HunyuanVideo 1.5 Technical Report

Tencent Hunyuan Foundation Model Team

Abstract

We present HunyuanVideo 1.5, a lightweight yet powerful open-source video generation model that achieves state-of-the-art visual quality and motion coherence with only 8.3 billion parameters, enabling efficient inference on consumer-grade GPUs. This achievement is built upon several key components, including meticulous data curation, an advanced DiT architecture featuring selective and sliding tile attention (SSTA), enhanced bilingual understanding through glyph-aware text encoding, progressive pre-training and post-training, and an efficient video super-resolution network. Leveraging these designs, we developed a unified framework capable of high-quality text-to-video and image-to-video generation across multiple durations and resolutions. Extensive experiments demonstrate that this compact and proficient model establishes a new state-of-the-art among open-source video generation models. By releasing the code and model weights, we provide the community with a high-performance foundation that lowers the barrier to video creation and research, making advanced video generation accessible to a broader audience. All open-source assets are publicly available at `https://github.com/Tencent-Hunyuan/HunyuanVideo-1.5`.

1 Introduction

Recent years have witnessed remarkable advancements in video generation models, with closed-source systems like Kling2.5 [1], Veo3.1 [2], and Sora2 [3] achieving state-of-the-art (SOTA) performance. In the open-source domain, models such as HunyuanVideo [4], StepVideo [5], and Wan2.2 [6] have emerged as notable contenders. However, despite these strides, most SOTA models remain proprietary, limiting accessibility and community-driven innovation. Wan2.2 [6] employs a hybrid Mixture of Experts (MoE) architecture that leverages two 14-billion-parameter expert models to enhance video quality. Although this MoE architecture enhances visual fidelity by leveraging specialized experts for different denoising stages, it inherently introduces computational inefficiencies, as the model must manage multiple large parameter sets (totaling 27B parameters with 14B activated), leading to significant resource demands. While Wan2.2 [6] has attempted to address this with a more compact 5-billion-parameter variant, which utilizes a high-compression 3D VAE to reduce memory footprint, this lightweight model still exhibits limitations. Specifically, its generative capabilities, particularly in maintaining motion stability across frames and achieving the nuanced aesthetic quality required for professional applications, continue to fall short of practical demands. This situation underscores a critical gap: the absence of an open-source model that successfully balances high-quality output with efficient inference.

To address these challenges, we introduce HunyuanVideo 1.5, an open-source 8.3B video generation model designed to achieve state-of-the-art visual quality while maintaining high inference efficiency. Our framework is built on a two-stage pipeline for high-fidelity video synthesis. The first stage employs a Unified Diffusion Transformer (DiT) that supports both text-to-video and image-to-video generation, producing initial video sequences with resolutions from 480p to 720p and durations of 5 to 10 seconds. The second stage utilizes a dedicated Video Super-Resolution Network to upscale the outputs to 1080p, significantly enhancing the final visual fidelity. The main contributions of this report are summarized as follows:

- **Lightweight High-Performance Architecture:** We propose an efficient architecture that integrates an 8.3B-parameter Diffusion Transformer (DiT) with a 3D causal VAE, achieving compression ratios of 16× in spatial dimensions and 4× along the temporal axis. This design minimizes parameter count while fully leveraging the potential of the model, yielding performance comparable to state-of-the-art video generation models.
- **Video Super-Resolution Enhancement:** We develop an efficient few-step super-resolution network that upscales outputs to 1080p. It enhances sharpness while correcting distortions, thereby refining details and overall visual texture.
- **Sparse Attention Optimization:** We introduce a novel Selective and Sliding Tile Attention (SSTA) mechanism that dynamically prunes redundant spatiotemporal tokens. This significantly reduces computational overhead for long video sequences and accelerates inference, achieving an end-to-end speedup of $1.87\times$ in 10-second 720p video synthesis compared to FlashAttention-3 [7].
- **Enhanced Multimodal Understanding:** Our framework utilizes a large multimodal model for precise bilingual (Chinese-English) understanding, combined with ByT5 for dedicated glyph encoding to enhance text generation accuracy in videos. Additionally, detailed bilingual captions are generated for both images and videos.
- **End-to-End Training Optimization:** We demonstrate that the Muon optimizer [8] significantly accelerates convergence in video generation model training, while the multi-phase progressive training strategy—spanning from pre-training to post-training stages—enhances motion coherence, aesthetic quality, and alignment with human preferences, thereby enabling the production of professional-grade content.

This report presents a comprehensive pipeline, beginning with meticulous data preparation—including curation, filtering, and captioning, as detailed in **Section 2**. **Section 3** introduces the core architecture and key algorithms, such as the unified diffusion transformer, the video super-resolution network, and the mechanisms of sparse attention. **Section 4** describes training strategies for text-to-video and image-to-video tasks, covering both pre-training and post-training methodologies. Finally, **Section 5** provides a rigorous evaluation of the model's performance against state-of-the-art models using quantitative metrics and qualitative benchmarks.

2 Data Preparation

2.1 Data Acquisition and Filtering

Our training dataset comprises both image and video data. For image data, we adopted the acquisition and processing pipeline outlined in [9], curating 5 billion images from a pool of over 10 billion for pre-training, with 1 billion reserved for subsequent stages. For video data, building upon the pipeline in [4], we significantly expanded the data volume while refining the filtering mechanisms to enhance quality. The specifics of these processes are detailed below.

Data Acquisition. To continuously improve the model's performance, we prioritized diversity and quality during data acquisition. We sourced raw videos from a variety of channels, ensuring comprehensive coverage across diverse content, filming techniques, camera movements, styles, and scenarios. After basic deduplication and the removal of corrupted files, we obtained more than 10 million hours of raw video data.

To address the high variance in raw video lengths and optimize training efficiency, we segmented the videos into consistent clips ranging from 2 to 10 seconds. Specifically, we utilized PySceneDetect [10] combined with a custom segmentation operator to detect scene boundaries. To eliminate any remaining clips containing transition effects, we applied a secondary filtering stage using a specialized transition classifier.

To mitigate artifacts such as subtitles, logos, and watermarks, we applied spatial cropping to exclude affected regions. To maintain compositional integrity, we discarded clips where the cropped area retained less than 60% of the original frame.

Data Curation. To guarantee the high quality of our training data, we implemented a rigorous, multi-stage filtering pipeline. This pipeline consists of three distinct levels:

- **Basic Filtering**: We employed specialized filters to remove videos with padding borders, stitching artifacts, grid layouts (collages), and static or low-motion scenes.
- **Visual Quality Filtering**: We developed a comprehensive quality assessment operator that evaluates videos across four dimensions: sharpness, detail retention, noise & artifacts, and dynamic range. This allows us to discard clips with poor visual quality.
- **Aesthetic Filtering**: We applied an aesthetic scoring operator based on [11] to evaluate the aesthetic quality of the videos, filtering out those with low aesthetic scores.

Upon completion of these filtering stages, approximately 800 million high-quality video segments remained for pre-training.

2.2 Captioning

High-quality captions play a crucial role in video generation, significantly influencing precise instruction following and generation quality. To generate precise and comprehensive descriptions across diverse data modalities, we developed three specialized captioning models, each tailored for a specific function:

- **Image Captioning**: Generates descriptions for static images, leveraging the same methodology in HunyuanImage-3.0 [9].
- **Video Captioning**: Produces a highly structured, multi-component description. These encompass multi-level textual narratives alongside a detailed set of cinematic and aesthetic properties (e.g., shot type, shot angle, composition, lighting, style, color palette, atmosphere).
- **Image-to-Video (I2V) Instructional Captioning**: This novel module generates text describing the temporal evolution or transformation from the initial frame, detailing changes in both foreground subjects and the background environment.

Addressing the Richness-Hallucination Trade-off. A significant and persistent challenge in this domain is the inherent trade-off between descriptive richness and factual accuracy. Specifically, attempts to generate greater detail often increase the propensity for hallucination or factual inconsistencies. To overcome this, we integrate Reinforcement Learning (RL), specifically OPA-DPO [12], into our caption model post-training pipeline, as shown in Figure 1. This approach is designed to strike an optimal balance, maximizing descriptive detail while simultaneously minimizing hallucinations and factual errors.

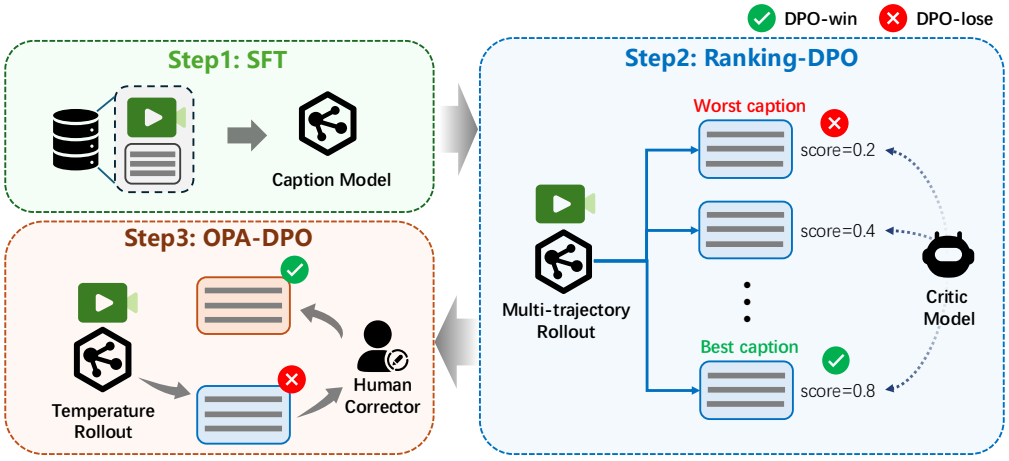

Figure 1: Caption Model Post-training Pipeline.

Camera Movement Description. We developed a specialized recognition model to analyze camera dynamics. This model is capable of identifying multiple camera movement types, operating at both a clip-level (for the whole video) and a sequential (over time) level. High-confidence predictions are converted into natural language descriptions and subsequently integrated into the structured video

Table 1: Architecture hyperparameters for the HunyuanVideo 1.5.

Dual-stream Blocks	Model Dimension	FFN Dimension	Attention Heads	Head dim
54	2048	8192	16	128

captions. The objective of this integration is to empower generative models with controllable camera movement.

3 Model Design

We propose a two-stage framework for high-fidelity video synthesis. In the first stage, we employ a diffusion transformer (DiT) model with 8.3 billion parameters, designed for multi-task learning. The architectural hyperparameters of HunyuanVideo 1.5 are detailed in Table 1. Subsequently, a video super-resolution network is utilized to further enhance visual quality. Comprehensive elaborations are provided in the following subsection.

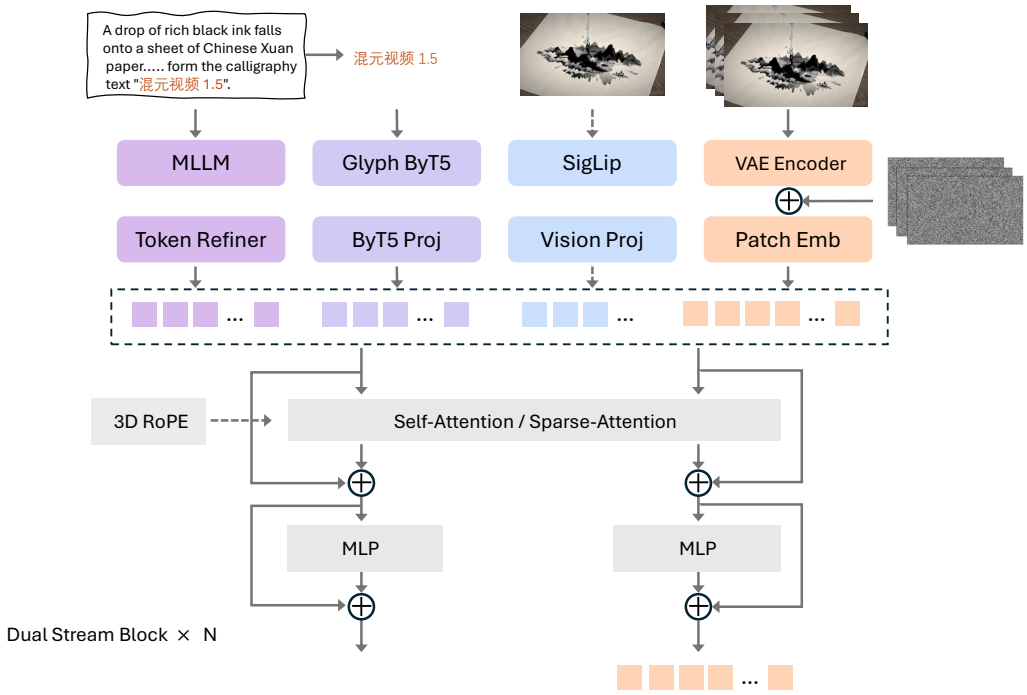

Figure 2: Architecture of the Unified Diffusion Transformer.

3.1 Unified Diffusion Transformer

Multi-Task Training. We developed a unified architecture for jointly training text-to-image (T2I), text-to-video (T2V), and image-to-video (I2V) tasks. For the I2V task, the reference image is integrated into the model via two complementary strategies: (1) VAE-based encoding, where the image latent is concatenated with the noisy latent along the channel dimension to leverage its exceptional detail reconstruction capacity; and (2) SigLip-based feature extraction, where semantic embeddings are concatenated sequentially to enhance semantic alignment and strengthen instruction adherence in I2V generation. A learnable type embedding is introduced to explicitly distinguish between different types of conditions.

VAE. We introduce a causal 3D transformer architecture designed for joint image-video encoding, which achieves a spatial compression ratio of 16× and a temporal compression ratio of 4×, with a latent channel dimension of 32.

Algorithm 1: Selective and Sliding Tile Attention

Require: Query/Key/Value tensor Q,K, $V \in \mathbb{R}^{h\times F\times H\times W\times D}$,
Block size $N = tile_t \times tile_h \times tile_w$,
Block num $B = (F/tile_t) \times (H/tile_h) \times (W/tile_w)$,
Window size $WS = (w_t, w_h, w_w)$,
Top-k.

(1) 3D Block Partition
Partition Q and K into blocks of size N
$Q_b, K_b \leftarrow$ split_blocks(Q, K, N)

(2) Selective Mask Generation
$\bar{Q}_b, \bar{K}_b \leftarrow$ adaptive_pool(Q_b), adaptive_pool(K_b)
$\text{Score}_s = \bar{Q}_b \bar{K}_b^\top \in \mathbb{R}^{h\times B\times B}$ {Q-K block-wise similarity}
$\text{Score}_r = \frac{1}{N-1} \sum_{\substack{i=1 \\ j\neq i}}^{N} [\bar{K}_b \bar{K}_b^\top]_{ij} \in \mathbb{R}^{h\times 1\times B}$ {K-K block-wise redundancy}
$\text{Score}_i = \lambda \cdot \text{Score}_s - \beta \cdot \text{Score}_r$ {Block importance}
Selected_indexes $\leftarrow$ Topk$(\text{Score}_\text{i}, k)$
$M_{sel} \leftarrow$ Indexes2mask(Selected_indexes)

(3) STA (Sliding Tile Attention) Mask Generation
for each block pair (i, j) **do**
  if j within local window of i defined by $WS = (w_t, w_h, w_w)$ **then**
    $M_{\text{sta}}[i, j] \leftarrow 1$
  else
    $M_{\text{sta}}[i, j] \leftarrow 0$
  end if
end for

(4) Combine Masks & Perform Block-Sparse Attention
$M_{combined} \leftarrow M_{sel} \wedge M_{sta}$ {Combine $M_{sel} \in \mathbb{R}^{h\times B\times B}$ and $M_{sta} \in \mathbb{R}^{B\times B}$}
$O \leftarrow$ flex_block_attention$(Q, K, V, M_{combined})$

return O

Text encoder. The model employs a dual-channel text encoder, combining a general textual encoder. This design leverages a visual-language multimodal encoder Qwen2.5-VL [13] to achieve a deeper understanding of scene descriptions, character actions, and specific requirements. Additionally, the incorporation of a multilingual Glyph-ByT5 [14] text encoder further strengthens the model's ability to render text accurately across various languages. This synergistic design enables the model to learn from both high-level semantic representations and fine-grained language-specific features, leading to a more comprehensive and robust text comprehension capability.

Sparse Attention. The computational complexity of standard self-attention scales quadratically with sequence length, posing significant efficiency bottlenecks for transformer-based video generation models. Given the inherent spatiotemporal redundancy in video data, replacing full attention with sparse attention presents a well-motivated optimization strategy. To synergistically combine the benefits of static local window priors and dynamic global adaptive selection, we propose a novel sparse attention method termed SSTA (Selective and Sliding Tile Attention). SSTA features a parameter-free design, enabling seamless integration at any training stage. In our implementation, sparse training is incorporated during the distillation phase, which more effectively preserves output quality while maintaining high computational efficiency. The SSTA algorithm comprises four key steps, as described in Algorithm 1: 3D Block Partition, Selective Mask Generation, STA [15] Mask Generation and Block-Sparse Attention. We propose an engineered acceleration toolkit [16] for sparse attention mechanisms, utilizing the ThunderKittens [17] framework to efficiently implement the flex_block_attention algorithm.

Optimizer. The Muon optimizer [8] is used in this work to achieve faster convergence. We observe that it attains a lower training loss than AdamW in only half the number of training steps, while also yielding superior performance across multiple text-to-image benchmarks. To ensure training stability, a weight decay of 0.01 is applied.

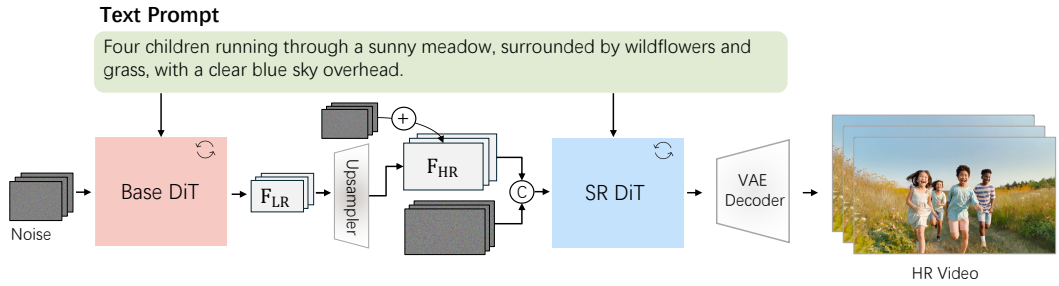

Figure 3: The pipeline of a cascaded video super-resolution model.

3.2 Video Super-Resolution

To further improve training and inference efficiency while meeting the demand for high-resolution video generation, we introduce a cascaded Video Super-Resolution (VSR) model that enhances fine-grained visual details and textures in videos generated by text-to-video and image-to-video based models. We follow the architecture of the main model—an 8.3B Video Diffusion Transformer—and perform super-resolution in the latent space. We inject the LR latents using a channel concatenation approach. To spatially align the LR and HR latents, we separately train a latent upsample block. The final high-quality 1080p video is obtained by decoding these HR latents through the VAE decoder, as shown in Figure 3.

4 Model Training

4.1 Pre-training

We establish a robust and versatile foundation for HunyuanVideo 1.5 through a structured pre-training strategy that systematically develops core generative capabilities across multiple modalities. This phase employs a progressive, multi-stage approach encompassing text-to-image (T2I), text-to-video (T2V), and image-to-video (I2V) tasks. By scaling spatial resolution, temporal length, and frame rate progressively and leveraging mixed-task training with carefully balanced data ratios, we equip the model with strong semantic alignment, visual diversity, and temporal coherence, forming a solid base for subsequent specialization and refinement.

T2I. We begin by training the T2I task, which includes two resolution stages (Stage I & II in Table 2): 256p and 512p. For each stage, we utilize training images with various aspect ratios and organize them into buckets to ensure efficient data sampling. The T2I task enables the model to learn robust semantic alignment between text and images, which we find significantly accelerates the convergence and performance of subsequent T2V & I2V training.

T2V & I2V. Throughout the pretraining phase, we employ a mixed training regimen combining T2I, T2V, and I2V tasks, with a task ratio of 1:6:3. This configuration leverages large-scale T2I data to enhance semantic understanding and generative diversity, while ensuring sufficient video-specific modeling through T2V and I2V tasks. We adopt a progressive multi-stage strategy (Stages III to VI in Table 2), starting from 256p resolution at 16 fps and progressively advancing to 480p and 720p at 24 fps, with video durations ranging from 2 to 10 seconds. This structured scaling of spatiotemporal resolution facilitates stable convergence and strengthens the model's capacity for detailed, coherent video synthesis. Similar to prior studies [18], we observe that flow matching-based training is particularly sensitive to the shift hyper-parameter when video token lengths vary across stages. To maintain training stability under such varying sequence conditions, we carefully design a series of shift scheduling strategies that adapt to different token lengths during the progressive training process.

4.2 Post-training

We conduct a comprehensive post-training pipeline to further enhance the generative quality and task-specific performance of HunyuanVideo 1.5. This multi-stage process consists of continuing training (CT), supervised fine-tuning (SFT), and human feedback alignment (RLHF), applied separately

Table 2: The training process for the three tasks: T2I, T2V, I2V. In the "Data volume" column, the number before the "/" indicates the amount of video data, and the number after the "/" indicates the amount of image data. "M" stands for million.

Stage	Stage	Training Resolution	Data Volumn	Task
I	Pretrain	256p	5 billion	T2I
II	Pretrain	512p	1 billion	T2I
III	Pretrain	256p 16fps 2$\sim$10s	800M	T2V/I2V/T2I
IV	Pretrain	480p 16fps 2$\sim$10s	200M	T2V/I2V/T2I
V	Pretrain	720p 16fps 2$\sim$10s	100M	T2V/I2V/T2I
VI	Pretrain	720p 24fps 2$\sim$10s	100M	T2V/I2V/T2I
VII	CT	480p/720p 24fps 2$\sim$10s	1M	T2V
VIII	CT	480p/720p 24fps 2$\sim$10s	1M	I2V

for text-to-video (T2V) and image-to-video (I2V) tasks. As shown in Figure 4, we present the visualization results during different post-training stages. By progressively refining the model distribution through these dedicated stages, we systematically steer the model toward the optimal output distribution for each task, significantly improving motion coherence, visual fidelity, and alignment with human preferences.

CT. We conduct both CT and SFT stage in 480p and 720p resolution. In the CT stage, we leverage 1 million high-quality video clips per task (as detailed in Table 2, Stages VII & VIII), balanced across diverse categories and sourced from premium datasets. For T2V, we further prioritize clips with high dynamic motion to strengthen temporal modeling. For I2V, we introduce instructional captions that focus exclusively on describing motion and transformations compared to the first frame, thereby enhancing action fidelity. Both T2V and I2V CT processes are initialized from the same optimally pre-trained checkpoint, enabling both tasks to build upon a shared high-quality foundation.

SFT. During SFT, we utilize rigorously selected clips per task, filtered strictly considering aesthetic appeal, clarity, and motion smoothness. Through CT and SFT, we observe evident improvements in output stability, visual quality, aesthetic appeal, and temporal consistency for both generation tasks.

RLHF. We employ Reinforcement Learning from Human Feedback (RLHF) strategies tailored for our I2V and T2V tasks, both mainly aimed at correcting artifacts and enhancing motion quality.

For the I2V task, we apply **online reinforcement learning (RL)** during post-training to correct structural and motion artifacts. Our approach begins with a curated prompt set spanning 100+ categories, constructed from high-aesthetic images. Candidate prompts are first generated via a vision-language model (VLM), then manually verified to ensure strict text-image consistency. We fine-tune a VLM-based reward model to evaluate videos across four key dimensions: text alignment, image alignment, visual quality, and motion dynamics. During RL training, we employ mixed sampling strategies by varying both random seeds and CFG scales, and adopt a hybrid ODE–SDE solver (MixGRPO [19]) to enrich exploration while maintaining sampling quality. This RLHF process yields consistent improvements across all evaluation metrics, with particularly notable gains in motion realism.

Building on this, our T2V alignment strategy uses a more comprehensive **hybrid offline-then-online** approach, designed to address the increased complexity of T2V motion artifacts. We found that existing reward models struggle to effectively differentiate fine-grained motion quality. Therefore, we first conduct an offline optimization stage using Direct Preference Optimization (DPO). For this, we curate a balanced $O(10K)$ prompt set (from LLMs-generated prompts and training video captions) covering diverse dimensions (motion, scene, subject, etc.). Using a selected high-quality SFT checkpoint, we generate N video candidates per prompt to create non-repetitive pairs. These pairs are manually annotated (GSB) for semantic fidelity, motion quality, and aesthetics. Applying Direct Preference Optimization (DPO) [20] to this high-quality paired data significantly reduces motion artifacts and establishes a superior policy starting point. Finally, we conduct an online optimization stage, adopting the exact same online RL framework developed for I2V to further enhance the model's visual quality and semantic-text alignment.

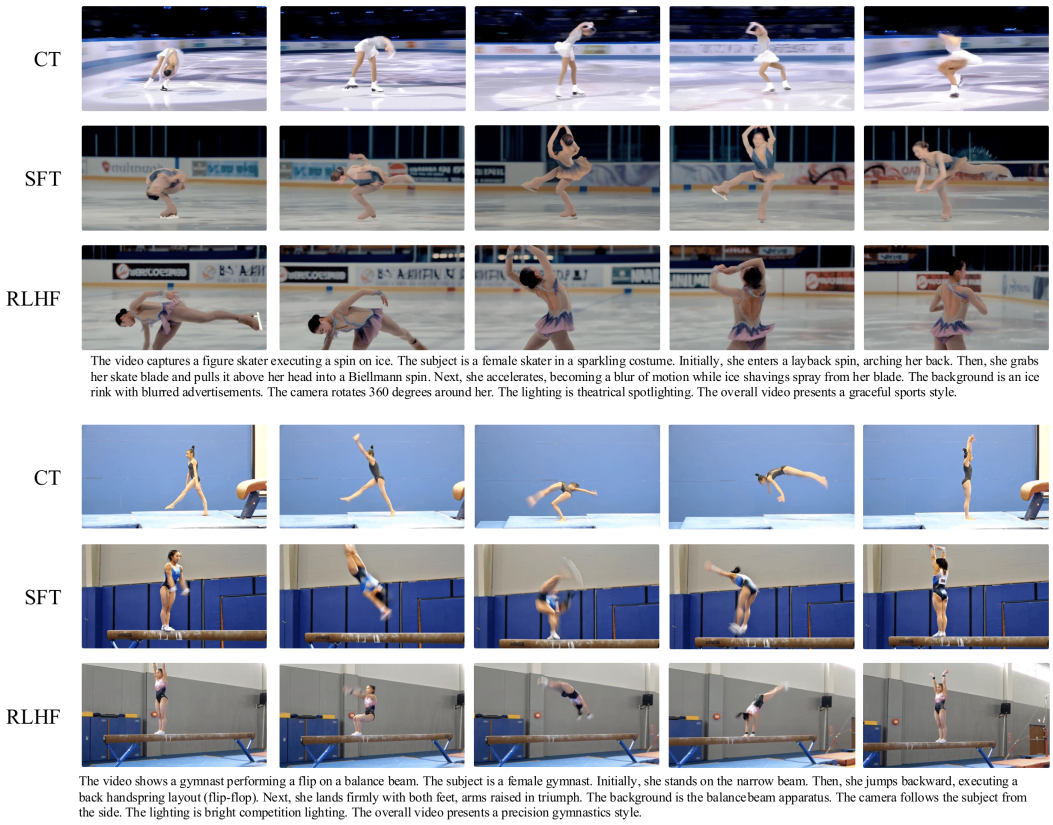

Figure 4: Visualization during different post-training stages

4.3 Video Super-Resolution.

We begin by initializing the weights of our video super-resolution model with a pretrained T2V video model to ensure effective training. For dataset construction, a diverse collection comprising 1 million high-quality video clips is curated. These clips span a wide range of scenarios and resolutions from 1K to 4K, each lasting between 3 and 10 seconds at 24 frames per second. In addition to video data, a set of high-resolution images is incorporated during training to enhance the model's capability to generate fine visual details.

For input preparation, we concatenate the low-resolution latents and noise along the channel dimension as input to the DiT model. All weights of the video super-resolution model are fully trainable throughout the training process, following the flow matching paradigm. As shown in Figure 5, our method significantly enhances visual quality, motion stability, and overall temporal coherence.

5 Model Performance

We adopt two evaluation methods: Rating and GSB (Good/Same/Bad). The Rating method provides a comprehensive assessment of the model from multiple perspectives, enabling us to better diagnose its shortcomings. The GSB approach is widely used to evaluate the relative performance of two models based on overall video perception quality.

5.1 Rating

As shown in Table 3 and Table 4, we assess text-to-video generation using a comprehensive rating methodology that considers five key dimensions: text-video consistency, aesthetic quality of individual frames, visual quality, structural stability, and motion effects. For image-to-video generation, the evaluation encompasses image-video consistency, instruction responsiveness, visual quality, structural stability, and motion effects.

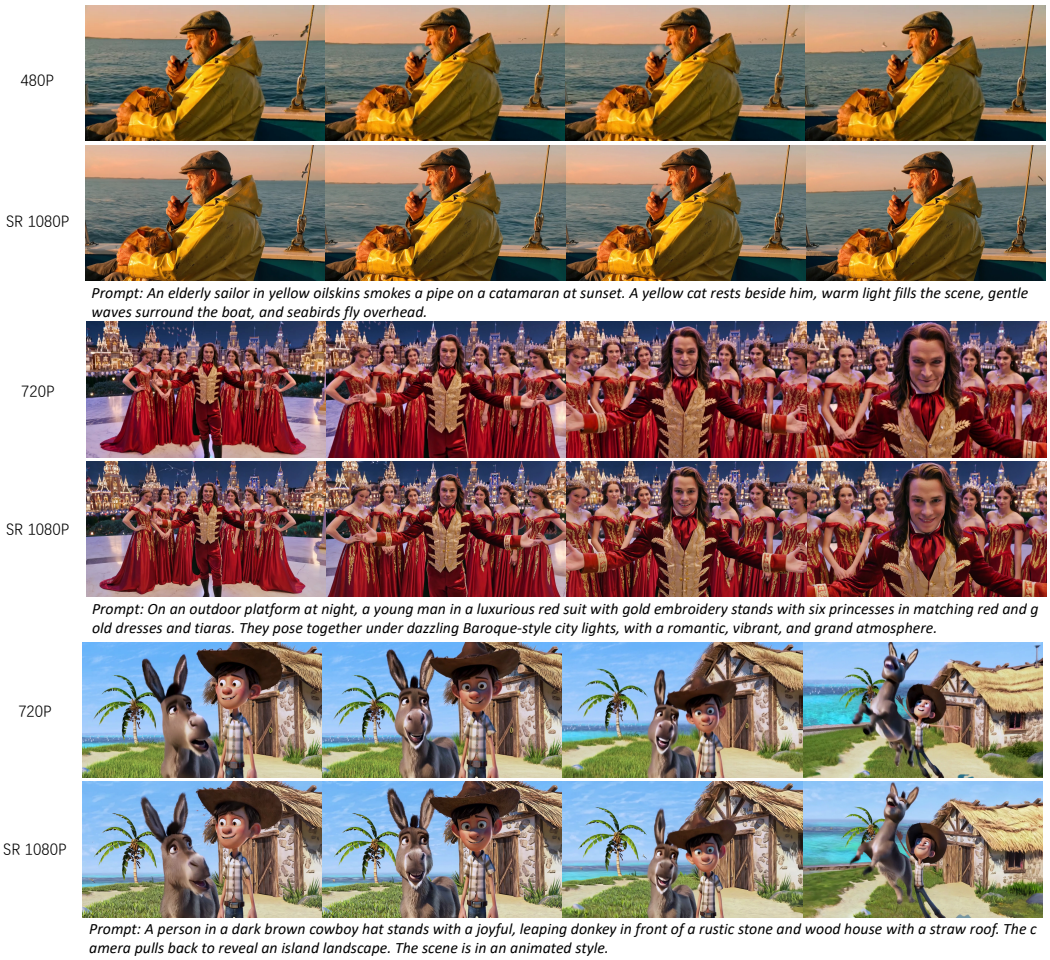

Figure 5: Visualization results of the cascaded video super-resolution model.

This systematic evaluation framework enables us to quickly diagnose model shortcomings and identify specific areas for improvement. By analyzing performance across these dimensions, we can effectively guide optimization efforts to enhance overall model quality and ensure the generated videos better fulfill task requirements.

Table 3: The rating of HunyuanVideo 1.5 T2V model and its competitors

Dimension	HY1.5 720p	Wan2.2	Kling2.1 Master	Seedance Pro	Veo3
Instruction following	61.57	44.07	50.03	53.19	73.77
Aesthetic quality	63.30	65.98	68.00	68.22	67.98
Visual quality	57.35	56.37	59.68	60.20	58.64
Structural stability	79.75	73.75	66.74	68.69	75.62
Motion effects	57.67	53.08	58.59	55.17	60.81

5.2 GSB

In our experiments, we carefully construct 300 diverse text prompts and 300 image samples to cover balanced application scenarios for both text-to-video and image-to-video tasks. For each prompt or image input, an equal number of video samples are generated by each model in a single run to ensure comparability. To maintain fairness, inference is performed only once per input without any

Table 4: The rating of HunyuanVideo 1.5 I2V model and its competitors. "HY1.5 480pSR" means we first generate 480p video and upscale it to 720p using our video super-resolution model.

Dimension	HY1.5 480pSR	HY1.5 720p	Wan2.2	Kling2.1 Master	Seedance Pro	Veo3
Instruction following	63.11	63.05	56.19	68.43	62.90	67.86
Image consistency	68.82	72.07	73.53	64.09	73.06	72.19
Visual quality	59.87	60.33	58.31	59.28	58.95	59.29
Structural stability	70.13	66.67	69.03	59.71	68.01	69.25
Motion effects	57.60	58.62	57.41	57.36	60.47	60.91

cherry-picking of results. All competing models are evaluated using their default configurations. The evaluation is conducted by over 100 professional assessors. The results are shown in Table 5 and 6.

Table 5: GSB of HunyuanVideo 1.5 720P T2V model and its competitors

Compared Model	Wan2.2	Kling2.1 Master	Seedance Pro	Veo3
HY Better	34.90%	31.55%	31.67%	24.64%
Other better	17.78%	18.95%	20.65%	34.96%
Equally Bad	40.70%	42.76 %	38.61 %	30.71%
Equally good	6.62%	6.75%	9.06%	9.69%
HY Win Rate	17.12%	12.6%	11.02%	-10.32%

Table 6: GSB of HunyuanVideo 1.5 720P I2V model and its competitors

Compared Model	Wan2.2	Kling2.1 Master	Seedance Pro	Veo3
HY Better	45.60%	40.60%	30.62%	37.44%
other better	32.95%	30.88%	36.39%	41.05%
Equally Bad	18.60%	22.73%	26.61%	18.44%
Equally good	2.85%	5.79%	6.38%	3.07%
HY Win Rate	12.65%	9.72%	-5.77%	-3.61%

6 Inference Speed and GPU Memory Requirements

This section presents a comprehensive analysis of the inference speed and memory requirements for HunyuanVideo 1.5 under various configurations. All speed measurements are conducted on the classifier-free guidance (CFG) distilled model using 8 NVIDIA H800 GPUs with context parallelism enabled across all devices.

Inference Speed without Engineering Acceleration. To better demonstrate the effectiveness of the proposed sparse attention mechanism, we evaluate inference speed without engineering-level acceleration. No engineering-level acceleration or custom optimization is applied beyond the described model methods and standard distributed inference. All reported results reflect unoptimized, out-of-the-box research code performance. For configurations without sparse attention, Flash Attention v3 [7] is adopted as the attention backend. Inference speed is measured as the wall-clock time per diffusion step, averaged across multiple runs. Table 7 presents the inference speed results under different configurations. It can be observed that without engineering acceleration, the model already achieves competitive inference speeds. This is attributed to the compact 8.3B-parameter architecture that is intrinsically efficient and the high VAE compression rate that significantly reduces the number of tokens required for computation. In addition, sparse attention effectively reduces inference time for long-context sequences, particularly evident in the configurations of 241 frames in Table 7.

Inference Speed with Engineering Acceleration. We also evaluate inference speed with basic engineering-level acceleration techniques enabled to demonstrate practical performance achievable in real-world deployment scenarios. Please note that in this experiment, we do not pursue the most extreme acceleration at the cost of generation quality, but rather to achieve notable speed improvements while maintaining nearly identical output quality. Several acceleration techniques

Table 7: Inference speed per diffusion step for HunyuanVideo 1.5 pipeline without engineering-level acceleration.

Task Type	Resolution	Total Frames	Sparse Attention	Time/Step (s)
T2V	480p (480 × 848)	121	✗	0.9064 ± 0.0062
T2V	480p (480 × 848)	241	✗	1.7015 ± 0.0203
T2V	720p (720 × 1280)	121	✗	2.0084 ± 0.0229
T2V	720p (720 × 1280)	121	✓	1.5638 ± 0.0150
T2V	720p (720 × 1280)	241	✗	5.5070 ± 0.0284
T2V	720p (720 × 1280)	241	✓	2.9475 ± 0.0206

Table 8: Total inference time for 50 diffusion steps for HunyuanVideo 1.5 with acceleration techniques enabled.

Task Type	Resolution	Total Frames	Sparse Attention	Time	Avg Time/Step
T2V	480p (480 × 848)	121	✗	13.90	0.2781
T2V	480p (480 × 848)	241	✗	27.08	0.5418
T2V	720p (720 × 1280)	121	✗	28.33	0.5667
T2V	720p (720 × 1280)	121	✓	26.41	0.5283
T2V	720p (720 × 1280)	241	✗	96.78	1.9356
T2V	720p (720 × 1280)	241	✓	58.39	1.1679

are employed in this experiment: SageAttention [21] is adopted for attention operations to reduce memory complexity and improve computational efficiency; the DiT is compiled using PyTorch's `torch.compile` to enable kernel fusion and operator optimization; a feature caching mechanism is implemented during diffusion sampling to reuse cached intermediate features for non-critical steps, reducing redundant computations. To better reflect practical performance, we report the total inference time for 50 diffusion steps. The results are presented in Table 8. It can be observed that even with engineering acceleration techniques enabled, sparse attention further reduces the inference time for long-context sequences, demonstrating its effectiveness in complementing engineering optimizations.

GPU Memory Requirements. The GPU memory requirements are contingent upon the hardware configuration and the employment of offloading techniques. All memory measurements reported herein are conducted on a single GPU configuration. When pipeline offloading, group offloading, and VAE tiling are enabled, the entire pipeline can complete end-to-end inference for 720p 121-frame T2V/I2V generation with a peak memory of 13.6 GB, thereby enabling inference on a single consumer-grade GPU (e.g., RTX 4090).

7 Conclusion

In this report, we present HunyuanVideo 1.5, a compact 8.3B-parameter Diffusion Transformer (DiT) for open-source video generation that competes with leading proprietary systems. Built via progressive training from a text-to-image model, it achieves high-quality text-to-video and image-to-video synthesis through meticulous data curation and a multi-stage pipeline. The efficient DiT framework integrates SSTA for accelerated inference and a dedicated video super-resolution network for high-resolution output, ensuring state-of-the-art visual quality, motion coherence, and strong multimodal alignment. HunyuanVideo 1.5 exhibits exceptional bilingual prompt understanding, accurate text rendering, and reliable instruction-following. Comprehensive evaluations confirm it sets a new standard among open-source video generators. By releasing this high-performance, lightweight model, we provide an accessible foundation to lower barriers for creative and research applications, fostering broader exploration and innovation.

8 Project Contributors

- **Project Sponsors:** Jie Jiang, Linus, Yuhong Liu, Peng Chen
- **Project Leader:** Zhao Zhong
- **Core Contributors:** (Authors in bolds are project leaders, others are listed alphabetically)
  - **Captioner & Data: Xin Li, Bing Wu**, Duojun Huang, Hao Tan, Xinchi Deng, Xuefei Zhe, Zhenyu Wang
  - **VAE & Model Accleration: Songtao Liu**, Changlin Li, Chang Zou, Fang Yang, Jianbing Wu, Jack Peng, Patrol, Peizhen Zhang, Penghao Zhao, Weiyan Wang, Xiao He, Yang Li, Yuanbo Peng
  - **Pretraining & Post Training: Yue Wu**, Jiangfeng Xiong, Qi Tian, Weijie Kong, Yanxin Long, Zuozhuo Dai
- **Contributors (Listed alphabetically):** Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou

References

[1] Kuaishou Technology. Kling 2.5 turbo. https://app.klingai.com/cn/release-notes/2025-09-19, 2025.

[2] Google DeepMind. Veo 3.1. https://deepmind.google/technologies/veo/, 2025.

[3] OpenAI. Sora 2. https://openai.com/sora, 2025.

[4] Tencent Hunyuan Foundation Model Team. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

[5] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, and Deshan Sun et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. URL https://arxiv.org/abs/2502.10248.

[6] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, and Jingfeng Zhang et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[7] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision, 2024. URL https://arxiv.org/abs/2407.08608.

[8] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, and Jialei Cui et al. Kimi k2: Open agentic intelligence, 2025. URL https://arxiv.org/abs/2507.20534.

[9] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

[10] PySceneDetect Contributors. Pyscenedetect: A python library for video scene detection, 2020. URL https://github.com/Breakthrough/PySceneDetect. Accessed: 2023-11-20.

[11] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023.

[12] Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10610–10620, 2025.

[13] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[14] Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering, 2024. URL https://arxiv.org/abs/2403.09622.

[15] Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention, 2025. URL https://arxiv.org/abs/2502.04507.

[16] Yuanbo Peng, Penghao Zhao, Jiangfeng Xiong, Songtao Liu Fang Yang, Jianbing Wu, Zhao Zhong, Key, Linus, Peng Chen, and Jie Jiang. flex-block-attn: an efficient block sparse attention communication library. https://github.com/Tencent-Hunyuan/flex-block-attn, 2025.

[17] Benjamin F. Spector, Simran Arora, Aaryan Singhal, Daniel Y. Fu, and Christopher Ré. Thunderkittens: Simple, fast, and adorable ai kernels, 2024. URL https://arxiv.org/abs/2410.20399.

[18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[19] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde, 2025. URL `https://arxiv.org/abs/2507.21802`.

[20] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[21] Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. In *International Conference on Machine Learning (ICML)*, 2025.