



Phantom-Data: Towards a General Subject-Consistent Video Generation Dataset

Zhuowei Chen * Bingchuan Li *† Tianxiang Ma * Lijie Liu * Mingcong Liu
Yi Zhang Gen Li Xinghui Li Siyu Zhou Qian He Xinglong Wu

Intelligent Creation Lab, ByteDance

*Equal contribution, †Project lead

Abstract

Subject-to-video generation has witnessed substantial progress in recent years. However, existing models still face significant challenges in faithfully following textual instructions. This limitation, commonly known as the copy-paste problem, arises from the widely used in-pair training paradigm. This approach inherently entangles subject identity with background and contextual attributes by sampling reference images from the same scene as the target video. To address this issue, we introduce **Phantom-Data, the first general-purpose cross-pair subject-to-video consistency dataset**, containing approximately one million identity-consistent pairs across diverse categories. Our dataset is constructed via a three-stage pipeline: (1) a general and input-aligned subject detection module, (2) large-scale cross-context subject retrieval from more than 53 million videos and 3 billion images, and (3) prior-guided identity verification to ensure visual consistency under contextual variation. Comprehensive experiments show that training with Phantom-Data significantly improves prompt alignment and visual quality while preserving identity consistency on par with in-pair baselines.

Date: June 21, 2025

Correspondence: Zhuowei Chen, Bingchuan Li at {chenzhuowei.ustc, libingchuan}@bytedance.com

1 Introduction

In recent years, text-to-video generation models, exemplified by Sora [4], have made significant progress [23, 35, 37, 39, 48]. However, due to the limited controllability inherent in textual instructions, achieving fine-grained control over video generation remains a key challenge for practical applications. Among recent advances [11, 16, 20, 22, 46, 49], increasing attention has been paid to enforcing subject identity consistency in text-to-video generation. The subject-consistent video generation task (S2V) [6, 9, 10, 18, 27] aims to generate videos that not only follow the given text prompt but also faithfully preserve the identity of reference subjects, such as people, animals, products, or scenes. This capability has great potential in applications such as personalized advertising [5] and AI-driven filmmaking [44].

Despite encouraging progress in visual consistency, existing S2V approaches still suffer from limited text-following ability and suboptimal video quality, a phenomenon often referred to as the *copy-paste* problem. As shown in Fig. 2, the generated video directly replicates the reference subject from one of its frames, leading to



(a) Comparison on single-reference training data

(b) Multi-reference training data

Figure 1 Overview of training samples. (a) Single-reference setting: existing methods typically extract the reference image from the target video itself. In contrast, our approach uses reference images captured in distinct contexts. (b) Our dataset also includes multi-reference samples, presenting each subject in varied contextual settings.

the omission of the "boxing ring" background described in the prompt. This issue stems from the *in-pair* training paradigm, where the reference subject is sampled from the same target video, as illustrated in Fig.1(a). Consequently [15, 28], the model tends to preserve not only subject identity but also irrelevant contextual details. However, in real-world scenarios, such entangled features may contradict the actions or semantics described in the text prompt, causing the generated videos that either deviate from the prompt or exhibit noticeable artifacts.

To address the above issue, prior works [6, 18, 20, 21, 24] have explored various data normalization and augmentation strategies, such as background removal, color jittering, and geometric transformations. However, these methods struggle to unravel complex contextual factors, such as viewpoint and motion, due to limited variation. More recent approaches introduce *cross-pair* data, where identity-consistent reference and target frames are sampled from different sources. This setting encourages the model to focus on identity preservation while reducing overfitting to irrelevant visual contexts [33, 50]. However, existing cross-pair datasets are primarily limited to facial domains, making them difficult to generalize to general subject scenarios. Overall, current training datasets either provide insufficient reference variation or lack domain diversity, limiting the effectiveness of easing copy-paste problem.

In this work, we introduce *Phantom-Data*, a subject-to-video dataset specifically constructed to mitigate the prevalent copy-paste problem in the general scenarios. It is built around three core design principles for the reference subject: 1) **General and input aligned subjects**: Reference images should span a wide range of commonly encountered subject types and reflect the distribution of real-world user inputs. 2) **Different contexts**: Reference subjects appear in varied conditions—such as different backgrounds, viewpoints, or poses—relative to their counterparts in the target video. This encourages the model to generalize identity preservation under distribution shifts and reduces reliance on spurious identity-irrelevant correlations. 3) **Consistent identity**: Despite contextual variation, the reference subject must remain visually consistent with the target video subject in terms of shape, structure, and texture.

To fulfill these principles, we design a three-stage pipeline: Firstly, we perform **S2V Detection** by leveraging a vision-language model to conduct open-set object detection and identify candidate subjects of appropriate size. A second-stage filtering step further refines the results by retaining only subjects that are both semantically relevant and visually compact. Then, we conduct **Contextually Diverse Retrieval** by constructing a large-scale

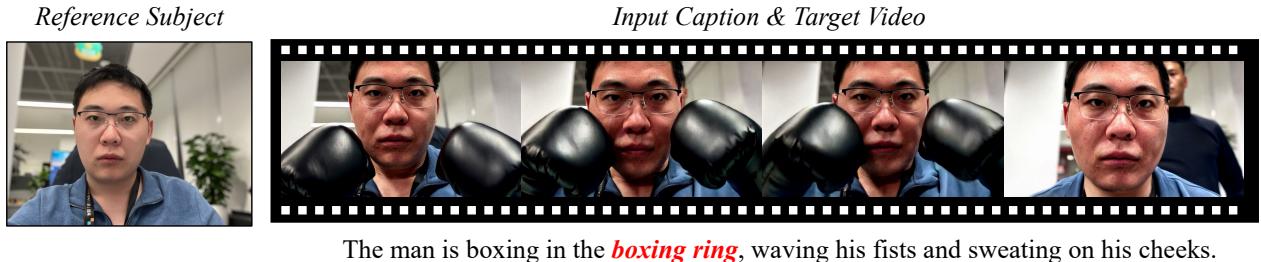


Figure 2 Illustration of the copy-paste problem. The shown result is generated by a SOTA video generation model (Kling [1]).

subject database comprising over 53 million video segments and 3 billion image samples, increasing the likelihood of retrieving the same identity under diverse backgrounds, poses, and viewpoints. Finally, we apply **Prior-Guided Identity Verification** to ensure identity consistency. For living beings (e.g., humans, animals), we mine temporal structures from long videos to construct cross-context pairs. For static objects (e.g., products), we perform category-specific retrieval. A final VLM-based pairwise check verifies that each selected pair maintains both identity consistency and contextual diversity. Through this pipeline, we construct a large-scale, high-quality cross-pair consistency dataset comprising approximately *1 million identity-consistent pairs* with over *30,000 multi-subject scenes*, offering a strong foundation for modeling general subject-to-video tasks. Representative samples are shown in Fig. 1.

To validate the effectiveness of our dataset, we conduct comprehensive experiments using open-source video generation models. The results demonstrate that, compared to prior data construction methods, our cross-pair approach substantially improves key metrics such as text alignment and visual quality, while maintaining identity consistency on par with in-pair baselines. Furthermore, we perform ablation studies to highlight the importance of a large-scale and diverse cross-pair dataset, showing that both data volume and scene diversity play a critical role in enhancing generation performance. We also validate the effectiveness of our data pipeline in preserving high identity consistency while ensuring sufficient contextual diversity.

Our main contributions can be summarized as follows:

- We introduce Phantom-Data, the first general-purpose cross-pair video consistency dataset, comprising approximately 1 million high-quality, identity-consistent pairs that span a wide range of subject categories and visual contexts.
- We present a structured data-construction pipeline purpose-built for subject-consistent video generation. It unifies a subject-centric detection module optimized for the S2V task, large-scale cross-context retrieval, and prior-guided identity verification, thereby securing strict identity fidelity while introducing rich contextual diversity.
- We conduct extensive experiments to validate the effectiveness of our dataset, demonstrating consistent improvements in text alignment, visual quality, and generalization over existing in-pair baselines.

Method	General Objects	Input-aligned Objects	Diverse Context	Publicly Available
MovieGen[33]	✗	✓	✓	✗
Video Alchemist[6]	✓	✗	✗	✗
ConceptMaster[18]	✓	✗	✗	✗
Ours	✓	✓	✓	✓

Table 1 Comparison between Phantom-data and datasets used in prior work.

2 Related Work

Text-to-Video Generation. Early diffusion-based video generators [3, 12, 42] were limited to producing short clips with constrained spatial and temporal resolution. However, the field has rapidly progressed with the introduction of large-scale latent diffusion models and transformer-based architectures. Notably, Sora [4] is capable of generating minute-long, high-fidelity videos, while contemporaneous systems such as Seaweed [37], Hunyuan-Video [23], CogVideo-X [48], MAGI [35], and others [39] have further advanced frame rate, resolution, scene complexity, realism, and motion smoothness. Despite their impressive visual quality, these generic text-conditioned models provide only coarse control: textual prompts alone cannot fully specify scene layout, subject appearance, or viewpoint, motivating research into finer control signals.

Subject-Consistent Video Generation. The task of subject-consistent video generation (S2V) [6, 9, 10, 18, 27] focuses on generating videos that not only align with the given text prompt but also preserve the visual identity of a reference subject, such as a person, animal, product, or scene. From a modeling perspective, one common strategy [6, 17, 18, 33] is cross-attention-based fusion, where visual features extracted from pretrained encoders[31, 34, 45] or VLMs, are injected into the generative backbone through dedicated attention layers. An alternative approach is noise-space conditioning, where identity features obtained from a VAE encoder are directly concatenated with the noise input of the diffusion model, without modifying the underlying architecture. This lightweight design enables nearly lossless injection of identity information, as seen in DIT-style models such as Phantom [27] and VACE [20]. Recent systems like SkyReels-A2 [10] explore combining both strategies, incorporating cross-attention guidance and noise-level conditioning within a unified framework.

Training Data in Subject-to-Video Generation. Training data plays a crucial role in subject-consistent video generation, as it directly influences a model’s ability to generate faithful and controllable results. Most existing approaches rely on *in-pair* supervision, where the reference and target frames are sampled from the same video clip. While this setup guarantees identity alignment, it often leads to the undesirable *copy-paste* effect—where the model reproduces not only the subject but also the background and pose of the reference frame, limiting its capacity to follow the input prompt. To mitigate this issue, several works [6, 18, 20, 21, 24] adopt data normalization and augmentation strategies, such as background removal, color jittering, and geometric transformations. However, these techniques, combined with the limited diversity inherent in in-pair training, are often insufficient to address complex contextual variations such as motion, viewpoint, and scene layout. Recent efforts have turned to *cross-pair* training, where identity-consistent reference and target frames are sampled from different videos. This setting encourages the model to concentrate on subject identity while reducing overfitting to specific visual contexts [33, 50]. Nevertheless, current cross-pair datasets are mostly restricted to narrow domains like human faces, limiting their generalizability to broader subject categories such as animals, products, or stylized characters. In summary, although cross-pair supervision offers a promising direction for addressing the copy-paste issue, the absence of high-quality, diverse, and identity-consistent training data across general domains remains a significant bottleneck for advancing S2V models. To bridge this gap, we introduce Phantom-data, a large-scale cross-pair dataset designed to support subject-consistent video generation across a wide range of real-world categories.

3 Phantom Data

We provide a detailed analysis of **Phantom-Data**, focusing on its statistical properties and a comparison with existing datasets for subject-consistent video generation.

3.1 Statistical Analysis

We analyze the dataset at both the video and subject levels.

Video-level properties. As shown in Fig. 3(a–c), our dataset spans a wide range of video durations, resolutions, and motion patterns. Around 50% of videos are 5–10 seconds long, and the majority are in 720p resolution. Motion levels also vary considerably, covering both relatively static and highly dynamic scenes.

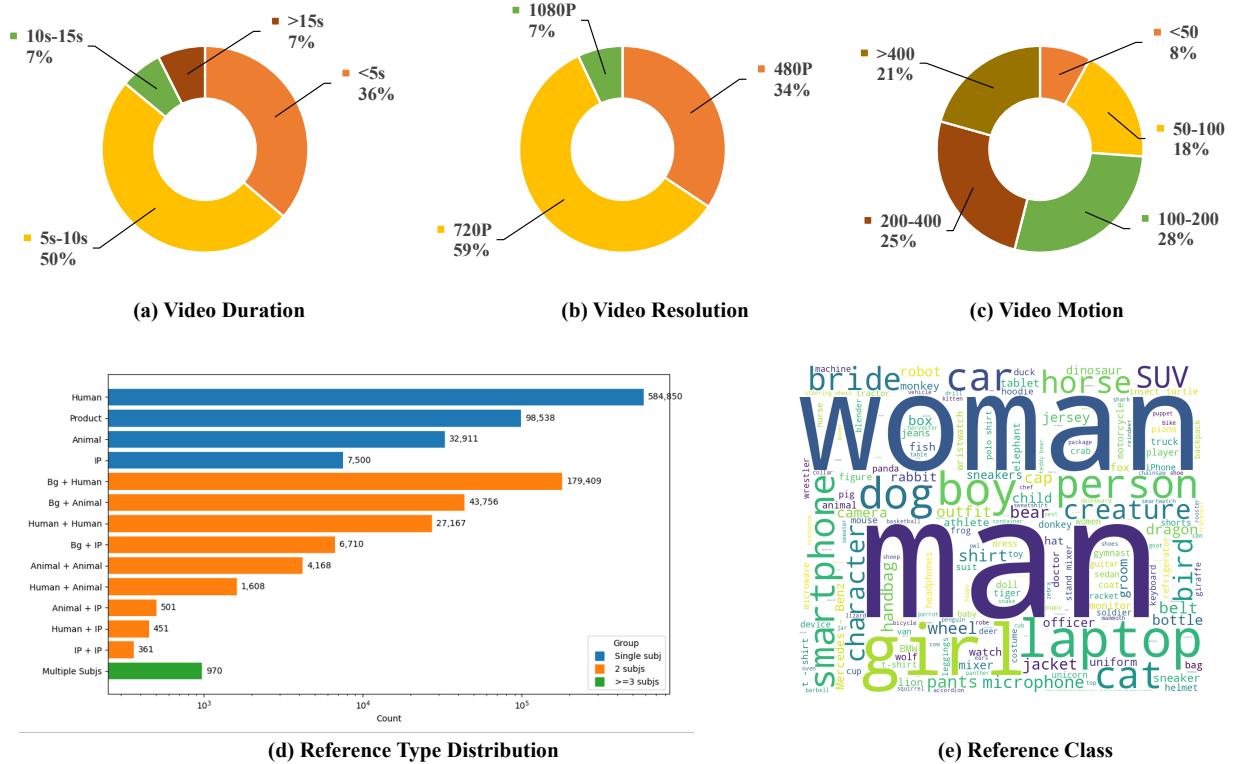


Figure 3 The statistical analysis of Phantom-Data.

Subject composition. Fig. 3(d) illustrates the distribution of subject types and their combinations. While the majority of samples (approximately 720,000) contain a single subject—such as a human, product, or animal—a substantial portion (around 280,000) involve two or more co-occurring entities, supporting multi-subject consistency modeling.

Reference diversity. As shown in Fig. 3(e), the dataset spans a broad semantic space of subject categories. Common reference entities include humans (e.g., woman, man, girl), animals (e.g., dog, bird), and man-made objects (e.g., smartphone, car, laptop), highlighting the dataset’s suitability for general-purpose subject-to-video modeling across varied domains.

3.2 Comparison with Prior Datasets

As summarized in Table 1, existing datasets for subject-consistent video generation either lack general object coverage, rely heavily on input-aligned references from the same video, or are limited in contextual diversity. In contrast, Phantom-Data offers a more comprehensive setting: it supports general object categories beyond faces, encourages cross-context modeling by sampling subject-reference pairs from diverse scenes, and is publicly available for research. This makes it the first open-access dataset to jointly support identity consistency and context diversity in a general-purpose, cross-pair setup.

4 Data Pipeline

4.1 Video Data Source

The Phantom-Data video dataset consists of clips collected from public sources such as Koala-36M [41], as well as proprietary internal repositories. Each video undergoes a rigorous quality control pipeline, including black border detection, motion analysis, and other filtering steps. Subsequently, long videos are segmented into short clips at the second level using scene segmentation. Each resulting clip is then annotated with a

corresponding video caption. The total number of videos is approximately 53 million.

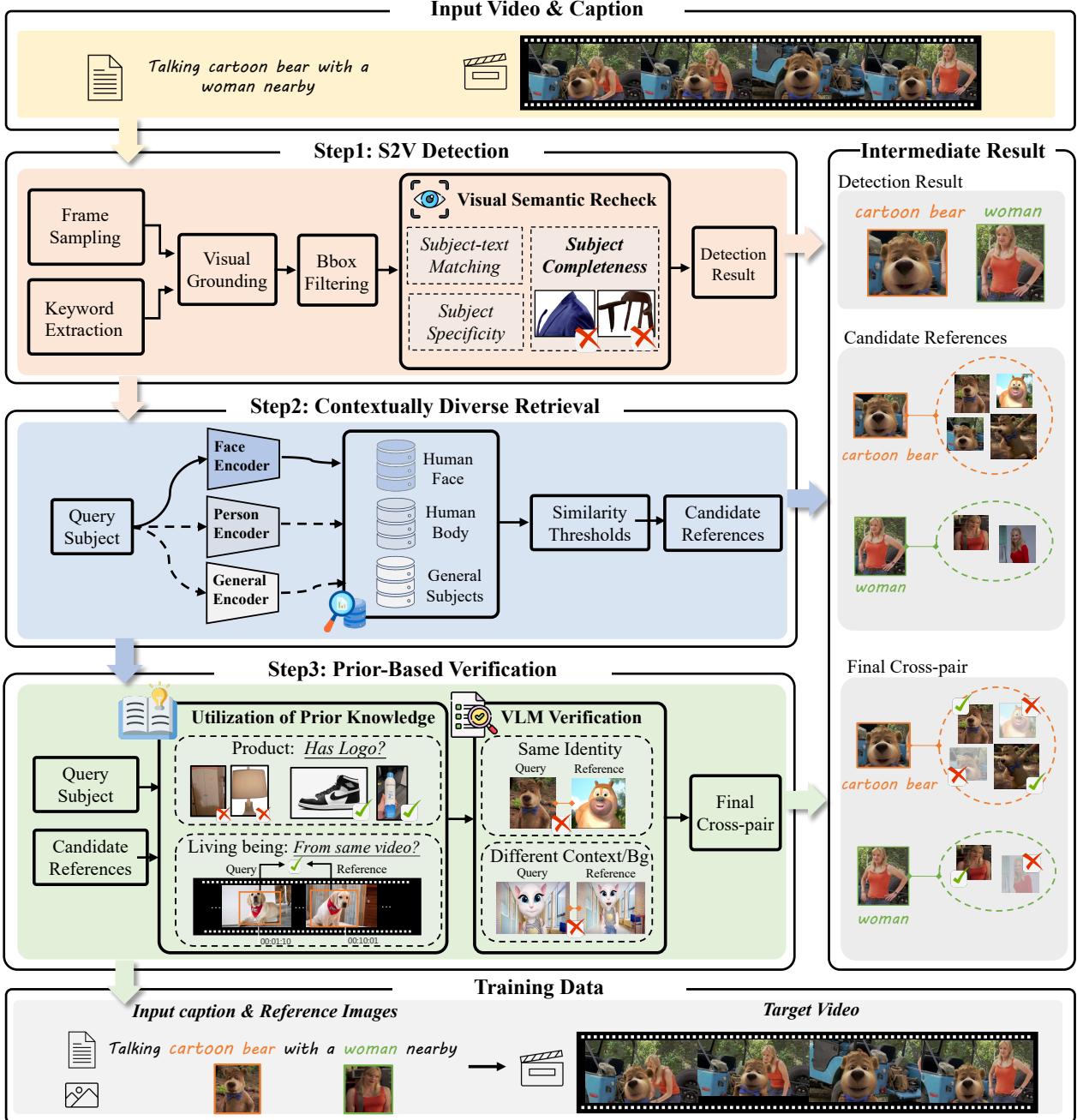


Figure 4 The overview of the data pipeline for constructing cross-pair training samples.

4.2 Data Pipeline

Given an input video and its associated caption, we focus on constructing a high-quality *cross-pair* dataset, where the same subject appears across different visual contexts while maintaining identity consistency. To this end, we design a structured data pipeline consisting of three key stages. As shown in Fig.4, firstly, we perform *S2V Detection* to identify high-quality subject instances from videos. Then, we propose a *Contextually Diverse Retrieval* module to recall candidate images that are likely to correspond to the detected subjects across

varying scenes. Finally, we apply *Prior-based Identity Verification* to filter the retrieved candidates, ensuring that only those sharing the same identity across different contexts are retained.

4.2.1 S2V Detection

This stage aims to identify diverse and qualified subjects from each video clip as candidates for cross-scene pairing. It consists of five major steps:

1. **Frame Sampling.** To reduce computation, we sample three frames at $t = 0.05, 0.5$, and 0.95 of each clip, following [6], ensuring temporal diversity while avoiding full-frame processing.
2. **Keyword Extraction.** We use Qwen2.5 [47] to extract key noun phrases (e.g., people, animals, products) from captions, serving as subject candidates for grounding.
3. **Visual Grounding.** Qwen2.5-VL [2] aligns each phrase to regions in the sampled frames. Ambiguous matches mapping to multiple regions are removed to reduce noise.
4. **Bbox Filtering.** We retain boxes covering between 4% and 90% of the image and at least 128×128 in size. Overlapping boxes ($\text{IoU} > 0.8$) are suppressed for clarity.
5. **Visual-Semantic Recheck.** To further ensure the quality of the grounded subjects, we employ another vision-language model, InternVL2.5 7B [7], to validate each detection against the following criteria: 1) Completeness: We observe that visual grounding often produces bounding boxes around partial or cropped objects, as a result of the underlying detection model's exhaustive labeling strategy. However, for S2V task, users typically provide complete reference subjects, making such incomplete detections unsuitable. We therefore filter out any region that fails to cover the full extent of the object. 2) Specificity: The subject must be visually distinct and identifiable. Vague or generic objects, such as trees, rocks, or background clutter, are excluded. 3) Subject-text Matching: The grounded region must be semantically consistent with the associated phrase. To improve alignment precision, we employ a separate instance of InternVL2.5 to reevaluate the consistency between the textual description and the detected subject.

As a result of this pipeline, we obtain a high-quality set of subject instances, each paired with a corresponding descriptive phrase. Since a subject may appear in multiple frames across the video, we select only one representative instance for visualization in the Intermediate Result section of Fig. 4.

4.2.2 Contextually Diverse Retrieval

Given the subject instances detected in the previous stage, we aim to find candidate reference images of the same subject appearing in different visual contexts. To achieve this, we construct a large-scale retrieval bank and use the detected subjects to perform identity-aware querying.

Large-Scale Retrieval Bank Construction. The retrieval bank comprises two essential components: diverse subject image sources to increase contextual variability, and feature representations tailored for identity-preserving retrieval.

Subject Source. We begin by registering every detected subject instance from the training videos into a retrieval bank. To further broaden candidate diversity, we augment this bank with an extra 3 billion images from the LAION dataset [36] beyond the original video corpus. These external images inject greater variation in scene, pose, and appearance, delivering broader contextual coverage during retrieval, an advantage that is particularly valuable for product-centric scenarios with substantial intra-instance variation.

Subject Representation. To support reliable cross-context identity matching, we employ expert-designed encoders to extract identity-preserving and context-invariant embeddings tailored to different subject categories. These embeddings are used for both indexing the retrieval bank and querying.

For facial representation, we adopt the widely used ArcFace encoder [8] to extract robust and discriminative identity embeddings:

$$V_{\text{face}} = E_{\text{arcface}}(I_{\text{face}}). \quad (1)$$

For general objects, inspired by ObjectMate [43], we employ a CLIP-based model fine-tuned on a consistency-focused image dataset [38] to extract identity-preserving embeddings:

$$V_{\text{subj}} = E_{\text{IR}}(I). \quad (2)$$

For human subjects, which are central to many downstream applications, we combine both facial and clothing features. Each individual is represented by concatenating the general appearance embedding with the corresponding facial embedding:

$$V_{\text{person}} = [E_{\text{IR}}(I), E_{\text{arcface}}(I_{\text{face}})]. \quad (3)$$

Query-Based Retrieval. To ensure the retrieved candidates are visually distinct from the query image yet share the same identity, we apply both upper and lower bounds on similarity. Specifically, we discard overly similar results (potential duplicates) by enforcing an upper similarity threshold, and exclude unrelated identities by applying a lower threshold.

4.2.3 Prior-Based Identity Verification

However, due to the large scale of the retrieval corpus, false positives frequently occur even within seemingly reasonable similarity ranges. To address this issue, we adopt a two-stage filtering strategy based on prior knowledge and VLM Verification.

Utilization of Prior Knowledge. We apply category-specific filtering strategies to improve cross-pair reliability: 1) *Non-living subjects* (e.g., products): These typically exhibit high intra-class variability, making identity verification more challenging. To improve precision, we retain only product instances that feature complete and recognizable brand logos (e.g., Nike, Audi), which remain visible across different scenes. 2) *Living entities* (e.g., humans, animals): For these subjects, we restrict retrieved candidates to those from different clips within the same long-form video. This constraint ensures natural variation in scene and pose while maintaining consistent identity.

VLM-Based Consistency Verification. To further ensure both identity consistency and contextual diversity, we apply a VLM-based verification procedure: 1) *identity consistency*: For non-living objects, we enforce strict similarity in visual details such as color, packaging, and textual elements, while allowing for background variation. For living subjects, especially humans, we verify facial identity consistency and, in the case of full-body samples, also ensure clothing alignment. 2) *Contextual diversity*. We keep only those cross-pair samples that exhibit substantial variation in background and scene context, thereby alleviating copy-paste artifacts during model training.

5 Experiments

5.1 Implementation

Model Architecture. We validate the effectiveness of our proposed data using the Phantom-wan [27] model. Built on the Wan2.1 [40] foundation, Phantom-wan is a leading open-source framework for subject-consistent video generation.

Training and inference. We train a 1.3 billion-parameter Phantom-wan model using Rectified Flow (RF) [25, 28] as the training objective. The training is performed on 64 A100 GPUs for 30k iterations with 480p resolution data, which yields stable performance. During inference, we apply Euler sampling with 50 steps and use classifier-free guidance [14] to decouple image and text conditions. All experiments follow the same training and inference settings to ensure fair comparisons.

Evaluation. We construct a test suite of 100 cases from diverse scenarios, covering humans, animals, products, environments, and clothing. These cases include both single- and multi-subject settings, paired with manually written text prompts that reflect natural user input.

We evaluate model performance across three dimensions: video quality, text-video consistency, and subject-video consistency. Subject-video consistency is evaluated using CLIP [13], DINO [30], and GPT-4o scores,



Figure 5 Qualitative comparisons across different training strategies.

following recent evaluation protocols inspired by [32]. Text-video consistency is measured via Reward-TA [26], both of which assess the semantic alignment between generated video content and the text prompt. Video quality is assessed using VBench [19] , which provide fine-grained evaluation along several aspects, including Temporal (temporal flickering and stability), Motion (smoothness of subject motion), IQ (overall imaging quality), BG (background consistency across frames), and Subj (temporal consistency of the generated subject).

5.2 Main Results

We evaluate our method against three representative baselines: (1) In-pair training, which samples the reference subject from the same video; (2) In-pair with copy-augmentation, which introduces spatial and appearance augmentations to reduce overfitting as in [6]; and (3) Face-based cross-pair, which utilizes face-level identity matching across videos. We report both quantitative and qualitative comparisons.

Quantitative results demonstrate that our cross-pair training paradigm achieves state-of-the-art performance in terms of text-video alignment and overall video quality, as measured by reward-based evaluation metrics. Furthermore, our method delivers competitive subject consistency, rivaling in-pair baselines despite the increased scene diversity. In contrast, In-pair settings suffer from poor text-following ability due to overfitting on narrow visual contexts. The Face cross-pair method performs slightly better on prompt following but is limited by the narrow domain of its face-centric training data, resulting in weaker identity preservation across diverse subjects.

Methods	Subject Consistency		Prompt Following		Video Quality				
	DINO \uparrow	GPT-4o \uparrow	Reward-TA \uparrow	Temporal \uparrow	Motion \uparrow	IQ \uparrow	BG \uparrow	Subj \uparrow	
In-pair	0.478	2.481	2.074	0.971	0.985	0.725	0.937	0.933	
In-pair + Data Aug	<u>0.473</u>	<u>2.792</u>	2.427	0.961	0.979	<u>0.730</u>	0.932	0.922	
Face Cross-pair	0.354	2.378	<u>3.022</u>	0.983	0.989	0.723	<u>0.937</u>	0.935	
Ours	0.416	3.041	3.827	0.975	0.986	0.739	0.948	0.944	

Table 2 Main results comparing prompt following, subject consistency, and video quality across different training paradigms. Bold denotes the best performance per column. The underline indicates the second-highest scores.

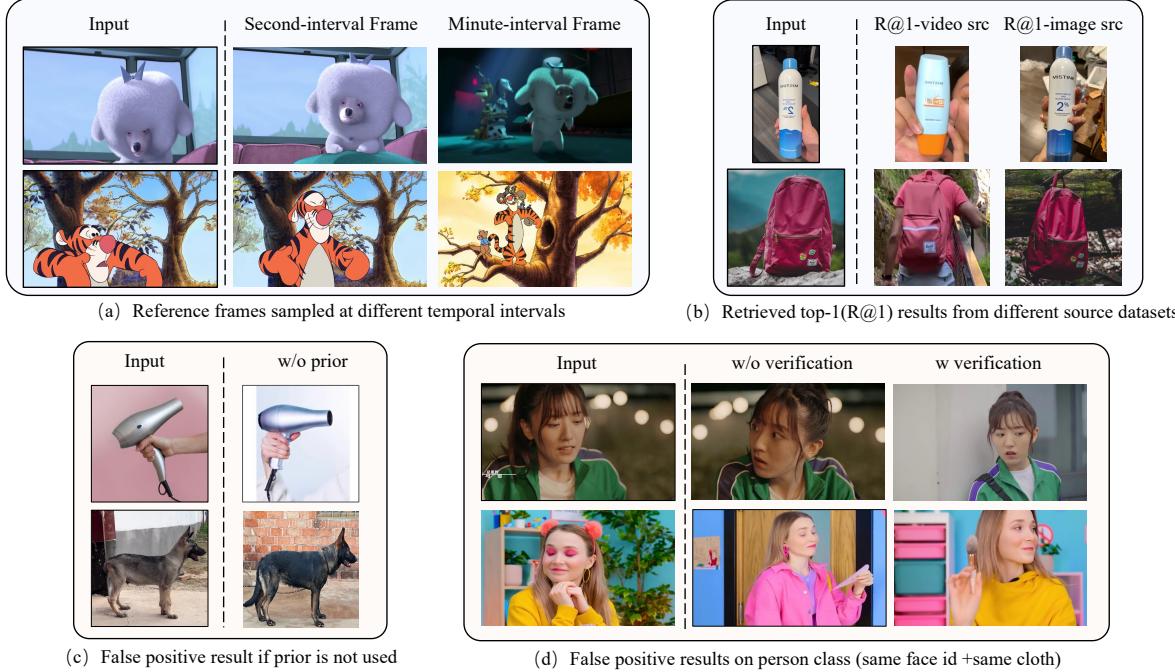


Figure 6 Ablation study on Contextually Diverse Retrieval and Prior-Based Identity Verification. (a) Reference frames from different timestamps show that longer videos offer more diverse contexts. (b) Retrieval from large-scale image datasets improves recall and candidate diversity. (c) Without prior filtering, false positives may be included. (d) Verification removes mismatched or overly similar identities, ensuring high-quality pairs.

Qualitative comparisons, as shown in Fig.5, further support our conclusions. Across multiple prompts and subject categories, models trained with in-pair data consistently fail to follow textual instructions, often generating videos with obvious artifacts. In contrast, our cross-pair trained model successfully aligns with the prompt across all cases, producing coherent and faithful subject-driven videos.

5.3 Ablation Studies

Subject Diversity. As shown in Table 3, enriching the training set with diverse subject types—including humans, animals, products, and multi-subject scenes—consistently improves subject consistency and prompt following, compared to the face-only baseline.

Data Scale. Table 4 illustrates the effect of data scale. Increasing the training set from 100 k to 1 million samples leads to further improvements across all metrics, highlighting the importance of both diversity and scale in building a robust subject-to-video generation dataset.

Contextually Diverse Retrieval. To assess the impact of contextual diversity in reference selection, we compare different sampling and retrieval strategies: (1) Temporal Sampling. As illustrated in Fig. 6(a),

Table 3 Ablation study on subject diversity.

Methods	Subject Consistency		Prompt Following reward-TA \uparrow
	DINO \uparrow	GPT-4o \uparrow	
baseline (face only)	0.354	2.378	3.022
+ human	0.401 \pm 0.047	2.747 \pm 0.363	3.726 \pm 0.702
+ IP/animal	0.416 \pm 0.062	2.795 \pm 0.411	3.407 \pm 0.383
+ product	0.386 \pm 0.032	2.662 \pm 0.288	3.572 \pm 0.58
+ multi-subject	0.418 \pm 0.064	2.901 \pm 0.525	3.512 \pm 0.498

Table 4 Ablation study on data scale.

Methods	Subject Consistency		Prompt Following reward-TA \uparrow
	DINO \uparrow	GPT-4o \uparrow	
100k	0.408	3.090	3.796
1 M	0.416	3.175	3.827

reference frames sampled at longer temporal intervals (e.g., minute-level vs. second-level) provides richer visual diversity. (2) Multi-source Retrieval. Fig. 6(b) compares retrieval from video-only sources and from a combined image+video retrieval bank. Incorporating large-scale image datasets improves both recall and candidate diversity.

Prior-Based Identity Verification. We further evaluate the role of prior filtering and identity verification in ensuring training quality: (1) Prior Filtering. Without prior-based constraints, visually similar but semantically incorrect matches (false positives) are often included (see Fig. 6(c)). (2) Verification Module. As shown in Fig. 6(d), our identity verification module further refines the candidate set by removing both near-duplicates (overly similar samples) and mismatched identities (overly dissimilar ones).

6 Conclusion

We propose **Phantom-Data**, a large-scale, general-purpose cross-pair dataset to improve subject consistency and text alignment in text-to-video generation. By introducing a structured pipeline—combining open-vocabulary detection, diverse cross-context retrieval, and identity verification—we address the limitations of in-pair training and reduce the copy-paste problem. Experiments show that our dataset significantly boosts generation quality while maintaining strong identity consistency. Phantom-Data provides a solid foundation for future research in controllable subject-to-video generation task.

References

- [1] Kling. <https://app.klingai.com/cn/image-to-video/frame-mode/new>.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. [arXiv preprint arXiv:2502.13923](#), 2025.
- [3] 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.
- [4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>.
- [5] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. [arXiv preprint arXiv:2502.04896](#), 2025.
- [6] 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.
- [7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. [arXiv preprint arXiv:2412.05271](#), 2024.
- [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.
- [9] Yufan Deng, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Angtian Wang, Shanghai Yuan, Yiding Yang, Bo Liu, Haibin Huang, and Chongyang Ma. Cinema: Coherent multi-subject video generation via mllm-based guidance. [arXiv preprint arXiv:2503.10391](#), 2025.
- [10] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. [arXiv preprint arXiv:2504.02436](#), 2025.
- [11] Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol-camera: Precise video camera control with adjustable motion strength. 2025.
- [12] 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. In [12th International Conference on Learning Representations, ICLR 2024](#), 2024.
- [13] 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.
- [14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. [arXiv preprint arXiv:2207.12598](#), 2022.
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. [Advances in neural information processing systems](#), 33:6840–6851, 2020.
- [16] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 8153–8163, 2024.
- [17] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. [arXiv preprint arXiv:2505.04512](#), 2025.
- [18] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruima 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.
- [19] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models.

In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21807–21818, 2024.

- [20] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. [arXiv preprint arXiv:2503.07598](#), 2025.
- [21] Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention. [arXiv preprint arXiv:2503.19907](#), 2025.
- [22] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: a large language model for zero-shot video generation. In [Proceedings of the 41st International Conference on Machine Learning](#), pages 25105–25124, 2024.
- [23] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuancode: A systematic framework for large video generative models. [arXiv preprint arXiv:2412.03603](#), 2024.
- [24] Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kumpeng 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.
- [25] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. [arXiv preprint arXiv:2210.02747](#), 2022.
- [26] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. [arXiv preprint arXiv:2501.13918](#), 2025.
- [27] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. [arXiv preprint arXiv:2502.11079](#), 2025.
- [28] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In [The Eleventh International Conference on Learning Representations](#).
- [29] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. [arXiv preprint arXiv:2504.16915](#), 2025.
- [30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, 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.
- [31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. [Transactions on Machine Learning Research Journal](#), 2024.
- [32] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. [arXiv preprint arXiv:2406.16855](#), 2024.
- [33] 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.
- [34] 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.
- [35] Sand-AI. Magi-1: Autoregressive video generation at scale, 2025. URL https://static.magi.world/static/files/MAGI_1.pdf.
- [36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. [Advances in neural information processing systems](#), 35:25278–25294, 2022.

- [37] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. [arXiv preprint arXiv:2504.08685](https://arxiv.org/abs/2504.08685), 2025.
- [38] Shihao Shao and Qinghua Cui. 1st solution in google universal image embedding. <https://www.kaggle.com/datasets/louieshao/guieweights0732>, april 2023.
- [39] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. [arXiv preprint arXiv:2503.20314](https://arxiv.org/abs/2503.20314), 2025.
- [40] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. [arXiv preprint arXiv:2503.20314](https://arxiv.org/abs/2503.20314), 2025.
- [41] Qiheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. [arXiv preprint arXiv:2410.08260](https://arxiv.org/abs/2410.08260), 2024.
- [42] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. [arXiv preprint arXiv:2401.04468](https://arxiv.org/abs/2401.04468), 2024.
- [43] Daniel Winter, Asaf Shul, Matan Cohen, Dana Berman, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectmate: A recurrence prior for object insertion and subject-driven generation. [arXiv preprint arXiv:2412.08645](https://arxiv.org/abs/2412.08645), 2024.
- [44] Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. [arXiv preprint arXiv:2503.07314](https://arxiv.org/abs/2503.07314), 2025.
- [45] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In [The Twelfth International Conference on Learning Representations](#).
- [46] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 1481–1490, 2024.
- [47] 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](https://arxiv.org/abs/2412.15115), 2024.
- [48] 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](https://arxiv.org/abs/2408.06072), 2024.
- [49] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 8850–8860, 2024.
- [50] Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-id: Towards universal identity-preserving video synthesis. [arXiv preprint arXiv:2503.14151](https://arxiv.org/abs/2503.14151), 2025.

Appendix

A The Limitations of Synthetic Data

We try two SOTA models, GPT4o and DreamO [29] to generate consistent subjects on different context. The results in Fig.7 shows these models could still generated inconsistent subject. However our real cross-pair data construction pipeline could provide exactly same subjects on different context.

B User study

To evaluate the generated videos under different training data regimes, we conducted a user study comparing four settings: in-pair training, in-pair with data augmentation, face-level cross-pair training, and our proposed full-object cross-pair approach. We ask six participants, each of whom independently evaluated 50 video groups, containing four videos generated in the different training settings. For each group, participants were asked to select the best video in terms of overall visual quality, subject consistency, and alignment with textual prompts. As summarized in Table 5, our method was overwhelmingly preferred, receiving **76%** of the votes. In contrast, all other baselines received less than 12%, highlighting the effectiveness of our cross-pair training design in producing videos that are faithful to the intent of users.

In-pair	In-pair + Data Aug	Face Cross-pair	Ours
6%	11%	7%	76%

Table 5 User study on the best video selection based on overall visual quality, subject consistency, and text alignment across different training data settings.

C Broader Impact

Our work aims to enhance identity consistency and contextual diversity in subject-to-video generation, thereby improving the controllability and realism of AI-generated content. This technology holds promise for a wide range of applications, including personalized media creation, digital asset generation, and educational or entertainment content production.

Nonetheless, we recognize the potential social risks associated with realistic identity-preserving video synthesis. Such capabilities may be misused for malicious purposes, including the creation of deepfakes, impersonation, or the dissemination of misinformation. We therefore emphasize the importance of responsible research and deployment practices. In particular, we encourage the use of watermarking, provenance tracking, and informed consent mechanisms to ensure ethical and transparent use—especially in scenarios involving human likeness or identity-sensitive content.

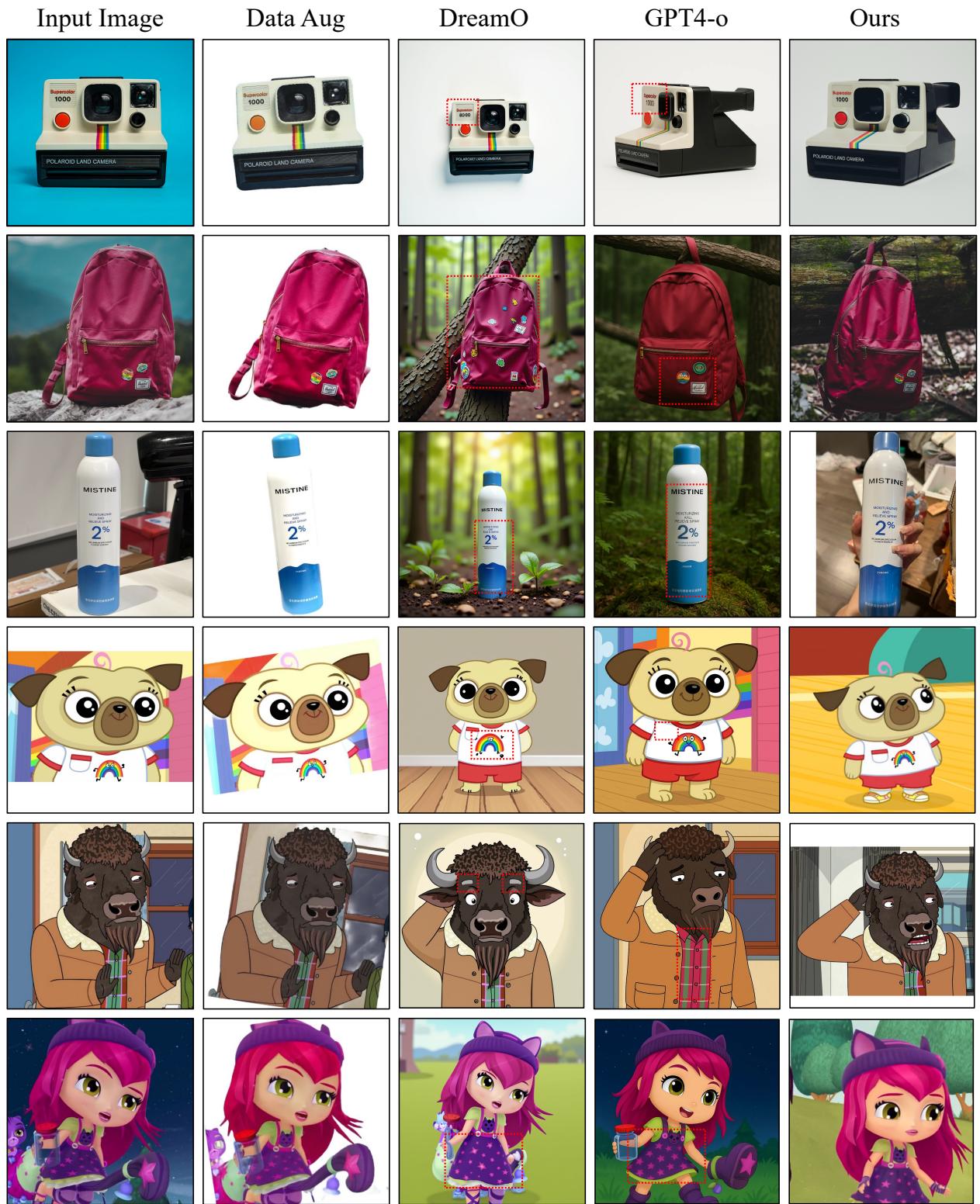


Figure 7 Comparison of reference-data construction strategies. Naïve data augmentation (Data Aug) offers only limited variation, while directly adopting state-of-the-art IP-consistent generators (e.g., DreamO, GPT-4o) can introduce appearance inconsistencies, highlighted in the figure. Best viewed in color with zoom.