OPENWAN-LIGHT: HIGH-QUALITY DATA ENHANCING AND ACCELERATING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in video generation have greatly improved visual quality and temporal coherence, but the growing scale of modern models makes inference increasingly expensive, limiting their use in real-time applications. While models and acceleration algorithms are becoming increasingly open, the data used to train high-quality accelerated video models remain comparatively underexplored. We introduce a 450K real-video dataset centered on rich and diverse motion while preserving fine visual detail and strong text alignment, curating from 4.55M clean, standardized candidate clips. Our data-curation pipeline combines standard filtering and agentic VLM-based review to identify and balance dynamic video content. To demonstrate its effectiveness, we further build a complete acceleration pipeline on Wan2.2 A14B, combining supervised fine-tuning and few-step distillation with a generator-matched one-step refiner for high-resolution detail recovery. Without quantization or sparse attention, our system achieves acceleration over the reference generation pipeline and generates a 5-second 480p video in approximately 4.9 seconds on single NVIDIA H100, while maintaining strong generation quality and user preference. We will publicly release the dataset, model checkpoints, and complete training and inference code to support reproducible research on efficient and dynamic video generation.

1 INTRODUCTION

Recent advances in large-scale generative models have substantially improved video generation in realism, temporal coherence, and visual detail (Brooks et al., 2024; Polyak et al., 2024; Seedance et al., 2026; Yang et al., 2025; Kong et al., 2024; Wan Team, 2025a; MiniMax AI, 2026). Beyond content creation, recent world models such as Genie 3 (Parker-Holder & Fruchter, 2025) and NVIDIA Cosmos (NVIDIA et al., 2025) further highlight the potential of video generation as a foundation for interactive world simulation, embodied agents, and physical AI. Real-time generation is particularly important for interactive applications and physical AI, yet the growing scale of modern video models makes inference increasingly expensive. This has motivated faster samplers and few-step distillation methods (Song et al., 2021; Lu et al., 2022; Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024b). Recent approaches such as DMD2 (Yin et al., 2024a) and rCM (Zheng et al., 2025) further improve few-step generation through adversarial supervision, score distillation, and consistency training, but aggressive acceleration can still compromise motion richness and fine-grained visual detail (Chen et al., 2026; Zheng et al., 2025).

While models and acceleration algorithms are increasingly open, open data for training high-quality accelerated video models remain comparatively underexplored. Existing acceleration approaches are often data-free or rely primarily on synthetically generated samples (Zheng et al., 2025), whereas real-video data can provide complementary supervision for motion and visual detail that may be difficult to obtain from teacher generation alone. This motivates the development of open real-video resources that better capture these challenging visual dynamics.

Existing datasets such as OpenVid-1M (Nan et al., 2025), UltraVideo (Xue et al., 2025), and DropletVideo (Zhang et al., 2025b) provide valuable collections of high-quality video-text data, but rich and diverse motion is not their primary curation objective. We therefore introduce a curated real-video dataset that places rich motion at the center of data selection while preserving fine visual detail and strong text alignment. Our curation pipeline combines standard filtering and VLM-based

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

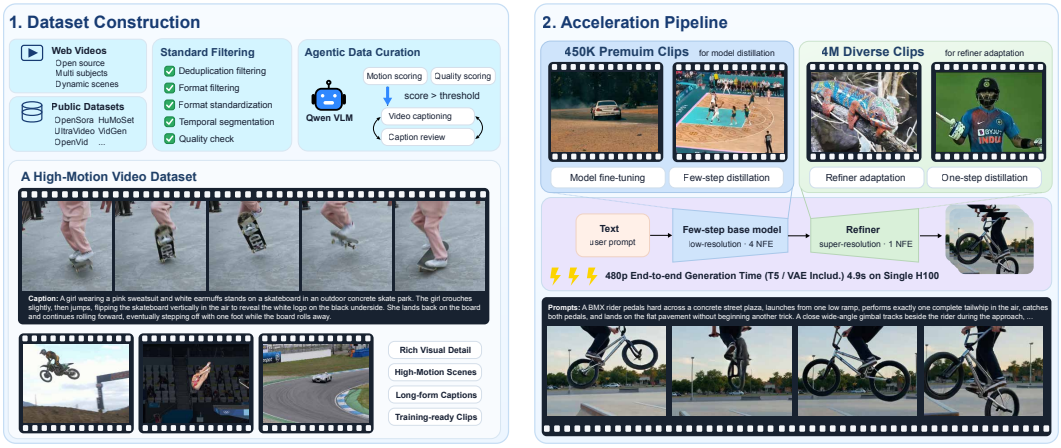


Figure 1: Overview of dataset construction and acceleration. Left: Coarse processing retrieves and standardizes heterogeneous videos, after which fine processing with agentic review selects useful dynamic content and constructs verified captions. Right: the resulting motion-centric data support base-model adaptation and distillation, while a broader dataset supports high-resolution refinement.

review to identify and balance dynamic video content. Although motivated by real-time generation, the resulting dataset is broadly applicable to motion-sensitive video learning, including supervised post-training, autoregressive modeling, super-resolution, and evaluation. Moreover, the dataset are composed of clean clips with standardized spatial and temporal resolutions, directly fit into common academic training pipelines with minimal additional preprocessing.

To demonstrate the effectiveness of the dataset, we further build a complete acceleration pipeline combining supervised fine-tuning, few-step distillation, and a 14B learned refiner Wan model (Figure 1). We observe clear improvements in complex dynamic scene generation and instruction following after SFT on our curated data. Furthermore, without quantization or sparse attention, the system achieves 76× acceleration over the reference generation pipeline and generates a 5-second video in approximately 4.9 seconds on single NVIDIA H100 hardware. Under our evaluation setting, it achieves the lowest latency and highest user preference among the evaluated open implementations while preserving competitive generation quality. To enable reproducible and extensible research, we will publicly release the dataset, model checkpoints, and complete training and inference codebase.

Our contributions are:

- **An two-stage video data-curation pipeline.** We develop a scalable two-stage curation framework that combines lightweight standard filtering with VLM-assisted semantic review. The agentic stage conditionally performs motion-aware selection, caption generation, and object grounding verification. Semantic clustering and category classification further improve diversity and balance across dynamic content.

- **A rich-motion video-text dataset.** Using this pipeline, we construct a curated real-video dataset centered on rich and diverse motion while preserving visual quality and caption alignment. The dataset is motivated by accelerated generation but is broadly applicable to supervised post-training, autoregressive video modeling, refinement, and evaluation.

- **An open end-to-end video acceleration pipeline.** We build a completeacceleration pipeline combining supervised fine-tuning, few-step distillation, and learned refiner model, achieving 76× acceleration and approximately real-time generation on single NVIDIA H100 GPU without quantization or sparse attention. To support reproducibility and further research, we will publicly release the dataset, model checkpoints, and complete training code.

2 Related Work

Video datasets. Large-scale video–text datasets have evolved from web-scale collections with weak supervision, such as WebVid and HD-VILA-100M (Bain et al., 2021; Xue et al., 2022), toward increasingly structured and automatically annotated corpora such as InternVid and Panda-70M (Wang et al., 2023; Chen et al., 2024). More recent generation-oriented datasets place greater emphasis on visual quality, motion, and caption fidelity: Open-Sora Plan develops a multi-dimensional curation pipeline (Lin et al., 2024), MiraData targets long-form and motion-rich content (Ju et al., 2024), OpenVid-1M introduces aesthetic, temporal, motion, and clarity filtering (Nan et al., 2025), UltraVideo focuses on native high-resolution videos with detailed descriptions (Xue et al., 2025), VideoUFO focuses on user-requested topics (Wang & Yang, 2025), and DropletVideo emphasizes continuous dynamic events together with subject and camera motion (Zhang et al., 2025b). High-resolution sequence collections such as Inter4K, UVG, REDS, and SJTU 4K additionally support interpolation, restoration, and video-quality research (Stergiou & Poppe, 2022; Mercat et al., 2020; Nah et al., 2019; Song et al., 2013). Our work builds on this direction but places rich and diverse motion at the center of data selection.

Video generation acceleration. Video diffusion acceleration spans faster numerical solvers, few-step distillation, coarse-to-fine generation, and systems-level optimization. DDIM and DPM-Solver reduce the number of sampling steps without retraining (Song et al., 2021; Lu et al., 2022), while progressive distillation, Consistency Models, LCM, DMD/DMD2, PCM, and rCM learn increasingly direct or distribution-matched generation trajectories (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Yin et al., 2024b;a; Wang et al., 2024; Zheng et al., 2025). Recent methods further improve few-step generation through intermediate distribution matching, temporal regularization and explore causal distillation (Fan et al., 2025; Zhang et al., 2026c; Huang et al., 2025). In parallel, cascaded approaches such as Imagen Video and Pyramidal Flow shift substantial computation to lower spatial or temporal resolutions (Ho et al., 2022; Jin et al., 2024), while SeedVR2 demonstrates efficient one-step diffusion refinement (Wang et al., 2025). Systems including FastVideo, LightX2V, TurboDiffusion, Helios, and Sparse VideoGen additionally exploit sparse attention, quantization, caching, parallelism, or optimized kernels for deployment (FastVideo Team, 2025; Zhang et al., 2026a; LightX2V Team, 2025; Zhang et al., 2025a; Yuan et al., 2026; Xi et al., 2025; Zhang et al., 2026b). Building on these advances, we develop an acceleration pipeline that push the achievable speedup of large video models to an extreme regime while maintaining generation quality.

3 Dataset Construction

Our dataset effort transforms heterogeneous video collections into clean and formatted clips for video generation. We develop a two-stage curation pipeline consisting of standard filtering followed by VLM-assisted agentic curation. The first stage standardizes video formats and removes invalid samples at scale, while the second performs semantic and aesthetic filtering, motion-aware selection, and captioning with multimodal models. For concreteness, the pipeline applies the Wan training format (480p, 5 s, and 16 fps), although the overall procedure is model-agnostic. As summarized in Table 1, 1.19M locally processed source videos yield 7.11M formatted clips. After standard filtering and deduplication, 4.55M clips remain. After careful curation, a motion-oriented subset contains 450,613 clips source videos, deriving from the larger formatted set. We plan to publicly release the wan-formatted curated data, together with additional clean high-resolution, longer-duration set and another audio-visual subset.

3.1 Standard Filtering

We begin with a broad mixture of existing open video datasets and web videos. Webvideos are collected using diverse semantic categories and motion-oriented search terms. At this stage, our goal is to obtain a large pool of clean clips in a consistent training format rather than aggressively filtering for specific semantic properties. To keep preprocessing scalable and inexpensive, we therefore rely primarily on lightweight CPU-based operations to segment and normalize videos to a fixed duration, resolution, and frame rate, while removing clearly unusable samples such as corrupted, severely blurred, or invalid clips. Detailed preprocessing settings are provided in Appendix A.1. As summarized in Table 1, this standard filtering stage retains 4.55M clean, standardized clips for

Table 1: Source statistics and curated subsets.

Source	Source videos	Encoded clips	Quality-filtered	Curated (450K)
WebVideo	30,114	4,375,330	2,427,169	6,972
Pexels	410,245	1,290,872	967,544	382,062
OpenVid	624,192	1,091,993	931,679	2,768
Pixabay	56,629	208,923	110,844	25,550
UltraVideo	55,885	129,941	101,374	28,861
Mixkit	7,834	14,856	9,931	4,400
Inter4K + REDS + UVG	1,273	1,550	1,283	0
Total	1,186,172	7,113,465	4,549,824	450,613

refiner training and subsequent motion-oriented curation. More selective judgments on motion, semantics, and caption quality are deferred to the agentic curation stage.

3.2 Agentic Curation Pipeline

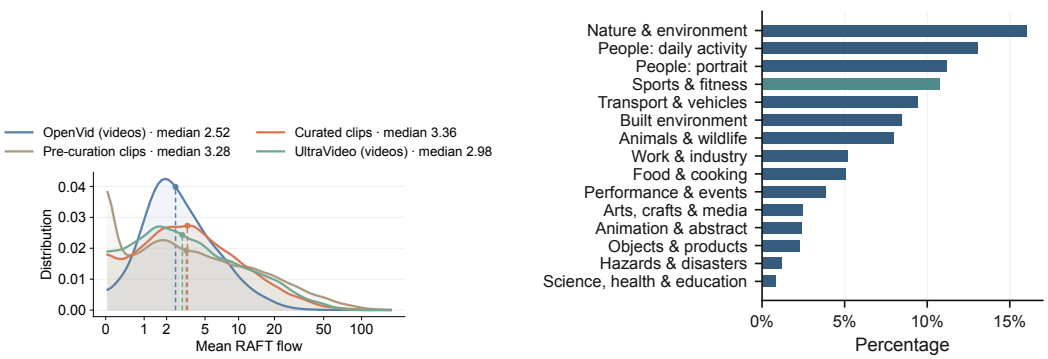
At this million-clip scale data curation pipeline, we target to address two problems at once selecting videos that contain useful visual dynamics and producing captions that faithfully describe them, while keeping the inference cost manageable. Existing pipelines typically combine several specialized metrics. For example, OpenVid separately evaluates aesthetics, clarity, temporal consistency, and optical-flow-based motion (Nan et al., 2025). Such signals are effective for large-scale filtering, but each captures only a narrow aspect of video quality. Motion is particularly difficult to characterize with low-level statistics alone, since similar pixel displacement may arise from subject motion, camera movement, scene cuts, or visual artifacts. We therefore use a general-purpose VLM as a semantic reviewer, allowing several visual, temporal, quality (aethetic) and semantic criteria to be evaluated jointly under programmable prompt-defined rules.

We provide the VLM reviewer with a temporally ordered sequence of uniformly sampled frames together with their timestamps. A single structured review jointly evaluates visual quality, camera motion, slow-motion playback, and shot discontinuities. Because all candidates have a fixed duration and are sampled uniformly in time, the model observes motion on a consistent temporal scale and dentify subject motion rather than background or camera motion. Only clips that pass this review proceed to caption generation, avoiding additional language-model inference on rejected samples.

For each accepted clip, the VLM generates a factual description of the visible subjects, scene, appearance, action progression, and also camera movement. Rather than generate the caption in one single review, we double verify the claims to avoid hallucination. In practice, we find that fine-grained spatial and temporal relations are the most challenging aspects for VLM-based captioning. The reviewer will first identify several semantic objects among the visual inputs, claims involving relative positions of those objects, left to right motion, or motion toward and away from the camera often suffer from insufficient visual grounding, especially when the model must infer changes across multiple frames. We therefore introduce verification for captions. The spatial-temporal verifier re-grounds and checks whether the described displacement corresponds to true object relationships, subject motion and camera movement. Unsupported claims are revised or even removed based on visual evidence.

This conditional routing forms an agentic curation workflow in which VLM reasoning is organized as an iterative verification loop. The initial review determines whether a clip is suitable for further processing, while the generated caption will be reviewed as verification. When verification identifies unsupported or ambiguous descriptions, the workflow selectively triggers re-grounding and revision based on visual evidence. This process is repeated until the annotation satisfies the verification criteria or reaches a predefined retry budget, after which uncertain semantic details are removed when necessary.

We additionally maintain an auxiliary subset as an audio-video set. Its candidate sources include audio-video samples from our collections, VidGen-1M (Tan et al., 2024), and HuMoSet (Chen et al., 2025). This branch applies basic audio quality checks, and uses an omni-modal model to describe jointly observed visual and acoustic events, while an ASR model extracts speech transcripts from



(a) Mean optical-flow distribution.

(b) Semantic-domain composition.

Figure 2: Motion and semantic coverage. (a) Smoothed distributions of mean RAFT optical flow for 10,000 uniformly sampled items per source; dashed lines mark group medians. (b) Distribution of the 450,613 curated clips over 15 semantic domains.

the audio track. The resulting multimodal annotations are subsequently verified against the visual content. The audio-visual subset will be included in the public release together with its alignment evaluation.

3.3 Semantic Balancing and Analysis

We analyze the curated captions to characterize the semantic coverage of our dataset and construct a balanced training mixture. We cluster caption embeddings with spherical k-means and manually review representative samples to derive a scene taxonomy. The resulting dataset covers 15 domains with 334 nonempty scene classes, spanning diverse categories including people, sports, nature, animals, transportation, and environments.

Figure 2(a) characterizes the motion distribution of our data with existing video datasets and pre-curated data using mean RAFT flow (Teed & Deng, 2020; Ju et al., 2024). Our curated clips exhibit a clear stronger motion, with a largest median flow value. Relative to the pre-curation pool, curation further reduces the concentration of low-motion samples while preserving a broad high-motion tail, resulting in a more motion-rich and better-balanced distribution. Figure 2(b) shows coverage across the 15 semantic domains. The scene taxonomy also supports controlled training sampling by limiting overrepresented concepts and increasing exposure to selected motion-rich categories. Further analysis is provided in Appendix A.

4 Acceleration Pipeline

4.1 Pipeline Overview

Cascaded and pyramidal designs have been widely explored in video generation to progressively construct videos across spatial and temporal scales (Jin et al., 2024; Yuan et al., 2026). Rather than performing the entire denoising trajectory at the target resolution, we build the two-stage pipeline illustrated in Figure 3. The **base model** adapts Wan2.2-A14B (Wan Team, 2025b;a) to the 450K motion-oriented manifest and is then distilled into a few-step low-resolution generator responsible for motion, layout, and coarse visual structure. The **refiner** operates at higher spatial resolution and draws from the 4.55M dataset, using degradations generated by the deployed base model to learn generator-specific detail recovery. The following sections describe the training of these two components.

4.2 Base Model

Our base-model training consists of two stages. We first adapt Wan2.2 T2V-A14B to low-resolution generation with supervised fine-tuning (SFT), and then compress the adapted model into a few-step

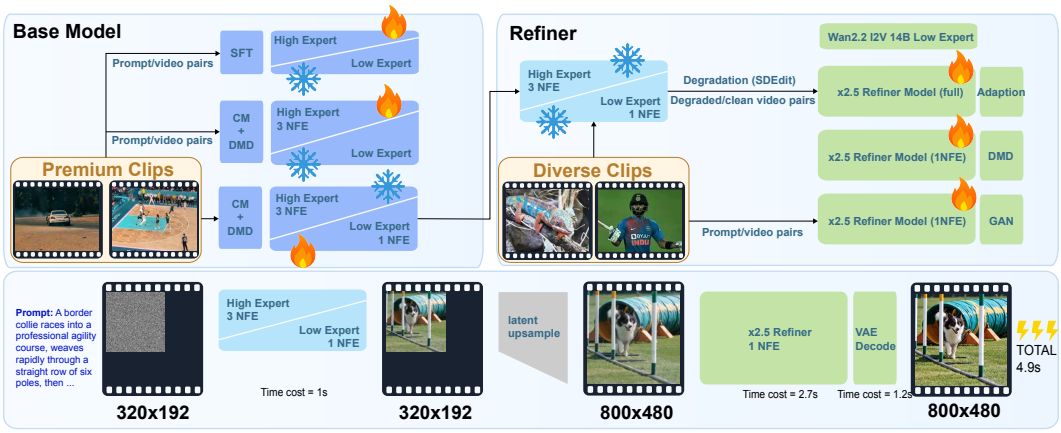


Figure 3: Training and inference pipeline. Curated video-text pairs adapt the base generator, while detail-rich videos provide SDEdit-derived degraded/clean pairs for refiner training. Flame and snowflake icons denote trainable and frozen modules.

generator. Since Wan2.2 is primarily post-trained at its recommended generation resolution (Wan Team, 2025b;a), directly reducing the spatial resolution may therefore weaken its generation capability. Low-resolution SFT first recovers the capability of intruction following and generation ability in complex scences. The supervise training also provides a direct test of whether our curated data can improve motion modeling. After tuning, we apply few-step distillation to obtain the accelerated base model. Full training configurations are provided in the appendix.

Low-Resolution Supervised Fine-Tuning. We fine-tune the high-noise expert of Wan2.2 T2V-A14B at our base resolution using the 450K example video-text subset with balanced sampling. The model is optimized in latent space with the standard rectified-flow velocity-matching objective, conditioned on the corresponding T5 caption embeddings, while the low-noise expert remains fixed. Rather than introducing a new objective, this stage restores generation quality under the resolution shift and adapts the model to the rich-motion distribution of our data. The resulting checkpoint serves as the initialization and teacher for subsequent few-step distillation.

Few-Step Distillation. Distribution-matching objectives such as DMD enable aggressive step reduction, but their mode-seeking behavior can reduce sample diversity and lead to mode collapse (Yin et al., 2024a). Following prior work on consistency-regularized distillation (Wang et al., 2024; Zheng et al., 2025), we combine distribution matching with phased consistency training (PCM). DMD aligns caption-conditioned student samples with the teacher distribution through a learned fake-score model, while PCM provides an anchored consistency objective on real video latents, preserving diversity and motion coverage and stabilize few-step training.

Wan2.2 partitions the denoising trajectory between high- and low-noise experts. Inspired by Fan et al. (2025); Luo et al. (2025), we first distill the high-noise expert to carry samples from noise to the boundary, and then distill the low-noise expert to complete the trajectory. For low-noise training, boundary states are generated by the frozen few-step high-noise student rather than the full-step teacher, matching the inputs encountered at inference and reducing train-inference mismatch. Exact schedules, objectives, loss scale, optimization and inference settings are reported in Appendices C–E.

4.3 Refiner

The refiner is designed to correct artifacts from the deployed few-step base model and recover fine spatial detail at higher resolution. A key challenge is therefore constructing degraded-clean training pairs that faithfully reflect the errors encountered at inference, whereas standard downsampling or synthetic blur does not capture generator-specific distortions, leading to a train-inference mismatch. We address this with an SDEdit-style procedure (Meng et al., 2021) that produces generator-matched degraded inputs from the trained base model. We first adapt a multi-step refiner on these

degraded-clean pairs, and then distill it into a one-step model using distribution matching followed by adversarial post-training for efficient high-resolution refinement.

Generator-matched refiner adaptation. Unlike super-resolution task, our refiner must also correct generator-induced artifacts and missing structure. Standard degradation such as downsampling or synthetic blur poorly represents these inference-time errors. However, training on SDEdit-derived degraded-clean pairs instead exposes the refiner to generator-matched corruptions and allows it to better exploit the generative prior inherited from the I2V model. The refiner can therefore reconstruct texture and repair local artifacts rather than merely sharpen the input. As for the degradation proccess, we first downsample a clean target video to the base resolution and encode it into latent space. We then inject rectified-flow noise near the few-step base model's final denoising boundary and apply the frozen few-step base model for one denoising step. This preserves the underlying content and temporal structure while introducing generator-specific blur, distortion, and local deformation. The resulting latent serves as the degraded condition, while the original high-resolution latent provides the clean refinement target.

We initialize the refiner from the Wan2.2 I2V A14B low-noise expert. Compared with the T2V model, the I2V model provides a stronger prior for visual conditioning, making it a natural initialization for refinement and enabling faster convergence. After bilinearly upsampling the degraded low-resolution latent to the target grid, we inject a small amount of noise before feeding it to the refiner. Rather than treating the refiner as a purely deterministic super-resolution module, it allows the model to retain its generative capability and actively correct generator-induced artifacts or recover missing spatial structure. The refiner is then optimized with the rectified-flow objective on degraded–clean pairs, and the same conditioning procedure is applied during inference.

One-step distillation. We further compress the adapted refiner into a one-step generator through a two-stage procedure. We first apply DMD, using the frozen multi-step refiner as the teacher and an online fake-score model to estimate the student distribution. Starting from the trained DMD checkpoint, we then apply adversarial post-training to further improve perceptual quality and fine detail. At this stage, we remove the teacher and DMD objectives and initialize both the generator and discriminator backbones from the distilled generator and fake critic. A time-conditioned discriminator compares generated and real target latents with a relativistic adversarial loss (Jolicoeur-Martineau, 2019). Training also uses intermediate discriminator feature matching (Wang et al., 2018) and finite-difference approximations to zero-centered gradient penalties on real and generated samples (Mescheder et al., 2018). We omit direct paired-latent reconstruction during this stage, allowing the refiner to prioritize perceptual realism over pointwise reconstruction. Appendix B provides the full objectives and training configuration.

5 Experiments

5.1 Effect of Curated Data on Base Model

To assess how motion-rich supervision affects complex actions and whether its gains survive distillation, we compare the official Wan2.2 teacher, our full-step SFT model, and the corresponding few-step model. Figure 4 presents generated sequences for two action-rich prompts: salsa dancing and a BMX trick. In the salsa example, the teacher produces a relatively limited rotation sequence with small skirt deformation, while the SFT model generates faster and more continuous spinning, with substantially larger skirt movement and more frequent changes in dancer orientation. The few-step distilled model retains these dynamic characteristics, preserving the stronger rotational motion and improved framing after trajectory compression. The BMX example demonstrates an even clearer improvement. While the teacher mainly generates a simple airborne jump with limited changes in bicycle pose, the SFT model captures the more complex trick described in the prompt, showing pronounced bicycle rotation. The few-step distilled model preserves this richer action progression, producing dynamic motion beyond the teacher's simpler interpretation while maintaining object coherence. These results validate the effectiveness of our curated data by exposing the model to real-world dynamic video set, as fine-tuning substantially improves low resolution generation quality, enabling more complex and faithful motion generation. Importantly, these improvements are largely preserved after aggressive few-step distillation.

We further compare few-step distillation objectives on motion-intensive prompts. The qualitative comparison in Appendix F shows that PCM gives the most stable motion and structure among the tested variants in these examples.

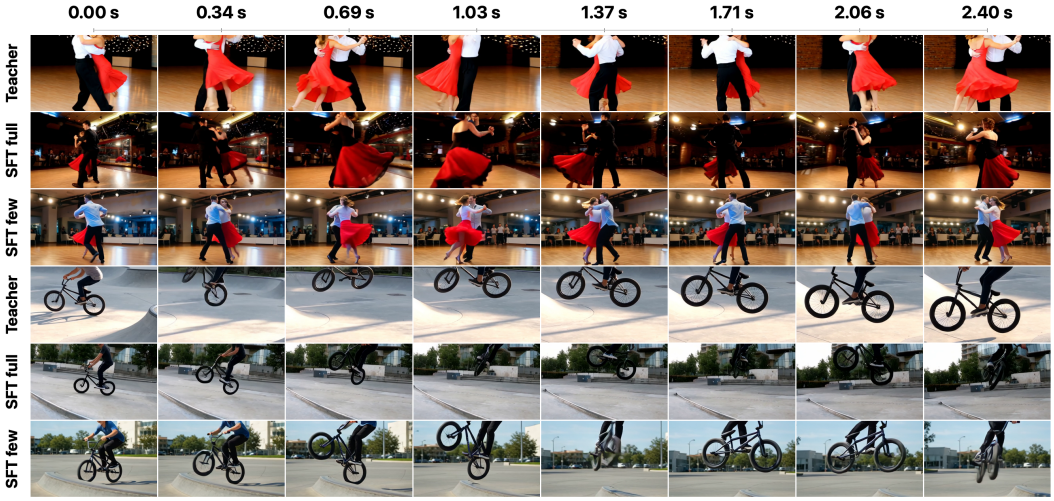


Figure 4: Qualitative comparison of the official teacher, full-step SFT model, and few-step distilled model on salsa dancing (top three rows) and a BMX trick (bottom three rows).

5.2 Effect of Generator-Matched Refiner

We isolate the effect of degradation-aware adaptation while keeping the low-resolution input and refiner backbone fixed. Figure 5 compares the low resolution base output with refiners trained with and without our generator-matched degradation, together with the few-step distilled variant of the degradation-trained refiner. Each row shows the full base frame and an enlarged region for direct comparison of local structure and detail.

As shown in Figure 5, generator-matched training consistently produces sharper and more coherent local structure, including clearer eye and skin texture on the chameleon, sharper butterfly-wing boundaries, and better-defined facial and tooth details of the bearded men. The improvement is also apparent in more challenging dynamic regions. For example, smoke around the explosion exhibits clearer structure, while the gymnast's blurred face and shoulders are restored more cleanly. Importantly, the few-step refiner retains these improvements and can even appear perceptually sharper. DMD and adversarial training preserves the full-step refiner's distributional behavior without imposing pointwise reconstruction, while further promotes realistic high-frequency detail and texture.

5.3 End-to-End Quality and Efficiency

Table 2: Aggregate VBench scores (%) at 480p. TurboDiffusion uses Wan2.1 14B; the other methods use Wan2.2 A14B.

Model	Resolution	NFE	Time↓	Speedup↑	Quality↑	Semantic↑	Total Score↑
Wan2.1-T2V-14B							
TurboDiffusion	480p	4	11.7 s	31.8×	84.07	78.57	82.97
Wan2.2-T2V-A14B							
Official teacher	480p	40	372.5 s	1.0×	84.72	80.92	83.96
LightX2V	480p	4	17.7 s	21.0×	84.60	**82.03**	84.09
OpenWan-light	480p	5	**4.9 s**	**76.0**×	**84.94**	81.23	**84.20**

We compare the final OpenWan-light pipeline with the official Wan2.2 A14B teacher, TurboDiffusion (Zhang et al., 2025a), and LightX2V (LightX2V Team, 2025) using VBench (Huang et al.,

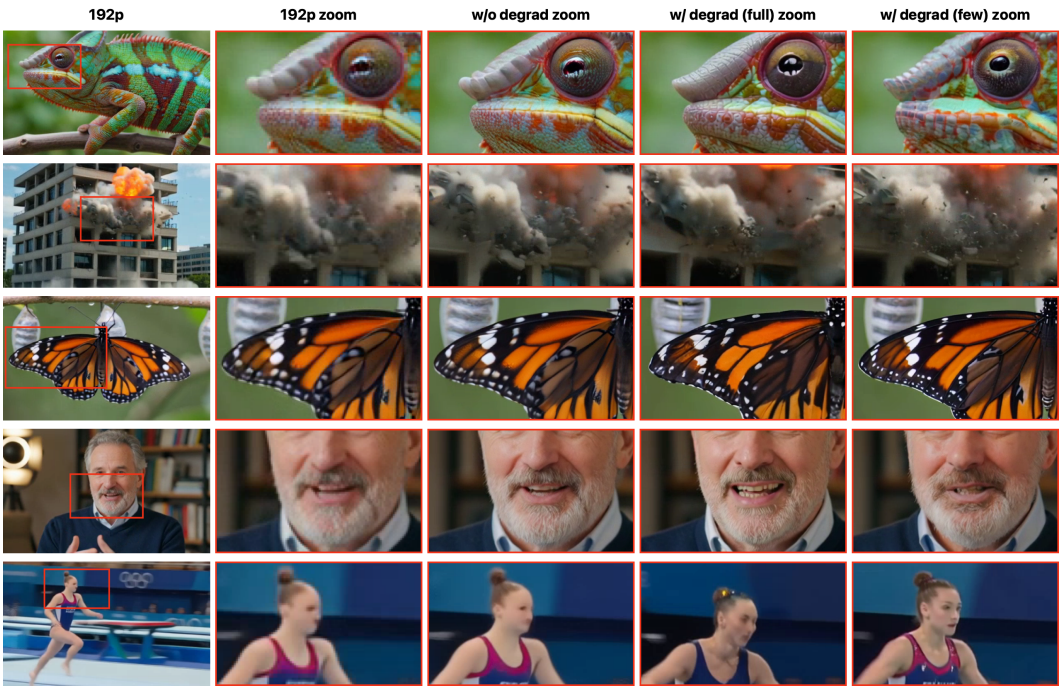


Figure 5: Qualitative effect of generator-matched degradation on refinement. The first two columns show the 192p base output and the enlarged region indicated in red. The remaining columns demonstrates refiner without degradation, generator-matched degradation using the full-step model, and its few-step distilled counterpart.

2024). As shown in Table 2, OpenWan-light achieves a total score of 84.20 and a quality score of 84.94, both the highest among the compared systems, while LightX2V obtains the highest semantic score of 82.03. Notably, OpenWan-light reduces the reported generation time from 372.5 s for the Wan2.2 teacher to 4.9 s, corresponding to a 76.0× speedup, while maintaining comparable or slightly improved aggregate VBench quality. Its reported latency is also substantially lower than LightX2V (17.7 s) and TurboDiffusion (11.7 s). Overall, OpenWan-light provides the best quality–efficiency trade-off among the compared methods, as it achieves the lowest latency together with the strongest aggregate VBench performance. LightX2V attains the highest semantic score, but OpenWan-light achieves the highest total score while operating in a substantially lower-latency regime. Per-dimension results are reported in Appendix G.

6 Conclusion and Limitations

We introduce a high-quality, high-motion aligned video-text dataset and demonstrate its utility through a complete Wan2.2 acceleration pipeline that combines SFT, few-step distillation, and generator-adapted refiner training. Beyond this particular system, the resource is intended to make data-intensive video research more accessible to academic and open-source communities. By releasing the curated dataset together with the training and inference code, we aim to provide a ready-to-use resource for academic-scale research, which lowers the barrier to reproducing our results, exploring alternative acceleration or training strategies and extending the pipeline to new models. Our current release nevertheless remains modest in scale because we prioritize quality over quantity. Moreover, strict quality and motion selection can reduce coverage of long-tail domains, static scenes, and subtle actions, so broader and more diverse data will be needed to establish generalization beyond the distributions and model studied here. Computational constraints led us to use a compact VLM for large-scale filtering. This choice makes the pipeline practical, but may be less reliable than larger models for subtle motion, ambiguous direction, and fine-grained semantic judgments.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for design or provide feedback on research methodology or experiments, implement methods, and proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, draft parts of a research paper, transcribe recordings of research material, summarize or analyse existing literature, brainstorming, sourcing and searching for information, edit a research paper to improve readability, identify relevant literature, format references, suggest a structure for a research paper, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, codes, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work involves the curation and release of video data collected from public sources and existing open datasets. We follow the usage policies and licensing requirements of the original sources and preserve provenance information for all collected data. Our curation pipeline is designed to remove corrupted or low-quality content and does not intentionally collect sensitive personal information. We will apply an access review process for dataset distribution and restrict its use to academic research purposes. This controlled release mechanism aims to reduce potential privacy risks and ensure responsible use of the collected data. Our release of the full video acceleration pipeline is intended to support research on efficient and controllable video generation rather than unrestricted deployment.

REPRODUCIBILITY STATEMENT

We aim to make all components of our work reproducible. We will publicly release the curated datasets, model checkpoints, training and inference code. The main paper describes the overall pipeline, including the dataset construction strategy, base-model acceleration procedure, and refiner training framework. Detailed implementation specifications and training configurations are provided in the appendix and supplementary materials. These resources enable researchers to reproduce the accelerated models, and further study on efficient video generation.

REFERENCES

Max Bain, Arsha Nagrani, G"ul Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2104.00650.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 2024.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. HuMo: Human-centric video generation via collaborative multi-modal conditioning. *arXiv preprint arXiv:2509.08519*, 2025. URL https://arxiv.org/abs/2509.08519.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, et al. Panda-70M: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024. URL https://arxiv.org/abs/2402.19479.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased DMD: Few-step distribution matching distillation via score matching within subintervals. *arXiv preprint arXiv:2510.27684*, 2025. URL https://arxiv.org/abs/2510.27684.

FastVideo Team. FastVideo: A unified inference and post-training framework for accelerated video generation. Open-source project, 2025. URL https://github.com/hao-ai-lab/FastVideo.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. URL https://arxiv.org/abs/2210.02303.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17982.

Yang Jin, Zhicheng Sun, Ningyuan Li, et al. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. URL https://arxiv.org/abs/2410.05954.

Alexia Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard GAN. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1807.00734.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

LightX2V Team. LightX2V: Light video generation inference framework. Open-source project, 2025. URL https://github.com/ModelTC/LightX2V.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, et al. Open-Sora Plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. URL https://arxiv.org/abs/2412.00131.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. URL https://arxiv.org/abs/2310.04378.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17719–17728. IEEE, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. URL https://arxiv.org/abs/2108.01073.

Alexandre Mercat, Marko Viitanen, and Jarno Vanne. UVG dataset: 50/120fps 4k sequences for video codec analysis and development. In *Proceedings of the ACM Multimedia Systems Conference*, pp. 297–302, 2020. doi: 10.1145/3339825.3394937.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3481–3490, 2018. URL https://proceedings.mlr.press/v80/mescheder18a.html.

MiniMax AI. Minimax-h3: An open general-purpose omni-modal generative model. https://github.com/MiniMax-AI/MiniMax-H3, 2026. Open-source model release.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. URL https://openaccess.thecvf.com/content_CVPRW_2019/html/NTIRE/Nah_NTIRE_2019_Challenge_on_Video_Deblurring_and_Super-Resolution_Dataset_and_CVPRW_2019_paper.html.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International conference on learning representations*, 2025.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind, August 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

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.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Li Song, Xun Tang, Wei Zhang, Xiaokang Yang, and Pingjian Xia. The sjtu 4k video sequence dataset. In *International Workshop on Quality of Multimedia Experience*, pp. 34–35. IEEE, 2013.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. URL https://arxiv.org/abs/2303.01469.

Alexandros Stergiou and Ronald Poppe. Adapool: Exponential adaptive pooling for information-retaining downsampling. *IEEE Transactions on Image Processing*, 32:251–266, 2022.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. VidGen-1M: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024. URL https://arxiv.org/abs/2408.02629.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a. URL https://arxiv.org/abs/2503.20314.

Wan Team. Wan2.2: Official model repository. https://github.com/Wan-Video/Wan2.2, 2025b. Official architecture description, accessed September 8, 2026.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.18407.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. SeedVR2: One-step video restoration via diffusion adversarial post-training. *arXiv preprint arXiv:2506.05301*, 2025. URL https://arxiv.org/abs/2506.05301.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_High-Resolution_Image_Synthesis_CVPR_2018_paper.html.

Wenhao Wang and Yi Yang. Videoufo: A million-scale user-focused dataset for text-to-video generation. *Advances in Neural Information Processing Systems*, 38, 2025.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. URL https://arxiv.org/abs/2307.06942.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, et al. Sparse VideoGen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025. URL https://arxiv.org/abs/2502.01776.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5026–5035. IEEE, 2022.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, and Dacheng Tao. UltraVideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025. URL https://arxiv.org/abs/2506.13691.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024a. URL https://arxiv.org/abs/2405.14867.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.pdf.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026. URL https://arxiv.org/abs/2603.04379.

Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E. Gonzalez, Jianfei Chen, and Jun Zhu. TurboDiffusion: Accelerating video diffusion models by 100–200 times. *arXiv preprint arXiv:2512.16093*, 2025a. URL https://arxiv.org/abs/2512.16093.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Faster video diffusion with trainable sparse attention. *Advances in Neural Information Processing Systems*, 38:152509–152534, 2026a.

Runze Zhang, Guoguang Du, Xiaochuan Li, Qi Jia, Liang Jin, Lu Liu, Jingjing Wang, Cong Xu, Zhenhua Guo, Yaqian Zhao, Xiaoli Gong, Rengang Li, and Baoyu Fan. DropletVideo: A dataset and approach to explore integral spatio-temporal consistent video generation. *arXiv preprint arXiv:2503.06053*, 2025b. URL https://arxiv.org/abs/2503.06053.

Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yifu Zhang, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Ping Luo. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12735–12743, 2026b.

Wenhu Zhang, Kun Cheng, Changyuan Wang, Shiyao Li, Yuechen Zhang, Wenbo Li, Jiajun Zha, Jingyi Zhang, Kang Zhao, and Jiaya Jia. CoDMD: Copula-aware distribution matching distillation for fast video generation. *arXiv preprint arXiv:2606.21982*, 2026c. URL https://arxiv.org/abs/2606.21982.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. *arXiv preprint arXiv:2510.08431*, 2025. URL https://arxiv.org/abs/2510.08431.