
PolyVivid: Vivid Multi-Subject Video Generation with Cross-Modal Interaction and Enhancement (Supplementary Material)

Anonymous Author(s)

Affiliation

Address

email

1 Overview

In this supplementary material, we offer further details on implementation, present additional experimental results, and provide more comprehensive analyses, structured as follows:

- Implementation details (Sec. 2);
- Multi-modal data curation (Sec. 3);
- More multi-subject comparison results (Sec. 4).
- More visualization results (Sec. 5)
- Limitations and societal impacts (Sec. 6)

A project page is also provided in the supplementary files, where additional results in video format can be found. Due to the 100MB file size limit, the videos on the project page have been downsampled.

2 Implementation details

Progressive training process. To enhance the efficiency of the training process, we divide it into two distinct stages. The first stage focuses on modeling the identity preservation capability, while the second stage targets the modeling of interaction generation. During the initial stage, the model is trained on single-subject data, concentrating solely on learning the target identity information without the complexity of interactions among multiple subjects. This stage involves 5,000 iterations. Once the model has effectively learned identity preservation, we proceed to the second stage, where the model is trained with multiple subjects as inputs. Here, the objective is to learn the interactions between the given subject images while maintaining their identities. This stage also comprises 5,000 iterations. Additionally, due to the extensive number of parameters in the pretrained HunyuanVideo model [10], each training iteration is time-consuming. To address this, in each stage, we initially train the model at reduced sizes for 1,000 iterations (included in the total 5,000 iterations), allowing the model to efficiently grasp the target objectives in less time. Subsequently, for the remaining iterations, we revert to the standard resolution to ensure the quality of the final output. All training processes are conducted on 256 GPUs, each with more than 80GB of memory, using a batch size of 256.

Evaluation Metrics. To comprehensively assess the performance of video customization, we adopt several metrics focusing on identity preservation, text-video alignment, and overall video quality:

- **Identity Similarity.** We utilize Arcface [4] to extract facial embeddings from both the reference image and each frame of the generated video, and then compute the average cosine similarity to evaluate how well the identity is preserved.

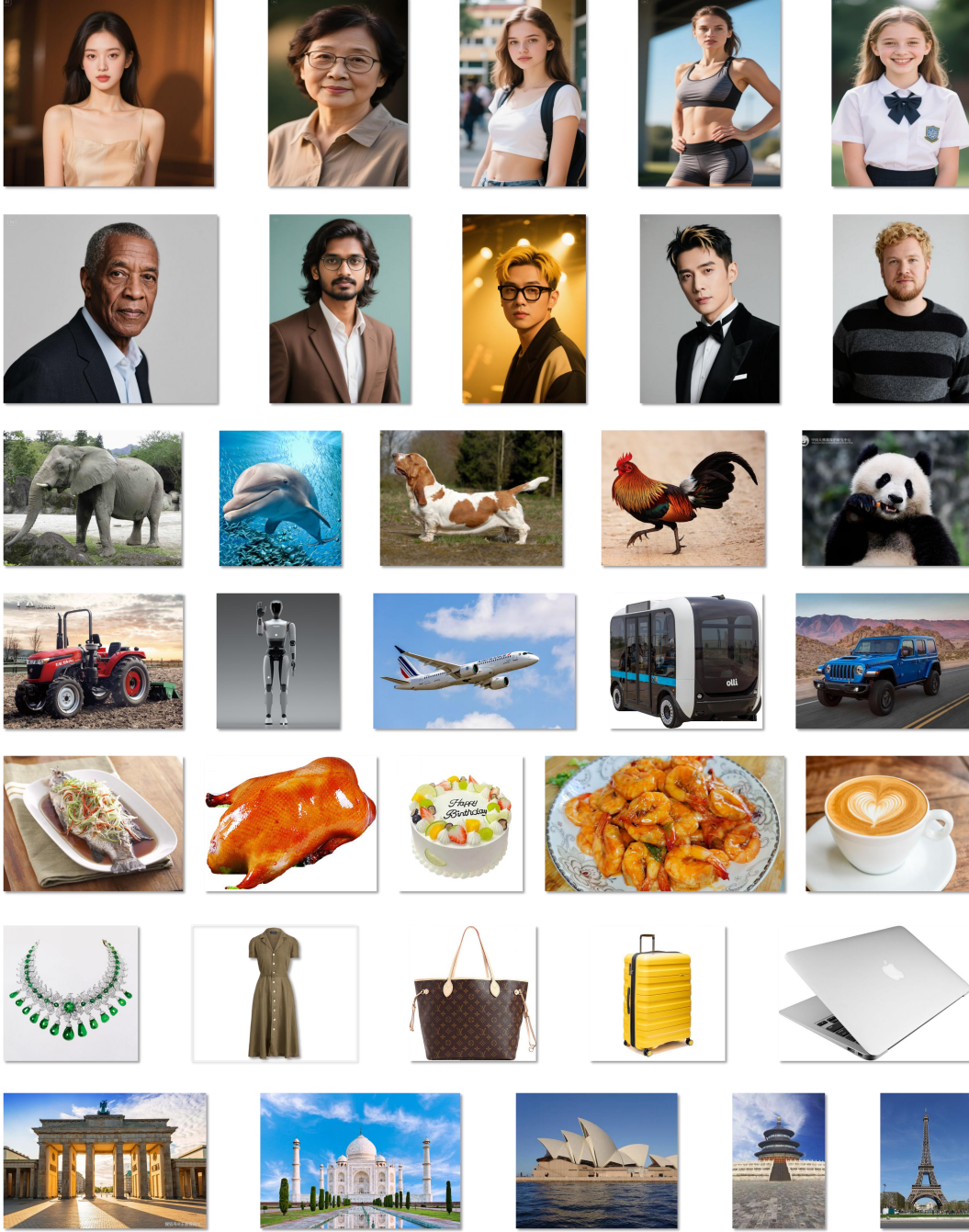


Figure 1: Examples of the test set, which contains images from diverse categories, such as human, animal, man-made machine, food, goods, and building.

- 31 • **Subject Similarity.** Each frame is segmented using YOLOv11 [9] to isolate the subject,

32 after which DINO-v2 [11] features are extracted. The similarity between these features and

33 those from the reference is calculated to assess subject consistency.
- 34 • **Text-Video Alignment.** We employ CLIP-B and CLIP-L [13] to measure the correspon-

35 dence between the provided text prompt and the generated video, evaluating how accurately

36 the video reflects the textual description.

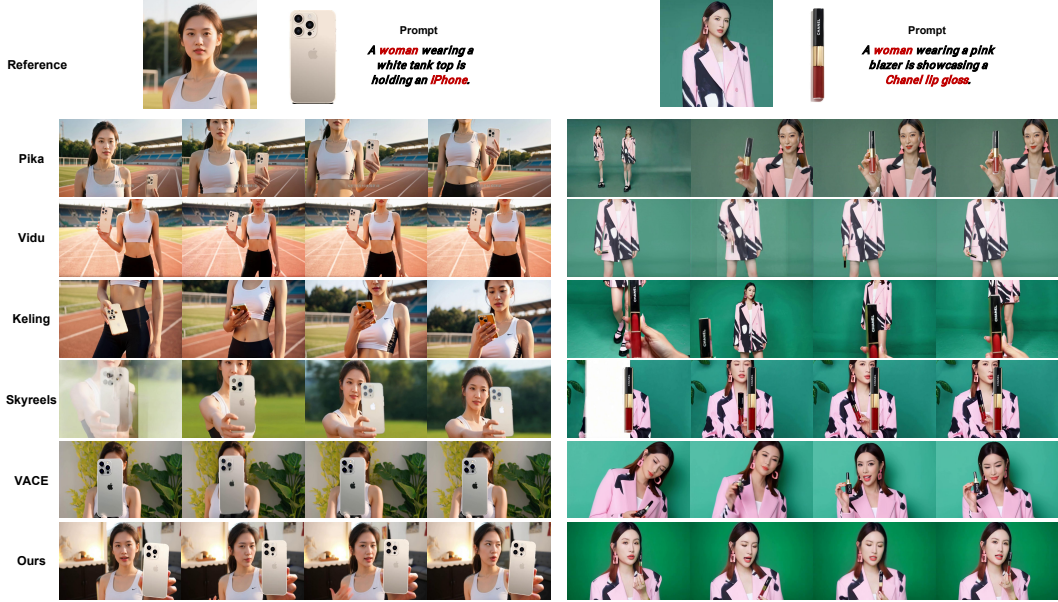


Figure 2: Comparison on human-object customization.

- **Fréchet Video Distance (FVD).** To assess the quality and diversity of the generated videos, we compute the FVD between generated and real videos. Video features are extracted using I3D [2], and the Fréchet Distance is then calculated.
- **Temporal Consistency.** Following the approach in VBench [6], we use the CLIP model to compute the similarity between each frame and its adjacent frames, as well as between each frame and the first frame, to evaluate the temporal coherence of the video.

Test Dataset. We manually collected 100 images of various objects, covering a wide range of categories such as man-made machines, food, goods, and buildings. In addition, we generated 100 human images using an image generation model. These images were then randomly paired to form 100 image pairs. For each pair, we utilized QWen2.5-VL [1] to generate corresponding interaction text prompts.

3 Multi-subject data curation

In the main paper, we have illustrated the MLLM-based Subject Segmentation stage and the Clique-based Subject Consolidation. In this section, we give more details for the preprocessing process, including the data source, data filtering and video captioning.

We curate a large set of high-quality data from open-source datasets, including Panda-70M [3] and Koala-36M [15], as well as our own collected data. Initially, we split the videos by dividing long videos into shorter clips. We then perform black border detection, subtitle detection, watermark detection, transition detection, and motion detection on these clips. Videos with black borders are cropped, and those with subtitles, watermarks, transitions, or low motion are removed. Further, we utilize Koala-36M [15] to filter out videos with scores below a certain threshold. We then perform structured video captioning on the remaining videos, generating long captions, short captions, and descriptions of background, style, and camera movement for each video. This structured combination is used during training to enhance caption diversity.

4 More multi-subject comparison results

Human-object customization. The ability to generate videos depicting human-object interactions is crucial, with broad applications in fields such as film production and advertising. We present qualitative results of human-object interaction in Fig. 2, where our method is compared against



Figure 3: Comparison on human-human and animal-animal customizations.



Figure 4: Comparison on three-subject customization.

several state-of-the-art approaches, including Pika [12], Keling1.6 [8], Vidu2.0 [14], Skyreels A2 [5], and VACE 1.3B [7]. As shown, Pika, Vidu, and Keling often focus primarily on the object, resulting in the human face disappearing from the generated frames. Skyreels A2, on the other hand, struggles with producing smooth transitions between frames, leading to lower overall video quality. VACE sometimes fails to capture the intended interaction between the human and the object (left example), and occasionally does not preserve the appearance of the specified object of the specified object (right example). In contrast, our model consistently maintains strong subject consistency for both the human and the object, while also generating natural and coherent interaction motions between them.

Human-human & animal-animal customization. We further provide comparative results for human-human and animal-animal customization tasks in Fig. 3. It can be observed that Pika still suffers from incorrect attention, focusing on hands rather than preserving human identities, and introduces artifacts such as generating three giraffes in the right example. Keling copies lighting

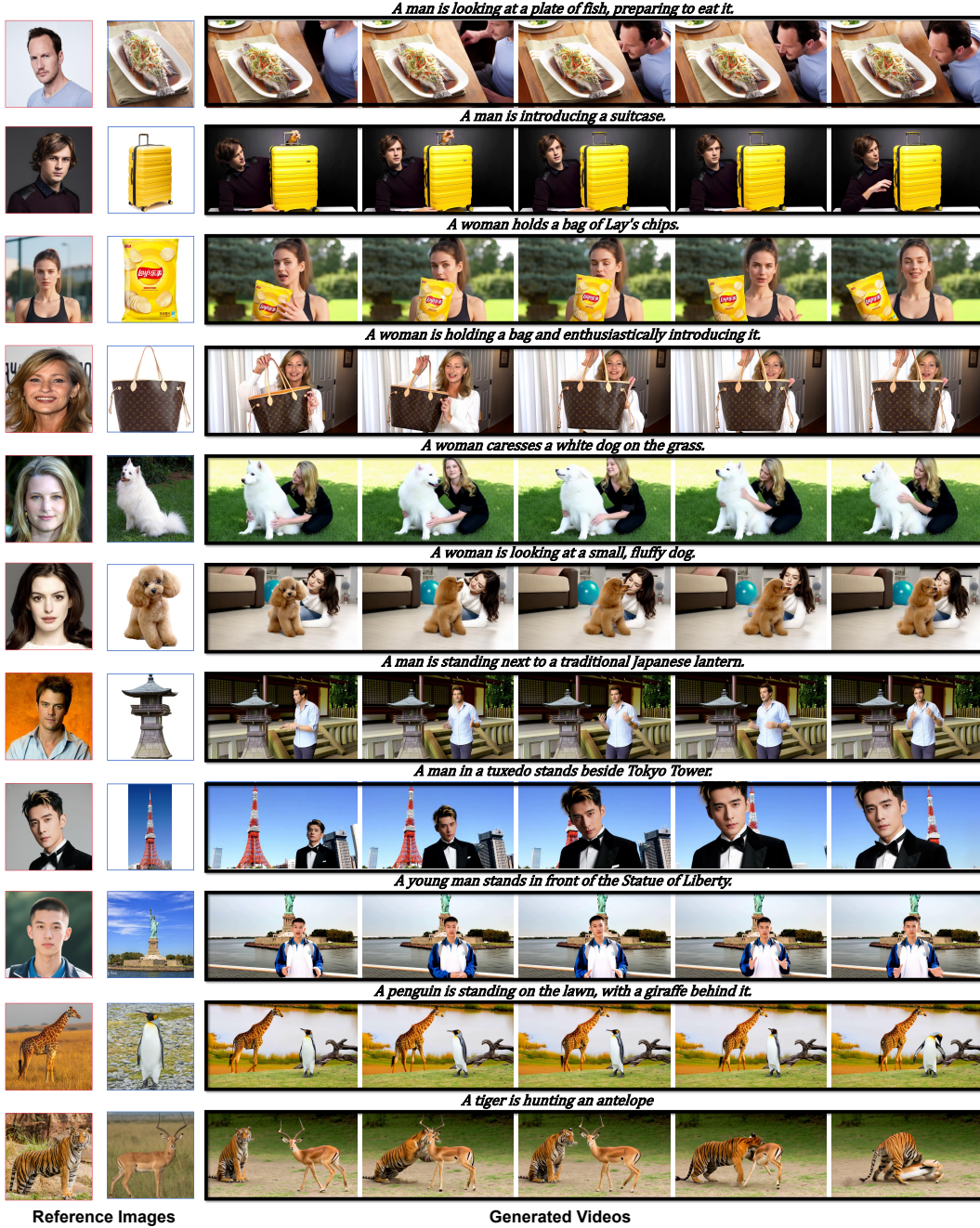


Figure 5: More results on multi-subject customization.

from the background of the man image, which contradicts the prompt 'road', and fails to capture the full body of the giraffe, focusing only on the head without depicting the fighting action specified in the prompt, indicating limited prompt adherence. Skyreels A2 fails to represent both subjects and exhibits poor identity preservation. VACE alters the generated human identities and does not follow the fighting prompt in the giraffe example. While Vidu demonstrates relatively better performance, its identity preservation remains suboptimal. In comparison, our model achieves the best identity consistency and prompt adherence, demonstrating superior capability in customized video generation.

Three-subject customization. Our model is not limited to two-subject customization. We present additional comparison results for three-subject video customization in Fig. 4. As shown, Pika, Vidu, Skyreels A2, and VACE all exhibit significant identity loss. While Pika and Vidu are able to



Figure 6: More results on three-subject customization.

87 generate correct interactions that adhere to physical rules, Skyreels A2 and VACE produce unrealistic
 88 frames in which the person and tiger appear pasted into the sky, violating physical plausibility.
 89 Keling maintains a relatively good identity preservation, but there is still room for improvement. In
 90 contrast, our model achieves the best identity preservation and is able to generate realistic interactions
 91 among multiple subjects while following physical rules, demonstrating our superior capability in
 92 multi-subject customization.

93 5 More visualization results.

94 In this section, we present additional visualization results of our model, covering a wide range of
 95 subject customization scenarios, including human-object interaction, human-scene interaction, human-
 96 animal interaction, and animal-animal interaction. We also provide more examples of three-subject
 97 customization.

98 The two-subject customization results are shown in Fig. 5. It can be observed that our model is
 99 capable of generating natural and realistic interactions between various types of inputs, demonstrating
 100 its potential effectiveness in applications such as advertising and movie production. Furthermore,
 101 beyond object interactions, our model can also generate specified subjects within assigned scenes,
 102 which is particularly useful for personalized content creation and other creative industries.

Next, we showcase more results of three-subject customization in Fig. 6, featuring diverse combinations such as human-animal-animal, human-object-animal, human-animal-scene, and human-object-object. These results illustrate that our model can effectively handle different combinations of inputs and generate complex interactions among multiple subjects, all while maintaining strong identity preservation. This demonstrates the superior capability of our model in customized video generation for multi-subject scenarios.

6 Limitations and societal impacts

Limitations. Despite the significant advancements introduced by PolyVivid in multi-subject video customization, several limitations remain. First, the quality and controllability of the generated videos are still constrained by the capabilities of the underlying video generation backbone and the pre-trained MLLM and VAE models. Second, while our framework demonstrates strong performance on a variety of subject types and interactions, it may encounter difficulties when handling highly complex scenes involving numerous subjects, intricate backgrounds, or fine-grained interactions that require detailed physical reasoning. Finally, although our MLLM-based data curation pipeline improves subject discriminability, it may still be susceptible to errors in grounding or segmentation, especially in cases of occlusion or ambiguous visual cues, potentially affecting the accuracy of subject alignment and interaction modeling.

Societal Impacts. PolyVivid enables more flexible and controllable video generation, which can benefit a wide range of applications, including creative content production, personalized education, digital marketing, and virtual reality experiences. By allowing users to customize videos with specific subjects and interactions, our framework empowers artists, educators, and businesses to efficiently create tailored visual content.

References

- [1] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [3] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-w. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.
- [4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [5] Z. Fei, D. Li, D. Qiu, J. Wang, Y. Dou, R. Wang, J. Xu, M. Fan, G. Chen, Y. Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.
- [6] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. 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.
- [7] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.
- [8] Keling. Keling. <https://klingai.com/cn/>, 2025.
- [9] R. Khanam and M. Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

- 149 [10] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, et al.
150 Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint*
151 *arXiv:2412.03603*, 2024.
- 152 [11] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza,
153 F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision.
154 *arXiv preprint arXiv:2304.07193*, 2023.
- 155 [12] Pika. Pika. <https://pika.art/>, 2025.
- 156 [13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell,
157 P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision.
158 In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- 159 [14] Vidu. Vidu. <https://www.vidu.cn/>, 2025.
- 160 [15] Q. Wang, Y. Shi, J. Ou, R. Chen, K. Lin, J. Wang, B. Jiang, H. Yang, M. Zheng, X. Tao, et al.
161 Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions
162 and video content. *arXiv preprint arXiv:2410.08260*, 2024.