Supplementary Materials

1. S2I+I2V *vs* S2V

Figure 9. Comparison of subject-to-image-to-video [9] and subject-to-video (ours).

As mentioned in the main text, combining subject-to-image (S2I) and image-to-video (I2V) can achieve similar effects to subject-to-video (S2V), but there are some difficult limitations. Firstly, existing methods [9, 14, 21] for generating subject-consistent images or ID-consistent images still exhibit noticeable artificial artifacts, and there is significant room for improvement in the dimension of subject consistency. Equally important, I2V cannot ensure consistency of the subject during motion. As illustrated in Figure 9, when inputting a reference portrait, S2I first generates a reference image for the initial frame of I2V. If the initial frame includes a back view or occlusions, I2V may "imagine" a false ID during the process of removing the occlusion, leading to a failure in maintaining consistency.

2. Copy-paste problem

In the field of video generation, the copy-paste issue is particularly prominent, manifesting as the leakage of image content into the generated video. Some methods sample keyframes from a video and use them as image conditions to reconstruct the video. However, this approach allows the model to employ shortcut learning strategies, simplifying the content understanding process. Figure 10 shows examples of the copy-paste issue, sampling from the initial, middle, and final frames: In the first row, the girl's expression

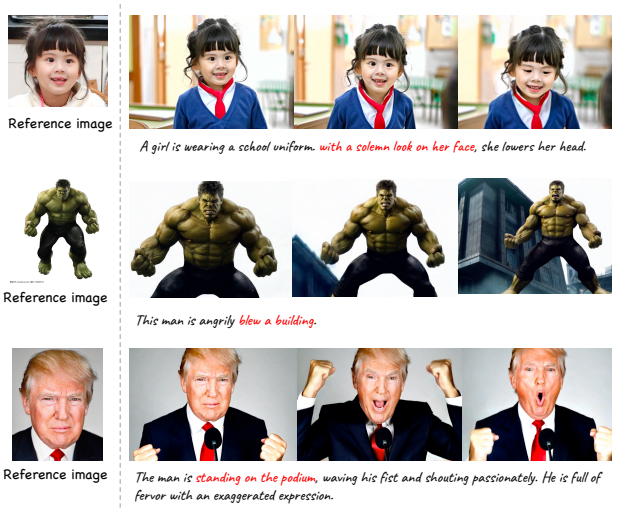


Figure 10. Intuitive cases of copy-paste problems. The red font in the text prompt does not function as intended.

remains unchanged, ignoring the text prompt. In the second row, the cartoon character's movements remain stiff and identical to the reference. The third row illustrates a common case where the generated video is too similar to I2V, diminishing the effectiveness of scene-related text and reducing content diversity. To address this, we focus on constructing cross-video multi-subject pairings, ensuring subjects match in content while allowing for non-rigid deformations and changes in color distribution, thereby avoiding the copy-paste problem.

3. Ablation study supplement

Multi-subject confusion issue. When multiple reference subjects are input simultaneously, appearance confusion may occur. Our solution aligns text descriptions with video subjects during training, ensuring distinct descriptions for each subject. During inference, a rephraser adjusts the input text prompts to align with the training data format. For example, in the first row of Figure 12, the original prompt "A family of three is having a meal at the table" caused confusion. The rephrased prompt "a woman in black, a young girl in white, and an elderly man in a suit eating together at the table" resolved this issue. In the second row of Figure 12, the original prompt "a girl in casual clothes walking by the beach" failed to match the reference. The rephrased prompt "a girl in a white T-shirt and jeans walking by the beach"

	w/o text-image alignment	***w/*** text-image alignment
Success rate	65%	95%

Table 4. Success rate of multi-subject generation with and without text-image alignment.

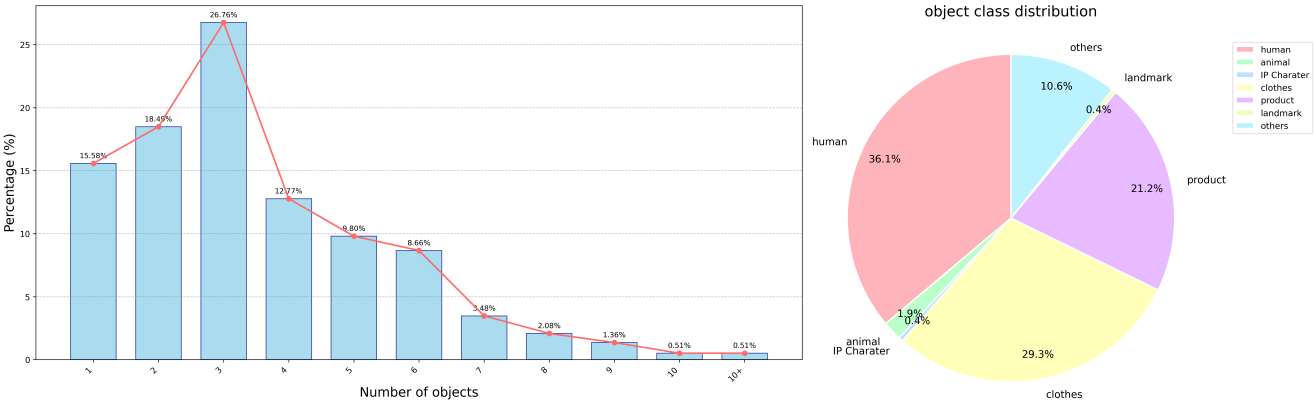


Figure 11. Distribution of object frequencies and class.

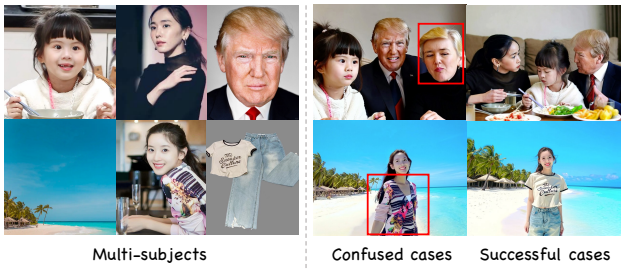


Figure 12. Examples of multi-subject confusion: On the left are the multi-subject reference images, while the right columns present the cases of confusion and the successful cases after improvement.

successfully matched the reference. Quantitative analysis, shown in Table 4, indicates a significant increase in the success rate of subject-consistent generation with this method. Aligning image and text is crucial for multi-subject generation tasks. This approach, which requires no additional complex data structures or model designs, significantly optimizes the multi-subject confusion problem.

4. Data pipeline for face ID

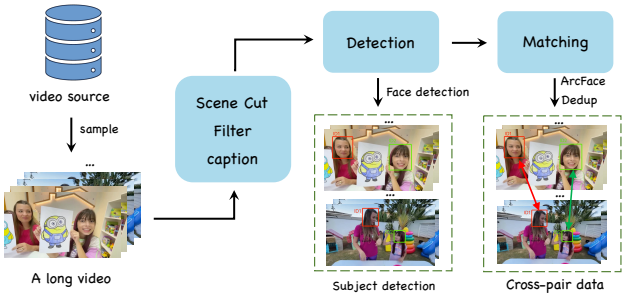


Figure 13. Facial data processing pipeline for constructing ID cross-pair

To enhance facial ID consistency, we developed an additional data pipeline for processing facial data. As shown in Figure 13, the facial data pipeline reuses the scene segmentation, video filtering, and annotation steps from the general subject pipeline. During the detection stage, we use an internal facial detection tool to identify each face in the video reference frames and calibrate it with the VLM [1] results from the captions using IOU (Intersection Over Union). In the matching stage, we calculate facial similarity using Arcface [8] features and add a deduplication operator [25] to further calibrate the recognition results.

5. Data distribution

Distribution of video object quantities. We sample three frames at [0.05, 0.5, 0.95] of the video timeline and perform object detection on these frames. We filter out objects that meet the following criteria: (1) objects that are small in size or occupy a small proportion of the frame; (2) objects with a high degree of overlap with other objects; and (3) incomplete objects judged by the VLM [1]. The final distribution of the number of objects per video is shown in the table on the left side of Figure 11.

Distribution of video object types. We use LLM [37] to classify the noun fields in all captions into the following categories: human, animal, clothes, product, landmark, IP character, and others. The distribution is shown in the accompanying Figure 11, with human, clothes, and product categories accounting for the majority.

6. Model architecture

The architecture of the *Phantom* model is shown in Figure 14, which supplements the missing details in the main text. As illustrated, it integrates the VAE and CLIP encoders to process reference images, while the text encoder handles captions. The encoded features are combined with added noise and processed through multiple MMDiT blocks, resulting in the final output. This design ensures a bal-

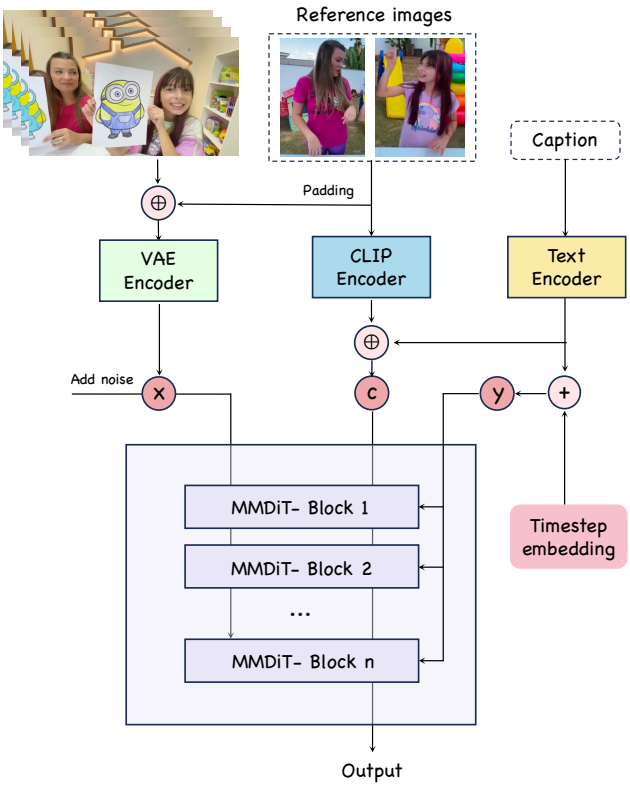


Figure 14. The supplementary diagram of the Phantom framework.

ance between detailed reconstruction and high-level information alignment while also guaranteeing a unified training paradigm for single and multiple subject inputs.

7. Qualitative analysis

Qualitative comparison results of single-subject consistency generation are shown in Figure 15. Firstly, Vidu [54] performs well in both image consistency and text following for the first two cases but fails in the third shoe case with two different seeds. The effectiveness of Pika [40] is evident, as the first two cases show significant disadvantages in maintaining subject consistency, tending towards a cartoonish appearance. The major issue with Kling [27] is that most cases resemble the I2V mode, where the initial frame directly replicates the reference image (as indicated by the red box in Figure 15), followed by subject motion generated based on text, thereby limiting the effectiveness of textual descriptions.

Figure 16 displays some qualitative comparisons of multi-subject consistency generation. Firstly, Kling still reflects the I2V pattern, appearing unnatural transitions in the first few frames of the video. Additionally, in the second example with three reference images of persons, confusion issues are evident in all methods except ours. Vidu shows the first man's clothes and the second man's face, and includes a person unrelated to the reference images. Pika misses one person, and Kling also lacks one person and shows the same issue as Vidu. The final case demonstrates that Vidu and Pika appear more realistic, indicating that their text responsiveness is stronger than their subject consistency.

8. Limitations and future work

Limitations. While *Phantom* demonstrates strong performance in subject-consistent video generation, several challenges persist. First, handling uncommon subjects (e.g., rare animals or niche objects) remains difficult due to biases in training data coverage. Second, complex multi-subject interactions (e.g., overlapping movements or fine-grained spatial relationships) often lead to partial confusion or inconsistent relative subject sizes. Third, generating videos that strictly adhere to intricate text responses (e.g., precise spatial layouts or nuanced temporal dynamics) is limited by the current cross-modal alignment mechanism. These issues stem from three core factors: (1) gaps in dataset diversity, particularly for non-human-centric scenarios; (2) the inherent rigidity of the reference image injection strategy, which struggles to disentangle entangled features from multiple subjects; and (3) biases inherited from pre-trained base models and visual encoders, such as CLIP's semantic oversimplification and VAE's over-referenced details.

Future work. Addressing these limitations will require multi-faceted innovations, and we propose the following directions:

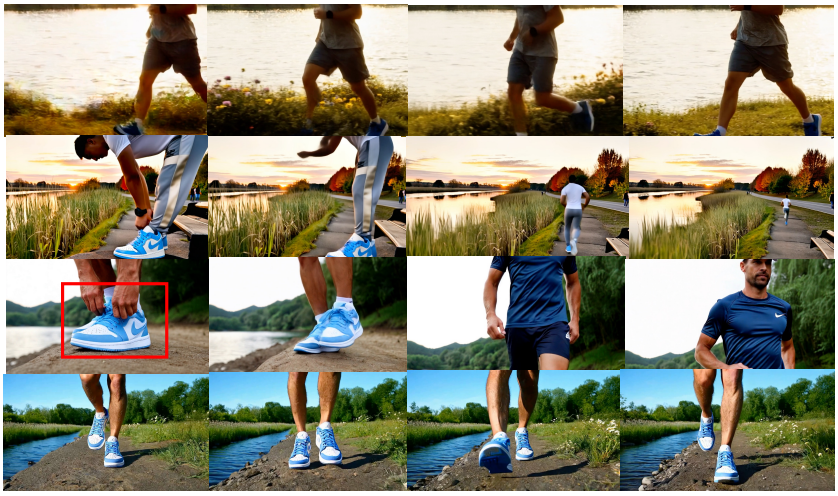
- Enhanced Cross-Modal Alignment: Develop adaptive injection mechanisms that dynamically prioritize text or image conditions based on task requirements, reducing content leakage and improving text responsiveness.
- Spatiotemporal Disentanglement: Integrate spatial-aware attention modules and physics-inspired motion priors to better model multi-subject interactions and enforce consistent relative scales.
- Bias-Aware Training: Mitigate dataset and model biases through adversarial debiasing techniques and synthetic data augmentation for underrepresented subjects.
- Granular Control: Explore auxiliary control signals (e.g., depth maps, segmentation masks) to complement text prompts, enabling precise alignment with complex instructions.
- Foundation Model Adaptation: Fine-tune pre-trained encoders on domain-specific data (e.g., medical imaging, animation) to broaden *Phantom*'s applicability while preserving generalization.

By advancing these areas, *Phantom* could evolve into a versatile tool for industrial applications such as virtual try-ons, interactive storytelling, and educational content creation, ultimately narrowing the gap between academic research and real-world demands.

Acknowledgments We would like to express our gratitude to the Bytedance-Seed team for their support. Special thanks to Haoyuan Guo, Zhibei Ma, Sen Wang and Lu Jiang for their assistance with the model and data. In addition, we are also very grateful to Siying Chen, Qingyang Li, and Wei Han for their help with the evaluation.

The sky is drizzling, and a woman is walking in a retro alley, reaching out to catch the rain. The raindrops are clearly visible.

The rabbit bounces on the soft mattress and falls from the bed to the wooden floor.



The man puts on the sneakers and starts jogging along the riverbank.

Figure 15. Comparative results of single reference subject-to-video generation.

Figure 16. Comparative results of multi-reference subject-to-video generation.