# LibraGen: Playing a Balance Game in Subject-Driven Video Generation

Jiahao Zhu[1,2*], Shanshan Lao[2*], Lijie Liu[2*], Gen Li[2*], Tianhao Qi[2*], Wei Han[2*], Bingchuan Li[2*,‡], Fangfang Liu[2], Zhuowei Chen[2], Tianxiang Ma[2], Qian He[2], Yi Zhou[1†], and Xiaohua Xie[1,3,4†]

[1] Sun Yat-sen University, China
[2] ByteDance, China
[3] Pazhou Lab (Huangpu), China
[4] Guangdong Province Key Laboratory of Information Security Technology, China

**Abstract.** With the advancement of video generation foundation models (VGFMs), customized generation, particularly subject-to-video (S2V), has attracted growing attention. However, a key challenge lies in balancing the intrinsic priors of a VGFM, such as motion coherence, visual aesthetics, and prompt alignment, with its newly derived S2V capability. Existing methods often neglect this balance by enhancing one aspect at the expense of others. To address this, we propose LibraGen, a novel framework that views extending foundation models for S2V generation as a balance game between intrinsic VGFM strengths and S2V capability. Specifically, guided by the core philosophy of "Raising the Fulcrum, Tuning to Balance," we identify data quality as the fulcrum and advocate a quality-over-quantity approach. We construct a hybrid pipeline that combines automated and manual data filtering to improve overall data quality. To further harmonize the VGFM's native capabilities with its S2V extension, we introduce a Tune-to-Balance post-training paradigm. During supervised fine-tuning, both cross-pair and in-pair data are incorporated, and model merging is employed to achieve an effective trade-off. Subsequently, two tailored direct preference optimization (DPO) pipelines, namely Consis-DPO and Real-Fake DPO, are designed and merged to consolidate this balance. During inference, we introduce a time-dependent dynamic classifier-free guidance scheme to enable flexible and fine-grained control. Experimental results demonstrate that LibraGen outperforms both open-source and commercial S2V models using only thousand-scale training samples. Project Page: `https://github.com/Phantom-video/LibraGen`

**Keywords:** Video Generation Foundation Models · Subject-to-Video · Post Training

## 1 Introduction

Recently, numerous open-source [23, 36, 39, 45] and commercial [1, 9, 15, 24, 30] video generation foundation models (VGFMs) have emerged, accelerating intelli-

* Equal contribution. ‡ Project lead. †Corresponding authors.

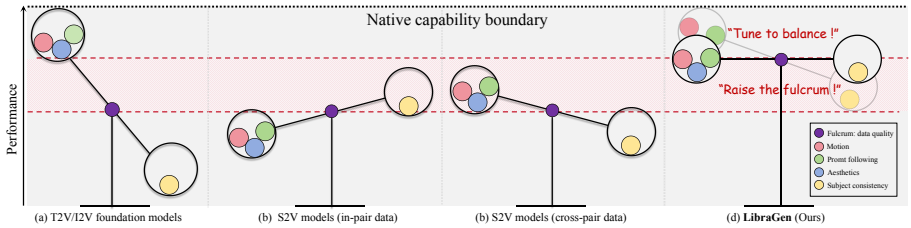


**Fig. 1: A balance game in S2V generation**. (a) T2V/I2V foundation models lack task-specific training data and thus exhibit poor S2V performance. (b) Previous S2V methods trained solely on in-pair data or (c) solely on cross-pair data often overlook the inherent balance trade-off. (d) LibraGen frames S2V generation as a balance game, achieving superior and well-balanced S2V performance.

gent video content creation. These models mainly target two tasks: text-to-video (T2V) and image-to-video (I2V). T2V models generate text-aligned videos, yet prompt-only control struggles to meet complex, multimodal, fine-grained user demands. I2V models turn static images into dynamic content, but the fixed initial frame constraint severely limits creative flexibility. To enable more customized control, commercial products (*e.g.*, Vidu [1], Kling [24], and Sora2 [31]) have been extended to the subject-to-video (S2V) task, which generates text-aligned, subject-consistent videos using textual prompts and reference images.

Extending a text-to-video (T2V) or image-to-video (I2V) foundation model to subject-to-video (S2V) generation represents a typical case of task-incremental learning, which inherently necessitates careful balancing between retaining prior task knowledge and acquiring new capabilities [35,37]. Consequently, the central challenge in S2V generation is to faithfully preserve the foundation model's intrinsic advantages, such as superior motion quality, appealing visual aesthetics, and strong prompt adherence, while simultaneously achieving robust subject-consistent video generation. As demonstrated in Fig. 1(a), off-the-shelf T2V/I2V foundation models exhibit markedly poor zero-shot performance on S2V tasks due to the complete absence of tailored training data. To overcome this limitation, a widely established training paradigm has emerged: fine-tuning these foundation models on carefully constructed `<prompt, video, reference>` triplets. Several studies [7,10,19,25,38,46] form such triplets by directly extracting the reference subject from frames within the target video clip itself (commonly referred to as in-pair data); however, this seemingly straightforward approach introduces severe reference content leakage during training, frequently resulting in undesirable copy-and-paste artifacts, which in turn cause the generated videos to suffer from degraded visual aesthetics, suppressed motion dynamics, and weakened prompt alignment, as illustrated in Fig. 1(b). More recent methods [4, 11, 27, 44] advocate the use of cross-pair data for fine-tuning, sourcing reference subjects from entirely separate video clips; while this strategy effectively eliminates the copy-and-paste issue and substantially improves prompt responsiveness, it significantly weakens the intrinsic correlation between the reference image and the

**Fig. 2:** We present LibraGen, a novel training paradigm that extends VGFMs to support both single-subject and multi-subject-driven video generation.

target video content, often leading to compromised subject consistency, as shown in Fig. 1(c). Furthermore, because most current S2V models are trained predominantly on large-scale, automatically curated (and thus noisy) datasets, ensuring consistently high human-aligned quality remains challenging, ultimately constraining overall S2V performance to suboptimal levels.

To address these issues, we propose **LibraGen**, a novel framework that extends T2V/I2V foundation models to support S2V generation by formulating the problem as a principled balance game. The framework is built upon a new training paradigm, termed "Raising the Fulcrum, Tuning to Balance," as shown in Figure 1(d), which comprises three key aspects:

**Data Curation**. Recent VGFMs exhibit strong T2V and I2V capabilities, greatly lowering the barrier to S2V learning and thus necessitating a paradigm shift from data quantity to data quality. We argue that data quality acts as a critical balancing fulcrum, and its careful refinement can significantly boost overall S2V performance. To this end, we pursue three key directions: (1) Data refinement—we distill a million-scale raw dataset into a thousand-scale, high-quality, human-aligned subset through a systematic automatic–manual hybrid pipeline encompassing video collection, precise subject extraction, and accurate captioning; (2) Data trait analysis—we strategically leverage the complementary strengths of in-pair and cross-pair data by thoroughly considering their inherent

characteristics; (3) Data labeling—we introduce a hierarchical tagging system that enables fine-grained control over the training data distribution.

**Model Training**. We propose a Tune-to-Balance post-training paradigm that harmonizes the foundation model's inherent strengths with its derived S2V capability. During supervised fine-tuning (SFT), models trained on in-pair and cross-pair data exhibit complementary strengths and weaknesses in subject consistency and foundation model capabilities. We adopt a weighted model merging strategy to achieve a trade-off. We further design two direct preference optimization (DPO) pipelines, termed **Consis-DPO** and **Real-Fake DPO**, and merge them to consolidate this balance, where the former enhances subject consistency and the latter restores foundation model performance.

**Model Inference**. We observe that early denoising steps primarily capture text-guided structural layouts and motion patterns, while later steps increasingly incorporate fine-grained visual details from the reference. Based on this, we propose a time-dependent dynamic classifier-free guidance (CFG) strategy that adaptively adjusts the relative strengths of text and reference conditioning throughout the denoising process, enabling precise and flexible control over the balance between prompt adherence and reference fidelity.

The contributions of this paper are summarized as follows:

- **Concept**. This paper identifies the fundamental challenge in S2V generation as an intrinsic trade-off between preserving native VGFM capabilities and ensuring robust subject-consistent video generation.
- **Technology**. We propose LibraGen, a novel training paradigm that regards S2V generation as a balance game. It introduces a quality-over-quantity data curation strategy and a Tune-to-Balance post-training framework.
- **Performance**. As shown in Figure 2, when trained on only thousand-scale data, LibraGen delivers impressive single- and multi-subject-driven video generation, outperforming both commercial and open-source S2V models.

## 2 Related Work

### 2.1 Video Generation Foundation Models

Recent advances in diffusion models have fueled rapid progress in video generation. Early video generation models [3, 17, 33, 43] typically adopted 2D U-Net backbones, integrating temporal modeling components (*e.g.*, attention mechanisms or convolutional layers) prior to fine-tuning on specialized video datasets. However, their generative performance is inherently constrained by architectural design and model scale. With the continuous expansion of computational resources, scaling laws [21, 40] have been empirically validated in the video generation domain, spurring the development of large-scale video generation models based on the DiT architecture [12, 32]; the videos synthesized by these models demonstrate remarkably superior visual fidelity. On the open-source front, representative models include Wan [36], HunyuanVideo [23], and CogVideoX [39], whereas closed-source counterparts, such as Veo 3.1 [9], Kling 2.5 [24], Hailuo

2.3 [30], Vidu [1], and Seedance 1.0/1.5 [5, 15], have attained commercial-grade performance levels. Fundamentally, foundation models for video generation primarily focus on T2V and I2V tasks, with their design objectives anchored in three core pillars: motion coherence, visual quality, and prompt alignment. Leveraging their scalability, these models can be adapted to a wide range of downstream tasks via customized post-training strategies.

### 2.2 Subject-to-Video Generation

The objective of S2V generation is to generate subject-consistent and prompt-aligned videos from given reference images and textual instructions. This can be achieved by fine-tuning state-of-the-art video generation foundation models on `<prompt, video, reference>` triplets. Some studies [7, 10, 19, 25, 38, 46] construct such triplets (*a.k.a.* in-pair triplets) by extracting reference subjects from frames within target video clips. However, this strategy often leads to copy-and-paste artifacts. Although some methods [7] alleviate this issue through data augmentation or prompt engineering, they do not fundamentally address the problem of unnatural integration of reference subjects, such as lighting inconsistencies between the subjects and the background in generated videos. To overcome these issues, subsequent studies [4,11,27,44] propose fine-tuning with cross-pair triplets, where reference subjects are extracted from different video clips. By weakening the direct correlation between target videos and paired reference images, this strategy effectively mitigates copy-and-paste artifacts.

## 3 Model Design

### 3.1 Lightweight Subject Injection

We build LibraGen upon a Multi-Modal Diffusion Transformer (MM-DiT) [15], which consists of low- and super-resolution DiTs with interleaved spatial and temporal attention blocks. To enable S2V generation with minimal impact on the foundation model, we adopt a lightweight subject injection framework that requires neither additional vision encoders [7,27,42] nor extra preprocessing steps such as subject segmentation [7,11,44]. As illustrated in Figure 3, during training, reference images and video frames are first encoded by the same VAE, yielding embeddings $\mathbf{z}^{\mathrm{r}_i} \in \mathbb{R}^{n\times c\times h\times w}$ and $\mathbf{z}^{\mathrm{v}} \in \mathbb{R}^{f\times c\times h\times w}$, respectively. The reference and frame embeddings are then concatenated along the temporal dimension:

$$\mathbf{z}_t^{\mathrm{noised}} = \mathrm{Concat}([\mathbf{z}_t^{\mathrm{v}}, \mathbf{z}_t^{\mathrm{r}_i}], \mathrm{dim} = 0) \in \mathbb{R}^{(f+n)\times c\times h\times w}, \tag{1}$$

where $\mathbf{z}_t^* = (1-t)\mathbf{z}^* + t\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The reasons for concatenating the reference frames after the video frames are twofold: first, to distinguish our approach from I2V tasks; second, to avoid introducing temporal RoPE offsets for the video tokens, which could potentially degrade performance.

Reference conditions are built by padding the reference frames with zeros:

$$\mathbf{z}^{\mathrm{cond}} = \mathrm{Pad}([\mathbf{z}^{\mathrm{r}_i}, \mathbf{0}], \mathrm{dim} = 0) \in \mathbb{R}^{(f+n)\times c\times h\times w}. \tag{2}$$

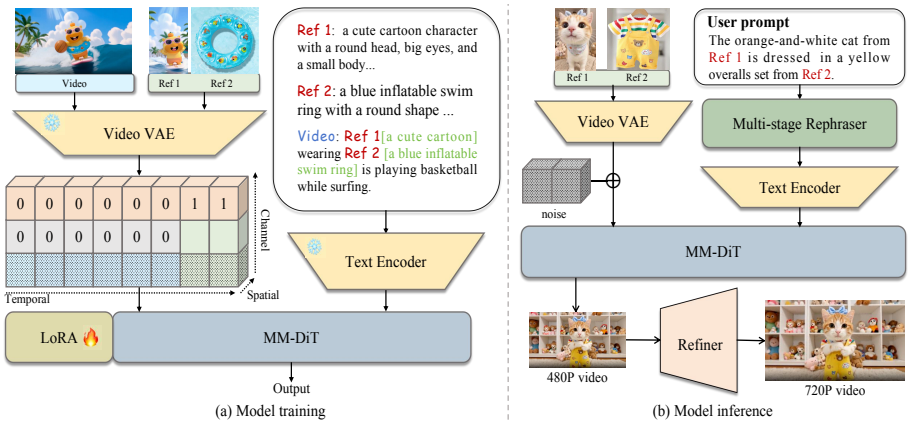


**Fig. 3: Training and inference pipelines of LibraGen**. During training, reference-related embeddings are concatenated with video frame embeddings along the temporal dimension, accompanied by dedicated flags. During inference, the user prompt is processed by a multi-stage prompt rephraser to better align with the training captions. The base MM-DiT produces 480P outputs, which can be further refined to 720P.

We then concatenate $\mathbf{z}^{\text{cond}}$ with $\mathbf{z}_t^{\text{noised}}$ along the channel dimension, together with dedicated binary flags $\mathbf{f}$ to indicate reference embeddings:

$$\mathbf{z}_t^{\text{input}} = \text{Concat}([\mathbf{z}_t^{\text{noised}}, \mathbf{z}^{\text{cond}}, \mathbf{f}], \text{dim} = 1) \in \mathbb{R}^{(f+n)\times(2c+1)\times h\times w}. \quad (3)$$

### 3.2 Multi-stage Prompt Rephraser

During inference, the user prompt, as shown in Figure 3, is often misaligned with the training captions (detailed in Section 4.1), which may degrade generation quality. To bridge this gap, we propose a multi-stage prompt rephrasing strategy. In the first stage, we use Qwen 3 [2] to generate fine-grained descriptions for each reference subject. In the second stage, the VLM extracts distinguishable coarse-level descriptions of the reference subjects and merges them with the user prompt. In the third stage, the processed prompt is appropriately elaborated by adding relevant details about subject motion and scene context.

## 4 Model Training

We frame the extension of T2V/I2V foundation models to S2V as task-incremental learning, which requires a balance between pre-trained knowledge and novel task acquisition [35, 37]. We propose a quality-over-quantity approach, featuring a data curation pipeline that generates thousand-scale human-aligned training samples (Figure 4). Furthermore, we introduce the Tune-to-Balance post-training paradigm to effectively harmonize existing and new model capabilities.

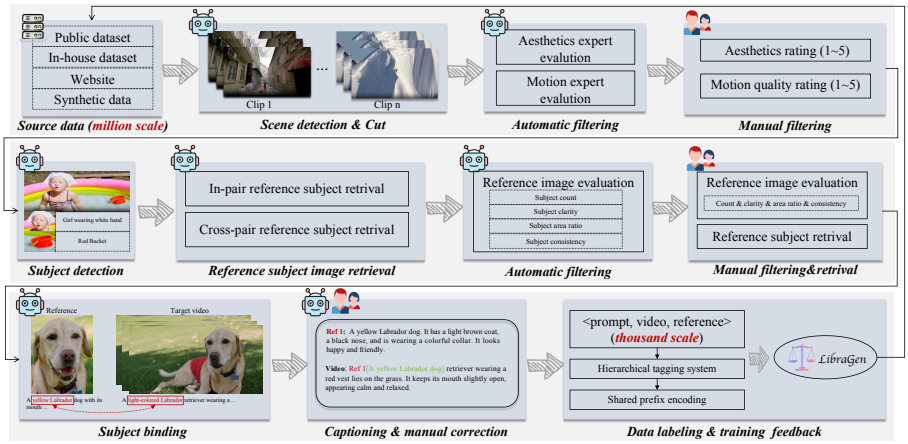


**Fig. 4: Human-aligned data curation pipeline**. The automatic–manual hybrid data curation pipeline follows a quality-over-quantity strategy and consists of four stages: video collection, reference subject extraction, data captioning, and data labeling with dynamic updates.

### 4.1 Human-Aligned Data Curation

**Video Collection**. Our video data originates from diverse sources, including Panda70M [6], an in-house collection of 3 million video clips, online platforms, and synthesized content. Long videos are segmented using AutoShot [47] and PySceneDetect [34] into 2–11-second clips. We employ a dual-stage filtering process: first, clips are scored by motion and aesthetics experts [26], retaining the top 5%. Second, we conduct manual evaluations for motion coherence, amplitude, and aesthetics on a 1–5 scale, keeping only those with an average score of 4 or higher. This pipeline ensures strong motion dynamics and high visual quality across our dataset.

**Reference Subject Extraction**. We adopt the S2V detection pipeline [8] to extract bounding boxes of primary subjects from keyframes. These subjects are further cropped and matched against a large-scale retrieval bank [8] to construct `<video, reference>` pairs. For cross-pair data, we retrieve the top 15 most similar samples per query subject, excluding the source video, and randomly select one as the reference. For in-pair data, subjects detected within the same video clip serve as references. To ensure quality, we filter ambiguous subjects, enforce subject consistency for cross-pair data, and constrain subject scales to prevent background leakage in in-pair data. This process combines automated VLM-based filtering with manual verification and data supplementation.

**Data Captioning**. We caption each `<video, reference>` pair in the form of $\texttt{<ref}_i\ \texttt{text, video text>}$. We first caption each reference subject and then perform subject binding using the video description and metadata `<subject, bounding box, reference>` provided by Stage 2. For in-pair data, where reference images are directly drawn from the target video, the subject name in

the video description (highlighted in red in Fig. 3) is replaced with the refined reference description (highlighted in green). For cross-pair data, we identify the corresponding reference for each subject using the metadata and perform the replacement accordingly. Finally, manual verification is conducted to ensure accurate data captioning.

**Data Labeling and Dynamic Updating**. Each `<prompt, video, reference>` triplet is annotated using a hierarchical tagging system. Categories at each level are assigned unique identifiers, and each triplet is represented via shared-prefix encoding. This data tagging system enables us to efficiently analyze statistics and dynamically adjust the training data distribution based on model feedback.

### 4.2 Supervised Fine-tuning

**Low-resolution**. To preserve the inherent capabilities of the foundation model to the greatest extent, we adopt Low-Rank Adaptation (LoRA) [18] for fine-tuning. Recent studies [11, 27] have demonstrated that fine-tuning exclusively on in-pair data gives rise to severe copy-paste artifacts and degraded prompt-following performance. In contrast, training solely on cross-pair data mitigates these issues but compromises subject consistency. Leveraging this complementarity, we propose fine-tuning two distinct models on in-pair and cross-pair data, respectively, followed by a linear interpolation of their respective LoRA weights. This interpolation yields a smoother loss landscape, thereby enhancing overall performance while achieving a flexible trade-off.

**Super-resolution**. To enable subject-consistent super-resolution, reference subjects are incorporated to guide the super-resolution process. We construct quadruple data `<prompt, 480P video, 720P video, reference>` for super-resolution SFT. Specifically, high-resolution videos are sampled from the dataset constructed in Sec. 4.1, and their low-resolution counterparts are synthesized via downsampling combined with slight noise-induced blurring to simulate real-world quality degradation. Since the super-resolution network shares a similar architecture with the low-resolution model [15], we initialize it by merging the LoRA obtained in the low-resolution SFT phase. Using the constructed quadruple data, we perform fine-tuning with an exponential moving average.

### 4.3 Tailored DPO

**Consis-DPO**. To further enhance subject consistency, we propose a subject-consistency-enhanced DPO strategy. The key challenge is to construct positive–negative pairs that differ significantly in subject consistency while minimally impacting other dimensions, such as motion and visual quality. To address this, we design a dedicated pipeline that synthesizes such pairs by manipulating the RoPE offsets of reference images during model inference. Specifically, the positive sample is generated by the SFT model trained on cross-pair data, since cross-pair data yield good visual and motion effects. The corresponding negative sample is produced using an identical inference configuration, including the same initial noise, sampling steps, and conditions; the only modification is increased

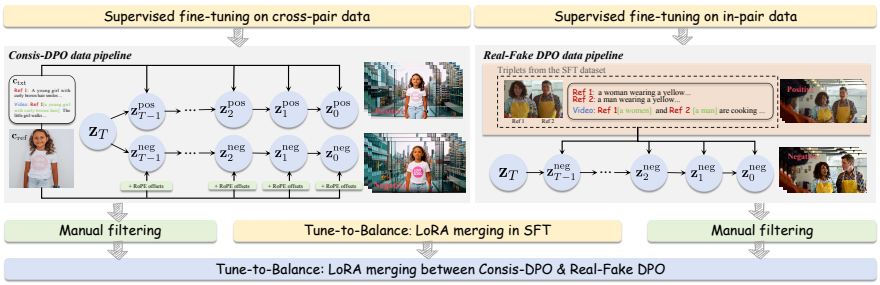


**Fig. 5: Tune-to-Balance post-training paradigm.**


**Fig. 6: The impact of temporal RoPE offsets on reference frames.**

RoPE offsets applied to reference-related embeddings in the temporal attention blocks, as illustrated in Figure 5. Figure 6 visualizes the impact of temporal RoPE offsets on reference frames. As the offset increases, fine-grained ID details gradually degrade, while the global structure and background generally remain intact, thereby justifying the Consis-DPO design. We generate a large number of such positive–negative candidate pairs and then carefully curate a high-quality preference dataset at the thousand-sample scale. Only pairs in which the positive sample demonstrates clearly superior subject consistency are retained.

**Real-Fake DPO**. While Consis-DPO effectively improves subject consistency, it tends to slightly degrade performance on other evaluation dimensions in practice, particularly motion dynamics. The detailed analysis is presented in Appendix B. This degradation arises mainly from the low quality of positive samples generated by the SFT model and DPO's inherent preference hacking issue [14, 16]. To address these problems, we propose a Real–Fake pair data curation pipeline. Specifically, positive samples are directly selected from training triplets during the SFT stage, which already exhibit strong visual quality and motion dynamics. The corresponding negative samples are generated by the model fine-tuned on in-pair data, using identical prompts and reference images as the positive counterparts. This design is motivated by the observation that the in-pair fine-tuned model tends to produce copy-and-paste artifacts, a desirable property for constructing effective negative samples. We then construct a large candidate

set of positive–negative pairs and carefully curate a high-quality dataset at the thousand-sample scale by retaining only pairs in which negative samples maintain strong subject consistency but suffer from poor visual and motion quality.

Building upon the LoRA-merged SFT model, we also adopt LoRA to enable efficient DPO training, with the training objective defined as

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{p_\theta(\mathbf{z}^{\mathrm{pos}}|\mathbf{c}_{\mathrm{txt}},\mathbf{c}_{\mathrm{ref}})}{p_{\theta_{\mathrm{ref}}}(\mathbf{z}^{\mathrm{pos}}|\mathbf{c}_{\mathrm{txt}},\mathbf{c}_{\mathrm{ref}})} - \beta\log\frac{p_\theta(\mathbf{z}^{\mathrm{neg}}|\mathbf{c}_{\mathrm{txt}},\mathbf{c}_{\mathrm{ref}})}{p_{\theta_{\mathrm{ref}}}(\mathbf{z}^{\mathrm{neg}}|\mathbf{c}_{\mathrm{txt}},\mathbf{c}_{\mathrm{ref}})}\right)\right]. \tag{4}$$

$\theta_{\mathrm{ref}}$ is initialized from the low-resolution SFT model and updated via an exponential moving average of $\theta$ after each iteration. We also employ LoRA to train two models separately on paired data generated by Consis-DPO and Real–Fake DPO. Finally, we merge the LoRA weights to achieve a desirable trade-off.

### 4.4 Time-dependent dynamic CFG

During inference, we follow prior methods [4, 27] and adopt CFG with separate guidance scales for each modality:

$$\begin{aligned}\mathbf{v}_\theta(\mathbf{z}_t,\mathbf{c}_{\mathrm{txt}},\mathbf{c}_{\mathrm{ref}}) &= \mathbf{v}_\theta(\mathbf{z}_t,\varnothing,\varnothing) + \omega_1\left(\mathbf{v}_\theta(\mathbf{z}_t,\mathbf{c}_{\mathrm{ref}},\varnothing) - \mathbf{v}_\theta(\mathbf{z}_t,\varnothing,\varnothing)\right)\\ &+ \omega_2\left(\mathbf{v}_\theta(\mathbf{z}_t,\mathbf{c}_{\mathrm{ref}},\mathbf{c}_{\mathrm{txt}}) - \mathbf{v}_\theta(\mathbf{z}_t,\mathbf{c}_{\mathrm{ref}},\varnothing)\right),\end{aligned} \tag{5}$$

where $\omega_1$ and $\omega_2$ control the guidance scales of the reference subjects $\mathbf{c}_{\mathrm{ref}}$ and text prompts $\mathbf{c}_{\mathrm{txt}}$, respectively. During the early stages of denoising, DiT predominantly captures structural information, including spatial object layouts and temporal motion trajectories, which are primarily guided by text prompts. In contrast, later stages emphasize visual details such as textures and colors, which are more strongly influenced by reference images [29]. Motivated by this observation, we introduce a time-dependent dynamic CFG strategy, in which $\omega_1$ progressively increases over time $t$ while $\omega_2$ correspondingly decreases. This strategy enables flexible and fine-grained control over both reference images and textual prompts without causing unexpected visual degradation.

## 5 Experiments

### 5.1 Setup

**Implementation Details**. In the low-resolution SFT phase, we train on 9,000 samples (5,500 cross-pair and 3,500 in-pair). We increase the reference image drop ratio for in-pair training (to mitigate copy-paste artifacts) and the text drop ratio for cross-pair training (to boost subject consistency), with early stopping at 8,000 and 5,000 iterations, respectively. For super-resolution SFT, 4,000 low-resolution samples are selected to build quadruple data, followed by 5,000 more fine-tuning iterations. In the DPO phase, Consis-DPO and Real-Fake pipelines generate 2,000 and 1,600 positive-negative pairs, respectively, with 4,000 iterations per training run. The SFT objective is based on Rectified Flow [28],

| Method | Motion Quality | | Visual Quality | | | Text Align. |
|---|---|---|---|---|---|---|
| | MS ↑ | MQ ↑ | AES ↑ | IQA ↑ | VQ ↑ | TA ↑ |
| Vidu Q1 [1] | <u>0.5373</u> | <u>0.9924</u> | <u>0.6491</u> | <u>71.82</u> | <u>2.795</u> | 3.315 |
| Kling O1 [22] | 0.4965 | 0.9865 | 0.6479 | **72.84** | **2.797** | 3.567 |
| MAGREF [11] | 0.3830 | 0.9853 | 0.6356 | 70.36 | 2.354 | 1.784 |
| Phantom [27] | 0.3844 | 0.9873 | 0.6410 | 69.35 | 2.373 | 1.998 |
| LibraGen (ours) | **0.5380** | **0.9930** | **0.6496** | 71.60 | <u>2.795</u> | **3.594** |

**Table 1: Quantitative evaluation of motion, visual quality, and text alignment** using text prompts and reference images as inputs. The best results are shown in bold, and underlined results indicate the second-best performance.

adopting the noise sampling strategy from [13]. During training and inference, each window in temporal DiT blocks only perceives a partial reference subject, which may impair subject consistency. We thus introduce an all-gather strategy to enable full reference subject capture.

**Evaluation Dataset and Baselines**. To evaluate LibraGen, we select Kling O1 [22] and Vidu Q1 [1] as commercial baselines, and Phantom [27] and MAGREF [11] as open-source baselines. Our evaluation utilizes an in-house benchmark of 200 diverse test cases encompassing both single- and multi-subject generation. The dataset is distributed across 1, 2, 3, and 4 reference images with proportions of 26%, 50%, 14%, and 10%, respectively. To test real-world robustness, the benchmark includes uncropped reference images where subjects may occupy only a small fraction of the frame, spanning categories such as animated IPs, humans, and consumer products. Additionally, we evaluate LibraGen on the public benchmark OpenS2V-Nexus [41], with detailed experimental settings and results presented in Appendix A.

**Evaluation Metrics.** To evaluate performance, we employ Motion Smoothness (MS) [20] and Motion Quality (MQ) [26] to quantify temporal fluidity and general motion fidelity. Visual appeal and perceptual integrity are assessed using the Aesthetic Score (AES) and Image Quality Assessment (IQA) from the VBench [20]. To capture human preferences more accurately, we incorporate the Visual Quality (VQ) reward model [26]. Furthermore, semantic consistency between text prompts and generated videos is measured via the Text Alignment (TA) reward model [26]. To assess subject consistency, we calculate the GSB Ratio, defined as $(G - B)/(G + S + B)$, to determine the relative win rate against baseline models. In this metric, $G$, $S$, and $B$ denote instances where our model is superior to, comparable to, or inferior to the reference, respectively. The quantitative evaluation results are summarized in Table 1.

### 5.2 Comparison with Other Methods

As reported in Table 1, LibraGen achieves superior performance across multiple dimensions. In terms of motion dynamics, it leads in both Motion Smoothness

| Baseline | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Vidu Q1 [1] | 0.154 | 0.140 | 0.143 | 0.100 |
| Kling O1 [22] | 0.077 | 0.080 | 0.071 | 0.100 |
| MAGREF [11] | 0.423 | 0.620 | 0.429 | 0.700 |
| Phantom [27] | 0.308 | 0.300 | 0.286 | 0.500 |

**Table 2: GSB ratio results relative to different S2V baselines**. We compare subject consistency in single- and multi-subject scenarios (2, 3, and 4 reference images). Higher values indicate better subject consistency performance of LibraGen.

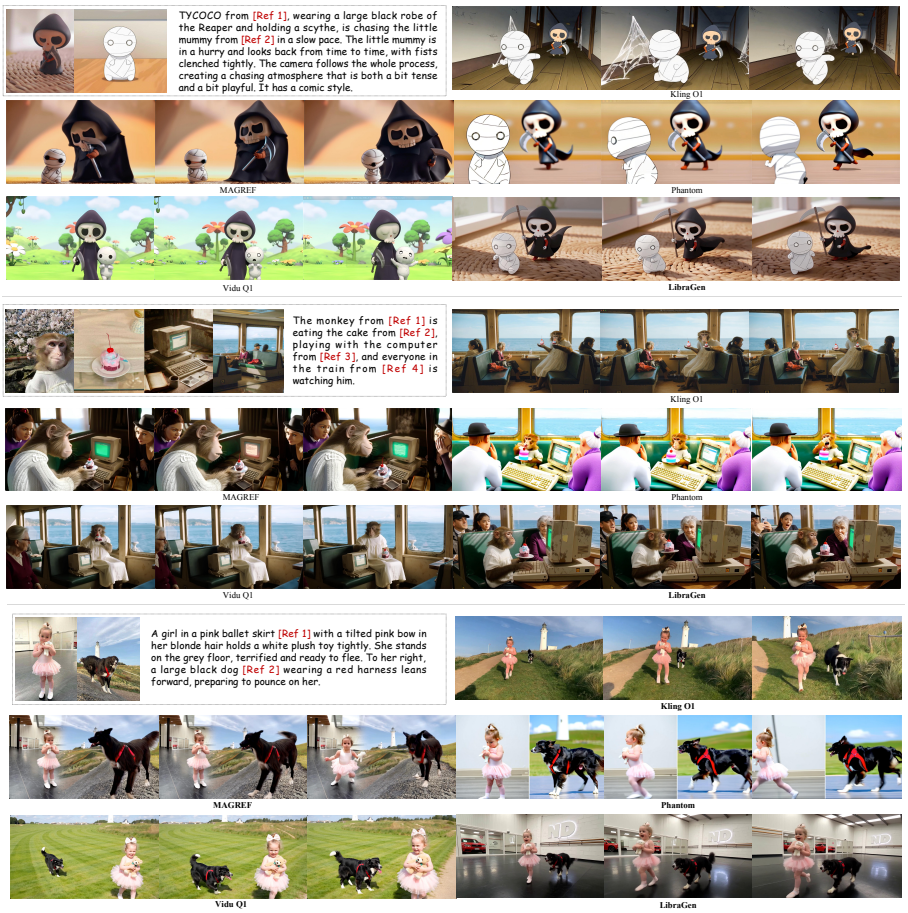


**Fig. 7: Visual comparison with state-of-the-art S2V models**. The first, middle, and last frames are sampled for each method for visualization. Zoom in for details.

and Motion Quality, underscoring its ability to generate fluid and physically consistent movements. Regarding visual perception, LibraGen obtains the best AES


**Fig. 8: Merging the LoRA module trained on in-pair data in the SFT phase**. As the merging coefficient increases, subject consistency gradually improves, but the copy-paste phenomenon becomes increasingly severe.

and remains highly competitive in IQA and VQ, performing on par with leading models like Kling O1 and Vidu Q1. Furthermore, LibraGen achieves state-of-the-art performance in prompt following, surpassing other S2V models and ensuring that the generated videos remain semantically faithful to the prompts.

The core strength of LibraGen in maintaining subject identity is further substantiated by the GSB ratio results reported in Table 2. Across all experimental settings, ranging from single-subject to intricate multi-subject configurations, LibraGen consistently achieves positive GSB ratios against all baselines, demonstrating a significantly higher win rate in terms of subject fidelity. Notably, our approach exhibits marked superiority over MAGREF, with GSB ratios peaking at 0.700. In particular, empirical observations indicate that Phantom outperforms MAGREF within our evaluation framework; this discrepancy is largely attributable to MAGREF's condition injection mechanism, which involves concatenating all reference images into a single frame, a constraint that requires high spatial occupancy by subjects for effective encoder capture. Consequently, when presented with test cases involving small-scale subjects in the reference images, MAGREF inevitably suffers from performance degradation. These findings highlight LibraGen's robustness in preserving fine-grained subject details and ensuring identity stability, effectively addressing the persistent challenges of identity blurring and semantic drift inherent in contemporary subject-driven generation frameworks. We further provide a visual comparison in Figure 7 to intuitively showcase the effectiveness of our approach. Notably, LibraGen maintains robust subject consistency when handling multi-subject generation tasks.

### 5.3 Training and Inference Strategy Analysis

**Tune-to-Balance in SFT**. In the SFT phase, the model is first trained on cross-pair data. However, as evidenced by the results with $l_{\text{in-pair}} = 0$ in Figure 8, this approach leads to suboptimal subject consistency. Cross-pair training diminishes the alignment between reference subjects and target videos, causing the model to overly prioritize text prompts at the expense of visual cues. To


**Fig. 9: Merging the LoRA module trained on data generated through the Consis-DPO pipeline**. Consis-DPO effectively improves human identity consistency without introducing the copy-paste phenomenon.

| Strategy | Motion Quality | | Visual Quality | | | Text Align. | Subject Consist. |
|---|---|---|---|---|---|---|---|
| | MS ↑ | MQ ↑ | AES ↑ | IQA ↑ | VQ ↑ | TA ↑ | GSB ratio ↑ |
| baseline | 0.9912 | 0.5350 | 0.6420 | 71.87 | 2.794 | 3.504 | - |
| + $l_{\text{in-pair}} = 0.15$ | 0.9894 | 0.5218 | 0.6306 | 71.45 | 2.708 | 3.466 | 0.190 |
| + $l_{\text{Consis-DPO}} = 0.5$ | 0.9907 | 0.5239 | 0.6334 | 71.10 | 2.714 | 3.490 | 0.286 |
| + $l_{\text{Real-Fake}} = 0.1$ | 0.9913 | 0.5342 | 0.6394 | 71.42 | 2.756 | 3.524 | 0.025 |
| + dynamic CFG | 0.9930 | 0.5380 | 0.6496 | 71.60 | 2.795 | 3.594 | 0.020 |

**Table 3: Analysis of training and inference strategies**. Note that "+" indicates a cumulative addition to the baseline. GSB ratios are benchmarked against the previous setup, where the metric is calculated regardless of the reference subject count.

mitigate this issue, we incorporate a LoRA module fine-tuned on in-pair data. As shown in Figure 8, increasing the merging coefficient $l_{\text{in-pair}}$ improves subject fidelity; however, excessively high values induce copy-and-paste artifacts due to overfitting to the reference pose. We therefore adopt $l_{\text{in-pair}} = 0.15$ to strike an effective balance. According to the ablation results in Table 3, merging the in-pair LoRA at this weight yields only a minor compromise in motion smoothness and visual quality.

**Tune-to-Balance in RL**. Simply merging the LoRA module trained on in-pair data still yields suboptimal subject consistency, as evidenced by the results in Figure 9. Figure 9 shows that human identity similarity improves progressively as the merging coefficient $l_{\text{Consis-DPO}}$ increases, demonstrating the effectiveness of Consis-DPO. Notably, Consis-DPO does not further compromise visual quality, motion quality, or text alignment, as reported in Table 3. This resilience is primarily attributed to our use of cross-pair data for constructing positive samples; this strategy mitigates "model hacking" (*e.g.*, background replication) and encourages the model to focus specifically on identity consistency. Finally, to bolster the motion and visual quality of the generated videos, we integrate a Real-Fake LoRA model with a coefficient $l_{\text{Real-Fake}} = 0.1$. As indicated in Table 3, this merging significantly enhances both motion and visual quality, while subject consistency and text alignment remain robust without degradation.

**Tune-to-Balance in Inference**. The aforementioned results were generated using a static CFG strategy with $\omega_1 = \omega_2 = 3.5$. In this section, we evaluate a dynamic strategy where $\omega_2$ linearly decreases from 5 to 1, while $\omega_1$ increases from 1 to 4. We also provide visual results in Appendix C. As shown in Table 3, this dynamic CFG schedule enhances text alignment without degrading other metrics, demonstrating its effectiveness. However, this strategy inevitably increases inference latency, as it requires two additional velocity function computations compared to the standard case where $\omega_1 = \omega_2$.

## 6 Conclusion

We present LibraGen, a principled framework for S2V generation that unifies the strengths of T2V/I2V foundation models with robust subject-consistent video generation. We show that data quality, rather than quantity, is a key factor in improving overall S2V performance. Furthermore, we propose a novel post-training paradigm guided by the philosophy of "Raising the Fulcrum, Tuning to Balance." Experiments demonstrate that LibraGen achieves state-of-the-art performance, surpassing competitive open-source and commercial methods. We believe that LibraGen offers a promising roadmap for adapting off-the-shelf T2V/I2V foundation models to S2V tasks without requiring full fine-tuning or million-scale training data. Moreover, the design principles of LibraGen are extensible and can be readily applied to emerging audio-video generation models.

## Acknowledgements

This project is supported by the National Natural Science Foundation of China (12326618, U22A2095), the National Key Research and Development Program of China (2024YFA1011900), the Major Key Project of PCL (PCL2024A06), and the Project of Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026).

## References

1. AI, S.: Vidu. `https://www.vidu.cn/create/character2video`, accessed June 30, 2026 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
4. Chen, L., Ma, T., Liu, J., Li, B., Chen, Z., Liu, L., He, X., Li, G., He, Q., Wu, Z.: Humo: Human-centric video generation via collaborative multi-modal conditioning. arXiv preprint arXiv:2509.08519 (2025)

5. Chen, S., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Chi, X., Cong, J., Cui, Q., Dong, Q., Fan, J., et al.: Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv:2512.13507 (2025)
6. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., et al.: Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13320–13331 (2024)
7. Chen, T.S., Siarohin, A., Menapace, W., Fang, Y., Lee, K.S., Skorokhodov, I., Aberman, K., Zhu, J.Y., Yang, M.H., Tulyakov, S.: Multi-subject open-set personalization in video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6099–6110 (2025)
8. Chen, Z., Li, B., Ma, T., Liu, L., Liu, M., Zhang, Y., Li, G., Li, X., Zhou, S., He, Q., Wu, X.: Phantom-data: Towards a general subject-consistent video generation dataset. arXiv preprint arXiv:2506.18851 (2025)
9. DeepMind, G.: Veo 3.1. `https://deepmind.google/models/veo/`, accessed June 30, 2026 (2025)
10. Deng, Y., Guo, X., Wang, Y., Fang, J.Z., Wang, A., Yuan, S., Yang, Y., Liu, B., Huang, H., Ma, C.: Cinema: Coherent multi-subject video generation via mllm-based guidance. arXiv preprint arXiv:2503.10391 (2025)
11. Deng, Y., Guo, X., Yin, Y., Fang, J.Z., Yang, Y., Wang, Y., Yuan, S., Wang, A., Liu, B., Huang, H., et al.: Magref: Masked guidance for any-reference video generation. arXiv preprint arXiv:2505.23742 (2025)
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
13. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
14. Fisch, A., Eisenstein, J., Zayats, V., Agarwal, A., Beirami, A., Nagpal, C., Shaw, P., Berant, J.: Robust preference optimization through reward model distillation. arXiv preprint arXiv:2405.19316 (2024)
15. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
16. Gupta, D., Fisch, A., Dann, C., Agarwal, A.: Mitigating preference hacking in policy optimization with pessimism. arXiv preprint arXiv:2503.06810 (2025)
17. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
19. Huang, Y., Yuan, Z., Liu, Q., Wang, Q., Wang, X., Zhang, R., Wan, P., Zhang, D., Gai, K.: Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. arXiv preprint arXiv:2501.04698 (2025)
20. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)

21. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
22. Kling Team, K.T.: Kling-omni: Omni video generation model from multi-modal visual language (Dec 2025), https://arxiv.org/abs/2512.16776, technical Report; Model ID: Kling-Omni (Kling-O1); accessed June 30, 2026 at https://app.klingai.com/global/omni/new
23. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
24. kuaishou: Kling 2.5. https://app.klingai.com/, accessed June 30, 2026 (2025)
25. Liang, F., Ma, H., He, Z., Hou, T., Hou, J., Li, K., Dai, X., Juefei-Xu, F., Azadi, S., Sinha, A., et al.: Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13146–13156 (2025)
26. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
27. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. arXiv preprint arXiv:2502.11079 (2025)
28. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
29. Lv, Z., Si, C., Pan, T., Chen, Z., Wong, K.Y.K., Qiao, Y., Liu, Z.: Dual-expert consistency model for efficient and high-quality video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14983–14993 (2025)
30. MiniMax: Hailuo 2.3. https://hailuoai.com/create/subject-reference-to-video, accessed June 30, 2026 (2025)
31. OpenAI: Sora2. https://sora.chatgpt.com/, accessed June 30, 2026 (2025)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
33. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
34. Team, P.D.: Pyscenedetect: An open-source video scene detection tool. https://www.scenedetect.com/, accessed June 30, 2026 (2024)
35. Van de Ven, G.M., Tuytelaars, T., Tolias, A.S.: Three types of incremental learning. Nature Machine Intelligence **4**(12), 1185–1197 (2022)
36. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
37. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5362–5383 (2024)
38. Xue, B., Duan, Z.P., Yan, Q., Wang, W., Liu, H., Guo, C.L., Li, C., Li, C., Lyu, J.: Stand-in: A lightweight and plug-and-play identity control for video generation. arXiv preprint arXiv:2508.07901 (2025)
39. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)

40. Yin, Y., Zhao, Y., Zheng, M., Lin, K., Ou, J., Chen, R., Huang, V.S.J., Wang, J., Tao, X., Wan, P., et al.: Towards precise scaling laws for video diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18155–18165 (2025)
41. Yuan, S., He, X., Deng, Y., Ye, Y., Huang, J., Ma, C., Luo, J., Yuan, L., et al.: Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. Advances in Neural Information Processing Systems **38** (2026)
42. Yuan, S., Huang, J., He, X., Ge, Y., Shi, Y., Chen, L., Luo, J., Yuan, L.: Identity-preserving text-to-video generation by frequency decomposition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12978–12988 (2025)
43. Zeng, Y., Wei, G., Zheng, J., Zou, J., Wei, Y., Zhang, Y., Li, H.: Make pixels dance: High-dynamic video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8850–8860 (2024)
44. Zhang, Z., Teng, J., Yang, Z., Cao, T., Wang, C., Gu, X., Tang, J., Guo, D., Wang, M.: Kaleido: Open-sourced multi-subject reference video generation model. arXiv preprint arXiv:2510.18573 (2025)
45. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
46. Zhou, Z., Liu, S., Liu, H., Qiu, H., An, Z., Ren, W., Liu, Z., Huang, X., Ng, K.W., Xie, T., et al.: Scaling zero-shot reference-to-video generation. arXiv preprint arXiv:2512.06905 (2025)
47. Zhu, W., Huang, Y., Xie, X., Liu, W., Deng, J., Zhang, D., Wang, Z., Liu, J.: 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. pp. 2238–2247 (2023)