

Phantom: Subject-Consistent Video Generation via Cross-Modal Alignment

Lijie Liu* Tianxiang Ma* Bingchuan Li[†] Zhuowei Chen* Jiawei Liu
 Gen Li Siyu Zhou Qian He Xinglong Wu

Intelligent Creation Team, ByteDance

<https://phantom-video.github.io/Phantom/>

Abstract

The continuous development of foundational models for video generation is evolving into various applications, with subject-consistent video generation still in the exploratory stage. We refer to this as Subject-to-Video, which extracts subject elements from reference images and generates subject-consistent videos following textual instructions. We believe that the essence of subject-to-video lies in balancing the dual-modal prompts of text and image, thereby deeply and simultaneously aligning both text and visual content. To this end, we propose **Phantom**, a unified video generation framework for both single- and multi-subject references. Building on existing text-to-video and image-to-video architectures, we redesign the joint text-image injection model and drive it to learn cross-modal alignment via text-image-video triplet data. The proposed method achieves high-fidelity subject-consistent video generation while addressing issues of image content leakage and multi-subject confusion. Evaluation results indicate that our method outperforms other state-of-the-art closed-source commercial solutions. In particular, we emphasize subject consistency in human generation, covering existing ID-preserving video generation while offering enhanced advantages.

1. Introduction

The rise of diffusion models [18, 39] is rapidly reshaping the field of generative modeling at an astonishing pace. Among them, the advancements in video generation brought by diffusion models are particularly remarkable. In the visual domain, video generation requires to pay more attention to the continuity and consistency of multiple frames compared to image generation, which poses additional challenges. Inspired by the scaling laws of large language models [37, 52, 59], the focus of video generation has shifted towards investigating foundational large models, similar to

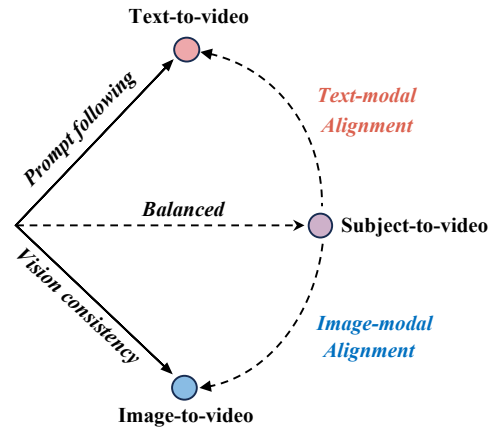


Figure 1. Relationship in cross-modal video generation tasks.

Sora [28, 36, 41, 60, 64], which have demonstrated promising visual effects and are paving the way for a new era in Artificial Intelligence Generated Content.

Currently, foundational video generation models focus mainly on two major tasks: text-to-video[36] and image-to-video [2]. Text-to-video (T2V) leverages language models to understand input text instructions and generate visual content describing the expected characters, movements, and backgrounds. While it allows for creative and imaginative content combinations, it often struggles with generating consistently predictable results due to inherent randomness. On the other hand, image-to-video (I2V) typically provides the first frame of an image along with optional text descriptions to transform a static image into a dynamic video. Although it is more controllable, the content richness is often limited by the strict “copy-paste” [4, 41] nature of the first frame. We term the process of subject-consistent video generation as subject-to-video (S2V) [4, 23, 26, 35], which involves capturing the subject from an image and flexibly generating a video based on text prompts, while combining the diversity and controllability of joint image and text inputs. As shown in Figure 1, its essence lies in balancing

*Equal Contributions

[†]Project Lead



Figure 2. Subject-consistent video generation examples using our method, with reference images and corresponding generated video frames (text prompts omitted). The last three rows show multiple reference subjects.

text-image dual-modal prompts, requiring the model to simultaneously align text instructions and image contents.

However, the research on subject consistency in video generation tasks still lags behind image generation scenarios. As text-to-image (T2I) foundation models [11, 29] have matured, subject-to-image (S2I) has evolved from parameter optimization methods [20, 45] to adapter-based training approaches [21, 61], to unified image editing approaches [5, 15, 58], achieving impressive results (refer to Sec 2.2). The most straightforward way to implement S2V is to combine S2I with I2V, but there are two main limitations. First, S2I has greater difficulty in learning subject consistency compared to S2V, as the S2V training data naturally include multi-view dynamic variations, allowing for better understanding of the subject. Second, transitioning from S2I to I2V can lead to information loss. For instance, when generating a back-to-front view motion, the subject’s ID information may be lost because the first frame lacks it, which hinders I2V from maintaining ID consistency (see supple-

mentary material). Therefore, subject-consistent generation requires a specialized video model for unified processing.

Specifically, the subject-consistent video generation task aims to deeply and simultaneously align the content described in the text and images. To achieve this, we propose a data pipeline for the S2V task, producing training data in the form of text-image-video triplets. Two key issues must be addressed. First, prevent the leakage of image content into the generated video. Some methods [4, 16, 23, 30, 62] sample key frames from a video as image conditions to reconstruct the video, but this allows the model to copy-paste image content, reducing text responsiveness. While some approaches enhance data through transformations [4], they fail to address the rigid properties and overall lighting of the image. We focus on constructing cross-video multi-subject pairs to ensure that subjects exhibit non-rigid deformations and color variations while maintaining content matching. Second, address the issue of confusion arising from multi-subject generation. Specifically, when similar

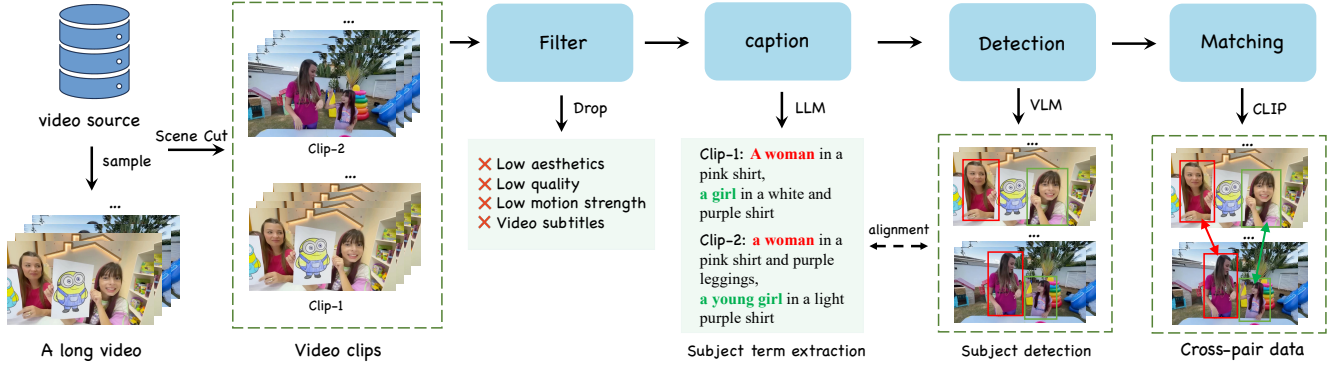


Figure 3. Data processing pipeline for cross-modal video generation. The process involves filtering, adding captions, detection, and matching stages to extract subjects from video clips and align them with the text prompts, ensuring consistent video generation.

subjects trigger identical textual descriptions, it can lead to content ambiguity. To resolve this, we emphasized distinct descriptions of the subjects’ appearances in the video. The appearances of multiple subjects should be distinguishable and precisely match the contents of the sampled reference images. Furthermore, we build a rephraser to rephrase user prompts to include detailed image descriptions.

Our model design is based on two primary considerations: (1) How to simply and effectively extend a video foundation model to support S2V capabilities; (2) How to achieve a unified framework for single- and multi-subject consistency generation. Thus, we redesigned the image-text joint injection model based on pre-trained T2V and I2V models [31] to ensure effective cross-modal learning. Specifically, our method is built on the MMDiT [11] architecture. Referring to [55], full self-attention is replaced with window attention to reduce computational costs. VAE [10] and CLIP [63] are used to encode the reference images, and the encoded results are fed into the video and text branches of MMDiT, respectively. The VAE latent provides low-level detail information, while CLIP offers high-level semantic information. Additionally, we introduced a dynamic information injection strategy during attention calculation, allowing the insertion of one or more reference images without affecting the window size and position encoding [49], achieving a unified model architecture for single- and multi-subject consistent video generation.

In addition, for the S2V task, we constructed evaluation datasets for portrait IDs, single subjects, and multiple subjects, and developed corresponding evaluation metrics. Since the performance of some open-source reproducible projects [4, 16, 23, 30, 62] has not yet matched that of closed-source commercial solutions [27, 35, 40, 41, 54], our focus is on comparing with commercial methods. Overall, our proposed *Phantom* has the following contributions:

Concepts. (1) We are the first to clearly define the subject-to-video (S2V) task and clarify its relationship with

text-to-video (T2V) and image-to-video (I2V); (2) *Phantom* offers a feasible path for S2V, focusing on high-quality alignment of both textual and visual content.

Technology. (1) A new data pipeline constructs cross-modal aligned triplet data, effectively addressing the issues of information leakage and content confusion; (2) *Phantom* offers a unified framework for generating videos from both single and multiple subject references, utilizing dynamic injection scheme of various conditions at its core.

Significance. (1) *Phantom* demonstrates superior generation quality, bridging the gap between academic research and proprietary commercialization; (2) the unified consistency generation paradigm covers subtasks such as ID generation and demonstrates significant advantages, indicating that *Phantom*-like solutions have broad prospects in scenarios such as the film industry or advertising production.

2. Related Work

2.1. Video foundation model

Early latent diffusion models (LDM) [43] typically utilized U-Net [44] architectures, such as the open-source Stable Diffusion 1.5 [43]. Temporal modules were later added to these models, evolving them into video generation models like Make-A-Video [48], SVD [2], and Animatediff [13]. The DiT [39] architecture, guided by scaling laws, has led to the development of more vision foundation models. Among these, Stable Diffusion 3 introduced MMDiT [11], a dual-stream DiT architecture, which has been adopted in open-source video generation projects such as CogvideoX [60], HunyuanVideo [28] and SeedVR [55].

2.2. Subject-consistency image generation

In recent years, significant progress has been made in subject-consistent generation for image tasks. Optimization-based methods [12, 19, 20, 45, 46] training bind image content with special identifiers for text-to-

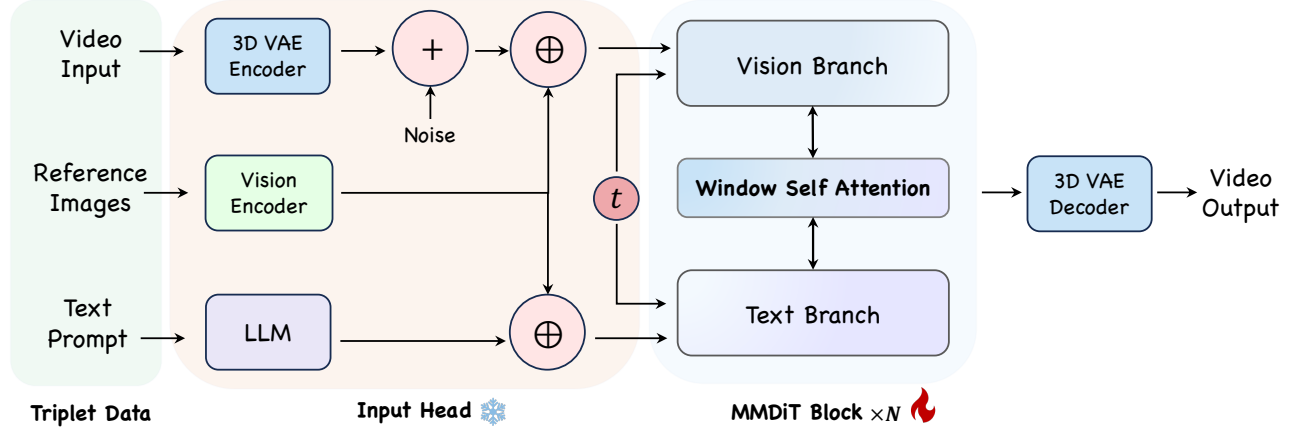


Figure 4. Overview of the *Phantom* architecture. Triplet data is encoded into latent space at the input head, and after combination, it is processed through modified MMDiT blocks to learn the alignment of different modalities.

image generation. A notable work in the training and inference paradigm is IP-Adapter [61], which freezes the base model weights while only training additional adapters to achieve subject-consistent generation. This approach is also widely used in tasks requiring facial ID consistency [6, 14, 56]. However, these solutions often rely on CLIP [7] or DINO [38] for extracting image semantics, leading to a trade-off between low-level detail reconstruction and flexible text response. Recent advancements have unified image generation and editing tasks [5, 15, 34, 58], enabling various types of editing tasks within a single model, including subject-consistent generation. Compared to adapter-based approaches, this method deeply learns image-text alignment, fully leveraging foundation models and resolving degradation issues from multiple adapters.

2.3. Subject-consistency video generation

From recent research developments, the advancement of video generation capabilities and algorithmic innovations tends to lag behind image tasks. Similar to image consistency techniques, Kling [27] has released an optimization-based video generation method for facial ID consistency, which requires uploading multiple videos of the same person for optimization, resulting in significant computational costs. Adapter-based approaches have also been attempted for video ID consistency tasks, such as ID-Animator [16] and ConsisID [62]. However, these works have been validated on small datasets (around 10k), which limits their ability to fully align facial information with text descriptions. Recent works like ConceptMaster [23], MovieWeaver [30], and VideoAlchemist [4] have demonstrated capabilities in generating consistent multi-subject videos in general scenarios. However, there are currently no open-source methods for the S2V task, and commercial software’s S2V capabilities [27, 35, 40, 41, 54] re-

main state-of-the-art. Therefore, comparing the performance with commercial closed-source solutions is crucial for evaluating the superiority of the proposed method.

3. Phantom

This section introduces the specific implementation of *Phantom*. The first subsection describes how to construct cross-modal alignment training data, emphasizing the creation of cross-pair text-image-video triplets to address the “copy-paste” issue. The second subsection presents the design and considerations of the *Phantom* architecture, focusing on how single and multiple subject features are dynamically injected into the framework. The third subsection introduces key training settings and inference techniques to ensure the efficient implementation of S2V capabilities.

3.1. Data Pipeline

To achieve subject-to-video (S2V) generation, we constructed a triplet data structure of text-image-video for cross-modal learning (Figure 3), ensuring that videos are paired with both images and text. First, we sampled long videos from Panda70M [3] and in-house sources. These videos were cut into single-scene segments using AutoShot [65] and PySceneDetect [51], and any clips with low quality, aesthetics, or motion levels were filtered out. Next, we used Gemini [50] to generate captions for the filtered video clips, focusing on describing the subjects’ appearance, behavior, and the scene. Further, the LLM [37] is utilized to analyze the caption and extract the subject words with appearance descriptions, which are used as prompts for the VLM [1] to obtain the subject detection boxes of the reference frames. At this point, the descriptions of the subjects in the captions can be exactly aligned with the subject elements detected in the reference images.

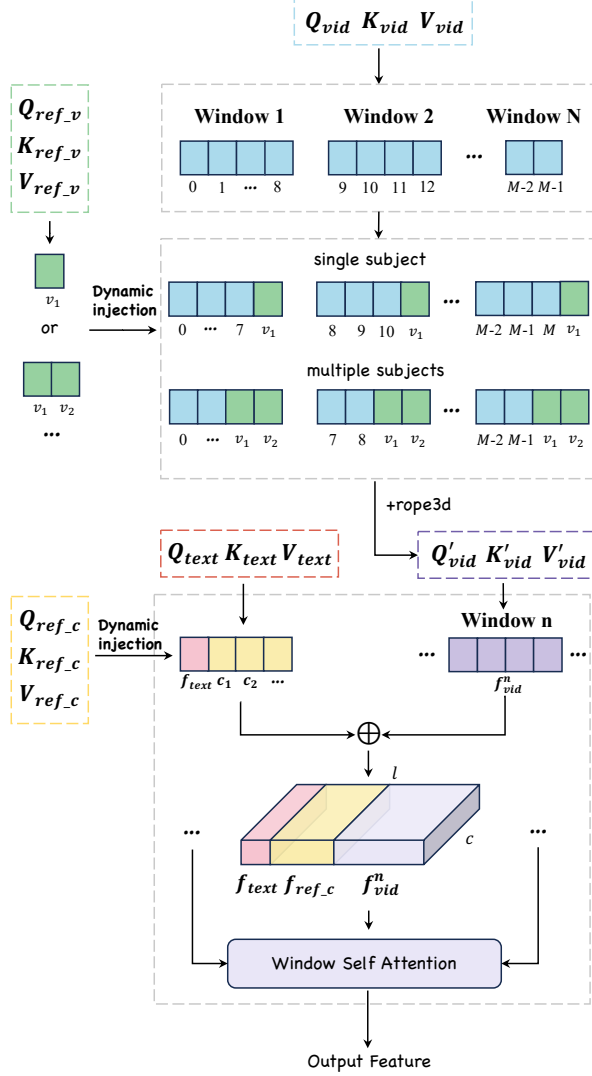


Figure 5. **Dynamic injection strategy** and attention calculation for single or multiple reference subjects in each MMDiT block.

Although the reference images and text are aligned, the reference images are taken from specific frames within the videos. These image-video pairs are termed "in-pair" data. Some existing methods [4, 23] use in-pair data to train S2V models, ensuring subject consistency between images and videos. However, high visual similarity might cause the model to disregard text prompts, resulting in generated videos that simply copy-paste the input images. To address this issue, we undertake an additional effort to further establish pairings between cross-video clips. We employ the image embedder [47] of the improved CLIP architecture to score and pair subjects detected across different videos. Pairs with scores that are excessively high (indicating a likelihood of copy-pasting) or too low (indicating different subjects) are eliminated.

After constructing the cross-paired data pipeline, further segmentation is required based on application scenarios. These primary elements include people, animals, objects, backgrounds, and more. Additionally, interactions between multiple elements can further categorize scenarios, such as multi-person interactions, human-pet interactions, and human-object interactions. By segmenting the data sources according to these application scenarios, we can quantitatively supplement missing data types. For example, virtual try-on applications require specific collections of model images and garment layouts. Ultimately, we obtained cross-pair data on the order of one million, among which the data containing human subjects accounted for the largest proportion. In addition, we also added a portion of paired image data to increase diversity. The data sources are Subject200k [5] and OmniGen [58].

3.2. Framework

The *Phantom* architecture, shown in Figure 4, consists of an untrained input head and a trained MMDiT module. The input head includes a 3D VAE [60] encoder and an LLM [59] inherited from the video foundation model [31, 55], which encode the input video and text, respectively. The vision encoder, critically, comprises both a Variational Autoencoder (VAE) [10] and CLIP [42, 63]. The image features $F_{ref.v}$ concatenated with the video latents F_{vid} reuse the 3D VAE to maintain consistency in the visual branch input. Meanwhile, the image CLIP features $F_{ref.c}$ concatenated with the text features F_{text} provide high-level semantic information, compensating for the low-level features from the VAE. Feature merging involves dimensional alignment, as detailed below,

$$F_T^{l_1+l_2,c} = F_{text}^{l_1,c} \oplus F_{ref.c}^{l_2,c}, \quad (1)$$

$$F_V^{t+n,h,w,c} = F_{vid}^{t,h,w,c} \oplus F_{ref.v}^{n,h,w,c}, \quad (2)$$

where $\oplus$ denotes concatenation. The concatenated features F_T and F_V are fed into the visual and text branches of MMDiT, and the model only separates the injected features during the calculation of attention.

Specifically, the MMDiT block is based on [31, 55] and improved for reference image input, primarily modifying the Attention [53] block, as shown in Figure 5. First, the $Q_{vid}, K_{vid}, V_{vid}$ features calculated from F_{vid} are divided into windows of size 9. Then, the $Q_{ref.v}, K_{ref.v}, V_{ref.v}$ features calculated from $F_{ref.v}$ are dynamically concatenated to the end of each window, while the in-situ features are sequentially shifted to the start of the next window. This approach maintains the window structure while ensuring interaction between video and subject features within each window, as well as adaptive input for single- or multi-subject. Meanwhile, the $Q_{text}, K_{text}, V_{text}$ features calculated from F_{text} and the $Q_{ref.c}, K_{ref.c}, V_{ref.c}$ features calculated from $F_{ref.c}$ are dynamically concatenated. After collecting all reference information, self-attention is calculated within each

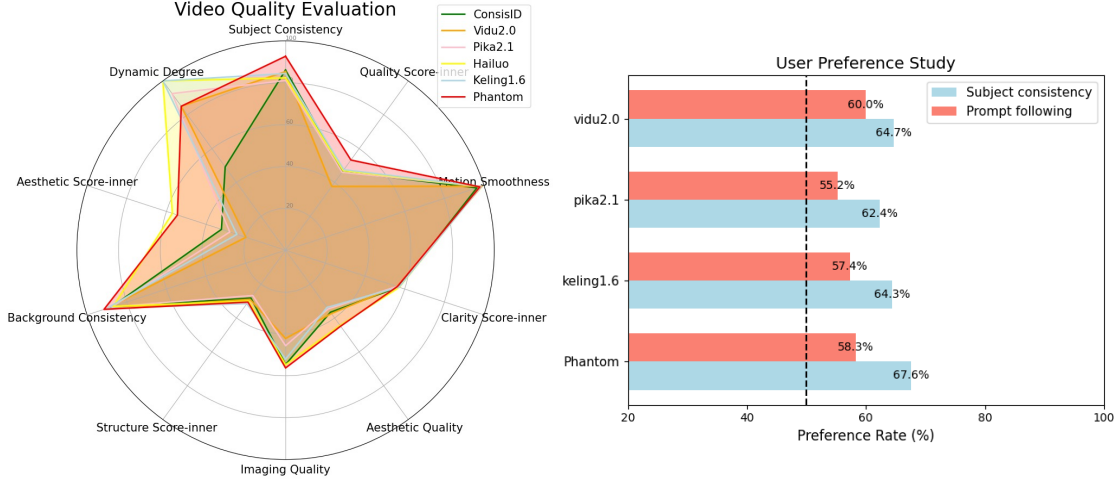


Figure 6. Video quality evaluation (left) and user study results for multi-subject consistency (right).

window. Then, the dynamically injected reference image features (including ref_v and ref_c) and the text features within each window are extracted from the output features and averaged. This process ensures that the dimensions of the input and output features within the current block remain consistent, thereby facilitating subsequent block computations.

3.3. Training and inference

Training setup. We employ rectified flow (RF) [32, 33] to construct the training objective and adjust the noise distribution sampling [11]. RF aims to learn an appropriate flow field, enabling the model to efficiently and high-quality generate meaningful data samples from noise. In the forward process of training, noise is added to clean data x_0 to generate $x_t = (1 - t) \cdot x_0 + t \cdot \epsilon$, where ϵ is Gaussian noise with the distribution $\mathcal{N}(0, I)$ and t is a randomly sampled step scaled to a value between 0 and 1 based on the total steps ($T=1000$). The model predicts velocity v_t to regress velocity $u_t = dx_t/dt$, and v_t is represented by,

$$v_t = \mathcal{G}_\theta(x_t, t, F_T, F_V). \quad (3)$$

Thus, the RF training loss is given by,

$$\mathcal{L}_{\text{mse}} = \|v_t - u_t\|^2. \quad (4)$$

Notably, v_t includes additional (n)-dimensional features at the tail (refer to Eq.2), which does not participate in the loss calculation. The model training is conducted in two phases: the first phase trains for 50k iterations at 256p/480p resolution, and the second phase incorporates mixed 720p data, training for an additional 20k iterations to enhance higher resolution generation capabilities. Additionally, since one of the training objectives of VAE is pixel-level reconstruction, the CLIP features can be overshadowed when trained

together with VAE features. Therefore, we set a relatively high dropout rate (0.7) for VAE during training to achieve balance. The total computational resources consumed approximately 30,000 GPU-hours on A100.

Inference settings. *Phantom* inference can accept 1 to 4 reference images and generate corresponding videos by describing the reference subjects using a given text prompt. Note that generating with more reference subjects may lead to unstable results. To align with the training data, the text prompt used in inference must first be adjusted by a rephraser to ensure it accurately describes the appearance and behavior of each reference subject, avoiding confusion between similar subjects (see supplementary materials). The Euler method is used for sampling over 50 steps, and the classifier-free guidance [17] separates the image and text conditions. The denoised output at each step is given by,

$$x_{t-1} = x_{t-1}^\circ + \omega_1(x_{t-1}^I - x_{t-1}^\circ) + \omega_2(x_{t-1}^{TI} - x_{t-1}^I), \quad (5)$$

where x_{t-1}° is the unconditional denoising output, x_{t-1}^I is the image-conditioned denoising output, and x_{t-1}^{TI} is the joint text-image conditioned denoising output. The weights ω_1 and ω_2 are set to 3 and 7.5, respectively.

4. Experiments

4.1. Evaluation materials

Phantom can be fine-tuned from any video generation base model [31, 55]. The T2V and I2V pre-training stages are excluded from this evaluation. We focus on assessing the subject consistency generation capability, with additional independent evaluations for face ID-based video generation. Due to the lack of an established benchmark for subject-to-

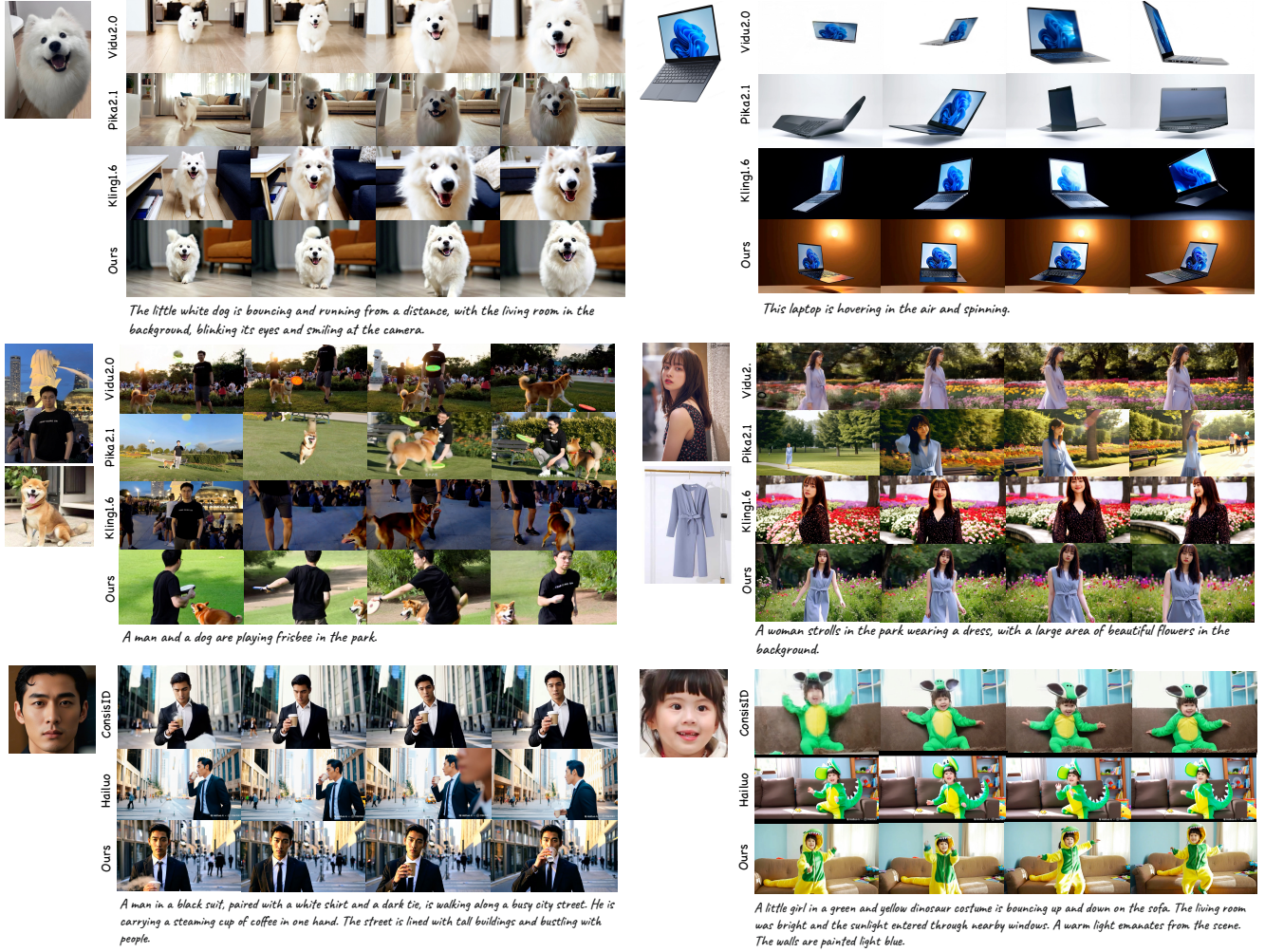


Figure 7. Comparison results showing, from top to bottom, single subject, multi-subject, and facial ID-consistent video generation, with four uniformly sampled frames displayed in each case.

video, we constructed a specific test set and defined evaluation metrics accordingly.

We collected 50 reference images from different scenarios, covering humans, animals, products, environments, and clothing. Each reference image is paired with 3 different text prompts. To ensure confidence in each case, each text-image pair is generated with three random seeds, resulting in a total of 450 videos. For scenarios with multiple reference images, we mixed the aforementioned reference images and rewrote the text prompts to obtain a test set of 50 groups. Additionally, considering the unique value of portrait scenarios, we collected an additional 50 portrait reference images, including both celebrities and ordinary individuals, for independent evaluation of ID consistency.

For the S2V task, the existing available state-of-the-art (SOTA) methods are closed-source commercial tools. Therefore, we evaluated and compared the latest capabil-

ities of Vidu [54], Pika [40], and Kling [27]. For the ID-preserving video generation task, the commercial tool Hailuo [35] demonstrated impressive results. We also evaluated an excellent open-source algorithm ConsisID [62].

4.2. Quantitative results

We classify the S2V evaluation metrics into three major categories: video quality, text-video consistency, and subject-video consistency. First, the visualization of video quality is shown in the radar chart on the left side of Figure 6. We selected six metrics provided by VBench [24] for testing and supplemented them with four inner model scores such as structure breakdown score. For text-video consistency, we used ViCLIP [57] to directly calculate the cosine similarity score between the text and the video. For single subject consistency, we uniformly sampled 10 frames from each video and calculated the CLIP [7] and DINO [38] fea-

Methods	Identity Consistency			Prompt Following
	FaceSim-Arc $\uparrow$	FaceSim-Cur $\uparrow$	FaceSim-glink $\uparrow$	ViCLIP-T $\uparrow$
ConsisID	0.538	0.417	0.470	21.76
Hailuo-ID	0.542	0.504	0.557	23.31
Phantom-ID	0.581	0.529	0.590	24.12

Table 1. Comparison of different methods based on identity consistency and prompt following

Methods	Subject Consistency				Prompt Following
	CLIP-I $\uparrow$	DINO-I $\uparrow$	CLIP-I-Seg $\uparrow$	DINO-I-Seg $\uparrow$	ViCLIP-T $\uparrow$
Vidu2.0	0.706	0.511	<u>0.724</u>	0.544	22.78
Pika2.1	0.697	0.498	0.712	0.534	<u>23.05</u>
Kling1.6	0.732	0.554	0.715	0.569	21.62
Phantom-IP	<u>0.714</u>	<u>0.523</u>	0.731	<u>0.538</u>	23.41

Table 2. Comparison of different methods based on single subject consistency and prompt following. **Boldface** indicates the highest scores in each column, and underline indicates the second-highest scores.

ture Direction Scores with the reference image. Additionally, we used grounded-sam to segment the subject part of the video and calculate the CLIP and DINO scores (excluding scene graphs). For ID consistency, we used three facial recognition models to measure similarity [8, 22].

The video quality evaluation results, shown on the left side of Figure 6, indicate that *Phantom* performs slightly worse [24], while excelling in other metrics. As shown in Table 1 and 2, *Phantom* leads in overall metrics for subject consistency (Identity Consistency) and prompt following. For multi-subject video generation, due to high error rates in automated subject detection and matching, we conducted a user study. We surveyed 20 users, who rated the methods on a scale of 1 to 3 (1: unusable, 2: usable, 3: satisfactory). The evaluation results, displayed in the bar chart on the right side of Figure 6, show that *Phantom*'s multi-subject performance is comparable to commercial solutions, with some advantages in subject consistency.

4.3. Qualitative results

We present the comparison results of several typical cases in Figure 7. Each generated video is displayed with four evenly sampled frames, including the first and last frames. The first two rows of Figure 7 respectively show the results of generating single- and multiple subject consistency. It can be seen that Vidu [54] and *Phantom* exhibit balanced performance in subject consistency, visual effect, and text response. Pika [40] performs poorly in subject consistency. Kling [27] has a notable issue: some cases exhibit characteristics analogous to I2V approaches. For instance, the first frame of character videos almost matches the input reference image, leading to low success rates in virtual try-on scenarios. Additionally, the laptop case shows that the compared methods tend to cause deformations in rigid body movements. The last row of Figure 7 shows the results

Methods	Subject Consistency		Prompt Following	Video Quality	
	CLIP-I $\uparrow$	DINO-I $\uparrow$	ViCLIP-T $\uparrow$	Aes score $\uparrow$	Clarity score $\uparrow$
w/o CLIP	0.693	0.519	23.63	62.03	71.40
w/o VAE	0.512	0.302	22.79	48.82	70.76
w/ All	0.714	0.523	<u>23.40</u>	64.32	71.72

Table 3. The ablation experiment results of VAE and CLIP.



Figure 8. Qualitatively display the ablation of VAE and CLIP.

of video generation for facial ID preservation. The open-source method ConsisID [62] tends to exhibit motion blur, and has weak text response. Hailuo [35] excels in visual aesthetics, but there is some loss in facial similarity. Our results are balanced across all dimensions, with particular advantage in ID consistency. More qualitative analyses are presented in supplementary materials.

4.4. Ablation study

Selection of visual encoder. Due to differences in training methods, CLIP aligns image-text pairs and tends to extract semantic information, while VAE aims for lossless reconstruction and focuses on detailed information. As shown in Figure 8, faces generated using only CLIP features are smoother and more refined but show decreased similarity. In contrast, faces generated using VAE features are sharper but may amplify undesirable details, making them contain more artifacts. In general object scenarios, CLIP is inadequate at reproducing details like text and patterns, thus primarily serving to supplement VAE's high-level information. Quantitative results in Table 3 show that combining VAE and CLIP features is more advantageous. Additional ablation studies are given in supplementary materials.

5. Conclusion

We propose *Phantom*, a method for subject-consistent video generation that achieves cross-modal alignment through text-image-video triplet learning. By redesigning the joint text-image injection mechanism and leveraging dynamic feature integration, *Phantom* demonstrates competitive performance in unified single/multi-subject generation and facial ID preservation tasks, outperforming commercial solutions in quantitative evaluations.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai 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. 4, 13
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 3
- [3] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4
- [4] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. *arXiv preprint arXiv:2501.06187*, 2025. 1, 2, 3, 4, 5
- [5] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. *arXiv preprint arXiv:2412.07774*, 2024. 2, 4, 5
- [6] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, and Zhendong Mao. Dreamidentity: Enhanced editability for efficient face-identity preserved image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1281–1289, 2024. 4
- [7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 4, 7
- [8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos 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. 8, 13
- [9] dreamina. Generate image and video capabilities. <https://dreamina.captcut.com/>, 2024. 12
- [10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3, 5
- [11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 6
- [12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3
- [13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3
- [14] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *arXiv preprint arXiv:2404.16022*, 2024. 4, 12
- [15] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024. 2, 4
- [16] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 2, 3, 4
- [17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1
- [19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3
- [20] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 2, 3
- [21] Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom: Narrowing real text word for real-time open-domain text-to-image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7476–7485, 2024. 2, 12
- [22] Yuge Huang, Yuhang Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 8
- [23] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025. 1, 2, 3, 4, 5
- [24] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and ver-

- satite benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 7, 8
- [25] Tanuj Jain, Christopher Lennan, Zubin John, and Dat Tran. Imagededup. <https://github.com/idealo/imagededup>, 2019. 13
- [26] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 1
- [27] keling. Image to video elements feature. <https://klingai.com/image-to-video/multi-id/new/>, 2024. 3, 4, 7, 8, 14
- [28] 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, 3
- [29] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 2
- [30] Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, et al. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. *arXiv preprint arXiv:2502.07802*, 2025. 2, 3, 4
- [31] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025. 3, 5, 6
- [32] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 6
- [33] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 6
- [34] Pengqi Lu. Qwen2vl-flux: Unifying image and text guidance for controllable image generation, 2024. 4
- [35] MiniMax. Hailuo s2v-01. <https://www.minimaxi.com/en/news/s2v-01-release/>, 2024. 1, 3, 4, 7, 8
- [36] OpenAI. Sora. <https://openai.com/>, 2023. Accessed: February 10, 2024. 1
- [37] OpenAI. Chatgpt (gpt-4 version). <https://chat.openai.com/>, 2024. 1, 4, 13
- [38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafranec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 7
- [39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1, 3
- [40] Pika. Pikascenes. <https://pika.art/ingredients/>, 2024. 3, 4, 7, 8, 14
- [41] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1, 3, 4
- [42] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3
- [45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 3
- [46] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2024. 3
- [47] Shihao Shao and Qinghua Cui. 1st place solution in google universal images embedding. *arXiv preprint arXiv:2210.08473*, 2022. 5
- [48] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3
- [49] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3
- [50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 4
- [51] PySceneDetect Development Team. Pyscenedetect: An open-source video scene detection tool. <https://www.scenedetect.com/>, 2024. 4
- [52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

- Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
- [54] Vidu. Reference to video. <https://www.vidu.com/>, 2024. 3, 4, 7, 8, 14
- [55] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. *arXiv preprint arXiv:2501.01320*, 2025. 3, 5, 6
- [56] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 4
- [57] Yi Wang, Kunchang Li, Yizhuo Li, Yanan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 7
- [58] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 2, 4, 5
- [59] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1, 5
- [60] 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, 3, 5
- [61] Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapt: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 4
- [62] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*, 2024. 2, 3, 4, 7, 8
- [63] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3, 5
- [64] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 1
- [65] Wentao Zhu, Yufang Huang, Xiufeng Xie, Wenxian Liu, Jincan Deng, Debing Zhang, Zhangyang Wang, and Ji Liu. Autoshot: A short video dataset and state-of-the-art shot boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2238–2247, 2023. 4