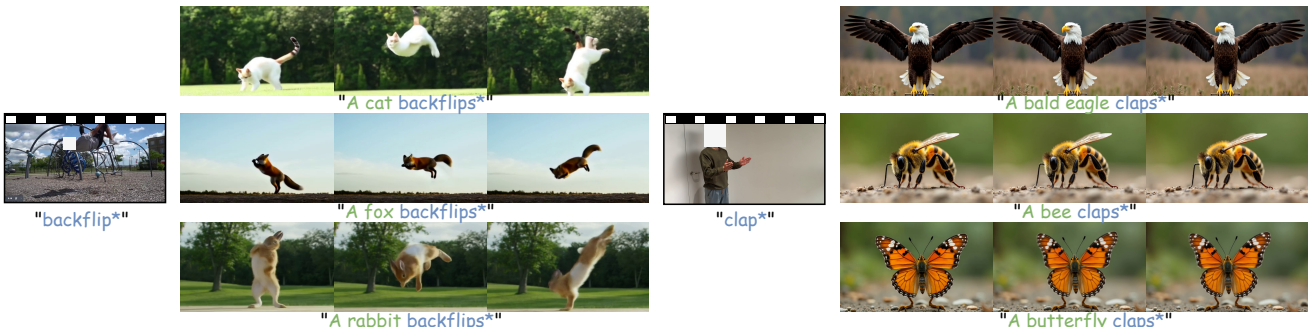


Figure 12. Failure cases generated by our method.

second layer of difficulty: cross-domain motion transfer. We task our model with applying these motions to novel subjects, such as animals performing unnatural behaviors (*e.g.*, "*a bear doing a handstand*"), which is a recognized and significant challenge in itself [4, 8].

To further validate our method's robustness, we benchmark it against FlexiACT [16] on their challenging dataset. This benchmark is particularly suitable as it contains a wide array of complex real-world motions, such as "*a kangaroo jumping across the screen*" and various fitness exercises. The table below presents a direct comparison, showing that our model consistently outperforms the original FlexiACT method. This confirms its strong capability to handle diverse and complex motion generation tasks.

3.3. Identity selection

The majority of our experiments deliberately focus on animal subjects. This choice is strategic, as animals serve as a rigorous testbed for motion customization. In contrast to human subjects, whose anatomy and motions are densely sampled in training data, animals exhibit a vast diversity of morphologies and kinematics that are often underrepresented. Consequently, successfully transferring motion to such out-of-distribution subjects constitutes a more stringent validation of a model's generalization capabilities. We consider our comprehensive investigation into this challenging domain to be a key contribution. Furthermore, to demonstrate the broad applicability of our approach, we also showcase its performance on a wide range of non-animal examples in Fig. 10.

3.4. Diverse and complex input

First, our current work focuses on solving the fundamental problem: customizing a single, novel motion concept where none existed before. Composing multiple motions is an important but distinct incremental challenge that can be built upon our framework. We believe establishing a robust solution for this fundamental task is a critical and necessary contribution. **Second**, we deliberately design the simple

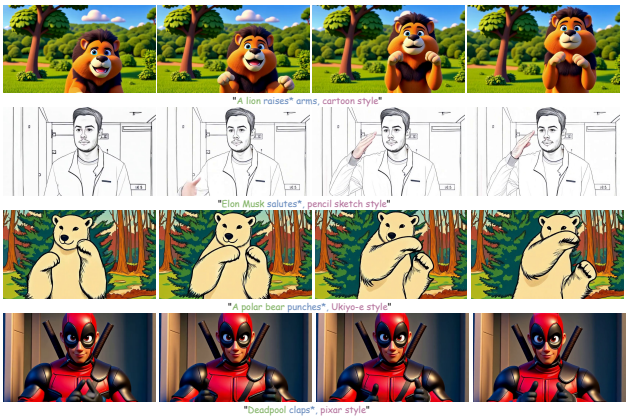


Figure 13. Stylized results. To demonstrate deverse and complex input, we take style instruction with $<$ subject, motion, style instruction $>$ as examples.

$<$ subject, motion $>$ format to create a clean, generalizable paradigm applicable to both T2V and I2V settings. This minimalist design provides a strong foundation without introducing confounding variables and additional external information. **Third**, the formulation does not preclude the adaptability to diverse or complex inputs. Our framework is inherently compatible with extended prompts like $<$ subject, motion, optional instruction $>$. We conduct preliminary experiments demonstrating that our model can successfully incorporate additional instructions for style (*e.g.*, "cut paper style") or background. As shown in Fig. 13, As shown in Fig. 13, the model not only preserves the correct identity and performs the specified motion customization but also seamlessly integrates the designated style, confirming the flexibility of our approach to handle more complex user inputs.

3.5. Discussion about related methods

Training-based approaches typically achieve precise control by explicitly teaching the model to interpret specific motion signals. For example, MotionCtrl [11] enhances a U-Net-based video diffusion model with Camera Motion and Ob-

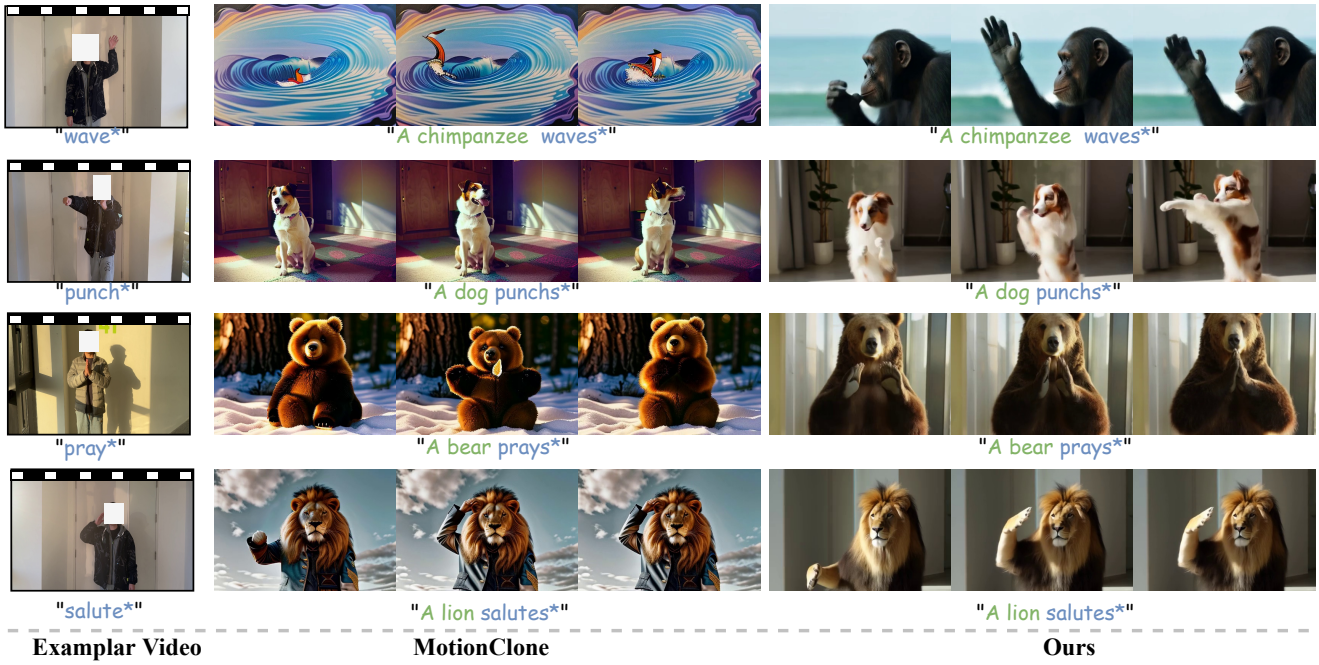


Figure 14. Comparison with MotionClone.

Method	CLIP-T (↑)	FVD-3DRN50 (↓)	FVD-3DInception (↓)	Flow Score (↓)	Temp. Consist. (↑)
MotionClone	0.296 ± 0.01	405.43	4619.52	1.48 ± 0.43	0.98 ± 0.00
Ours	**0.322 ± 0.02**	**212.05**	**3129.19**	**0.41 ± 0.02**	**0.99 ± 0.00**

Table 5. More quantitative comparison with MotionClone.

ject Motion Control Modules, trained respectively to process camera poses and object trajectories. This enables fine-grained control of both global and local movements. For transformer-based models that jointly process spatiotemporal information, VD3D [1] introduces a ControlNet-like mechanism that injects spatiotemporal camera embeddings into patch tokens during fine-tuning.

In contrast, training-free methods leverage the intrinsic capabilities of pre-trained models. MOFT [13] demonstrates that robust motion-aware features already exist within video diffusion models and proposes extracting them by removing content correlation. These features then guide generation through latent optimization without additional training. MotionClone [6] suggests that motion is mainly driven by sparse, dominant components in temporal attention maps and presents a framework that guides new video generation by aligning its temporal attention weights with those of a reference video.

While these methods provide valuable insights, MotionCtrl and VD3D primarily address camera control, which differs from our focus on motion customization. MOFT operates in a different setting, requiring user-defined masks and a dedicated motion feature extraction pipeline. Therefore, we select MotionClone as the most relevant baseline for comparison. Since its experimental settings are not entirely consistent with ours, we have not included it in the main paper.

As shown in Fig. 14, even when MotionClone is provided with an exemplar video, its results remain strongly influenced by the input text and fail to reproduce certain desired motions. Moreover, the generated videos exhibit noticeably lower visual quality. In contrast, our method accurately customizes and transfers the specified motions to the target subject, producing high-quality video results. The quantitative results in Tab. 5 further demonstrate that our approach outperforms MotionClone across all evaluation metrics.

3.6. Ethical Considerations

Our work on motion-customized video generation offers several potential benefits across a range of domains, including creative content production, virtual avatars. How-

ever, this technology also raises potential concerns. The ability to generate videos where arbitrary subjects perform customized actions could be misused for deepfake generation, disinformation, or defamation, particularly when applied to public figures or unauthorized likenesses. The inclusion of recognizable individuals (*e.g.*, celebrities or fictional characters) in synthetic videos may raise ethical and legal questions regarding consent, intellectual property, and identity rights. To mitigate these risks, we incorporate visible or invisible watermarking into all generated videos, making it easier to trace and verify synthetic content.

3.7. Limitation

Despite the promising results achieved by our framework, there are still several aspects that warrant further investigation. (1) When the reference video involves multiple moving entities or complex scene dynamics, our method may struggle to distinguish and isolate the intended motion pattern. (2) While we utilize vision-language model (VLM)-based evaluation metrics to assess motion accuracy, these automatic metrics do not always correlate with human perception of customization quality.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 11

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[4] Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7797–7806, 2024. 10

[5] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1

[6] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 11

[7] Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, Ajmal Mian, Mubarak Shah, and Chang Xu. Generative physical ai in vision: A survey. *arXiv preprint arXiv:2501.10928*, 2025. 8

[8] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024. 10

[9] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025. 8

[10] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[11] 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. 10

[12] WanTeam. Wan: Open and advanced large-scale video generative models. 2025. 1

[13] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *Advances in Neural Information Processing Systems*, 37:76115–76138, 2024. 11

[14] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2

[15] 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. 2

[16] Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. Flexiact: Towards flexible action control in heterogeneous scenarios. *arXiv preprint arXiv:2505.03730*, 2025. 2, 5, 10

[17] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024. 2