ByteDance | Seed

Seedance 1.0: Exploring the Boundaries of Video Generation Models

ByteDance Seed

Abstract

Notable breakthroughs in diffusion modeling have propelled rapid improvements in video generation, yet current foundational model still face critical challenges in simultaneously balancing prompt following, motion plausibility, and visual quality. In this report, we introduce **Seedance 1.0**, a high-performance and inference-efficient video foundation generation model that integrates several core technical improvements: (i) multi-source data curation augmented with precision and meaningful video captioning, enabling comprehensive learning across diverse scenarios; (ii) an efficient architecture design with proposed training paradigm, which allows for natively supporting multi-shot generation and jointly learning of both text-to-video and image-to-video tasks. (iii) carefully-optimized post-training approaches leveraging fine-grained supervised fine-tuning, and video-specific RLHF with multi-dimensional reward mechanisms for comprehensive performance improvements; (iv) excellent model acceleration achieving 10× inference speedup through multi-stage distillation strategies and system-level optimizations. Seedance 1.0 can generate a 5-second video at 1080p resolution only with 41.4 seconds (NVIDIA-L20). Compared to state-of-the-art video generation models, Seedance 1.0 stands out with high-quality and fast video generation having superior spatiotemporal fluidity with structural stability, precise instruction adherence in complex multi-subject contexts, native multi-shot narrative coherence with consistent subject representation.

Official Page: https://seed.bytedance.com/seedance

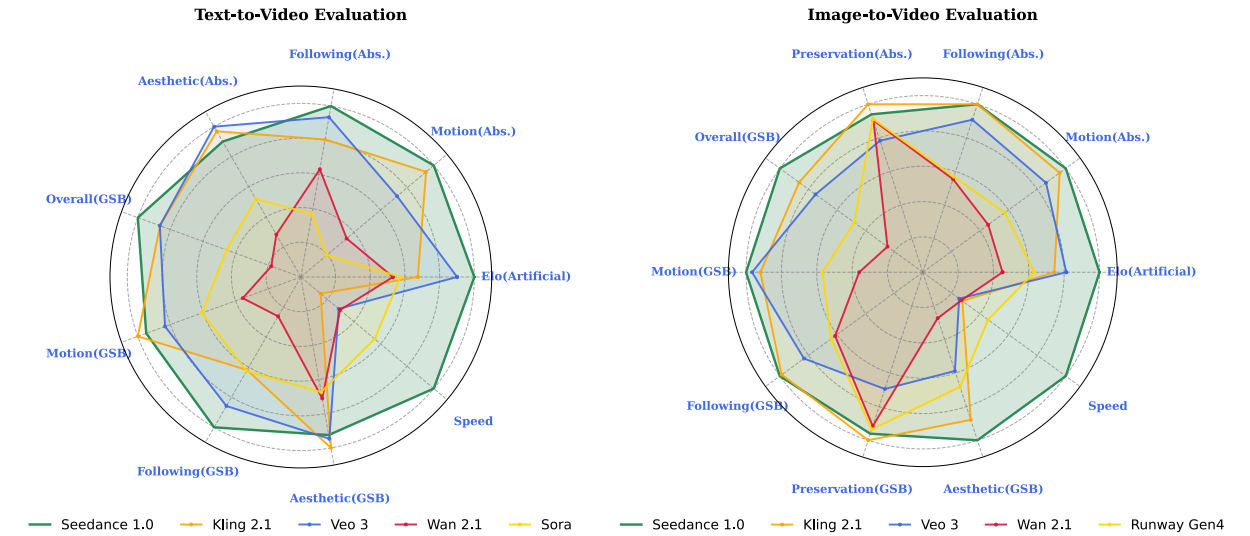


Figure 1 Overall evaluation. Left: Text-to-Video; Right: Image-to-Video. Seedance 1.0 ranks first on both the two video generation leaderboards of Artificial Analysis on Jun 10, 2025 (Due to unavailable public data, the Elo score for Kling 2.1 is taken from Kling 2.0). "Speed" denotes the inverse of the average generation time per second of video (from API).

Contents

1 Introduction

With recent advances in diffusion models, the progress of video generation has been accelerated considerably. Leading open-source frameworks including Wan [26], Huanyuan Video [15], and CogVideoX [30], complemented by commercial systems such as Veo and Keling, have catalyzed broad academic and industrial adoption. However, current video generation foundation models still have critical challenges in balancing multidimensional requirements, particularly in prompt following, motion plausibility, and visual fidelity. To address these limitations, we present **Seedance 1.0**, a foundational video generation model with native support bilingual (Chinese/English) video generation and multi-task versatility encompassing text-to-video synthesis and image-guided video generation. Seedance 1.0 integrates four key technical improvements:

- **Multi-Source Data with Comprehensive Video Captioning.** Through multi-stage, multi-perspective curation and dataset balancing, we construct a large-scale high-quality video dataset spanning diverse categories, styles, and sources. This enables a comprehensive learning of rich scenarios, topics, and action dynamics. Our precision video captioning system ensures accurate interpretation of user instructions while enabling fluent generation of complex video narratives.
- **Efficient Architecture Design.** In our design, we decouple spatial and temporal layers with an interleaved multimodal positional encoding. This allows our model to jointly learn both text-to-video and image-to-video in a single model, and natively support multi-shot video generation. In particular, the decoupled layers are integrated with carefully-designed window attentions which further improve model efficiency considerably in both training and inference.
- **Enhanced Post-Training Optimization.** We use a small set of carefully collected data for SFT, which is followed by a video-tailored RLHF algorithm (Reinforcement Learning from Human Feedback). We develop feedback-driven learning algorithms using multiple well-developed reward models, which allow us to considerably improve our performance on both T2V and I2V, in terms of motion naturalness, structural coherence, and visual fidelity.
- **Inference Acceleration.** We proposed a multi-stage distillation framework to reduce the number of function evaluations (NFE) required for generation, with inference infrastructure optimization techniques, achieving over 10× end-to-end speedup with no degradation in model performance.

Compared with contemporary models, Seedance 1.0 exhibits four distinguishing characteristics:

- **Comprehensive Generation Capabilities.** Seedance 1.0 achieves superior spatiotemporal coherence and structural stability, demonstrating exceptional motion fluidity and physical plausibility. The model produces photorealistic visuals with nuanced textures and compositional richness, attaining state-of-the-art performance across both proprietary evaluation suites and authoritative third-party benchmarks.
- **Precision Instruction Following.** Through comprehensive learning of diverse scenarios, entities, and action semantics, Seedance 1.0 precisely interprets complex user specifications. It robustly handles multi-agent interactions, adaptive camera control, and stylistic variations while maintaining narrative continuity.
- **Multi-Shot Narrative Capability.** Seedance 1.0 natively supports coherent multi-shot storytelling with stable view transitions while maintaining consistent subject representation across temporal-spatial transformations.
- **Ultra-Fast Generation Experience.** With multiple model acceleration techniques, Seedance 1.0 significantly reduces inference costs: it can generate a 5-second video at 1080p resolution only with 41.4 seconds (NVIDIA-L20), which is substantially faster than other commercial counterparts.

Seedance 1.0 will be integrated into multiple platforms in June 2025, including Doubao[1] and Jimeng[2]. We envision it becoming an essential productivity tool for enhancing professional workflows and daily creative applications.

[1] https://www.doubao.com/chat/create-video
[2] https://jimeng.jianying.com/ai-tool/video/generate

2 Model Design

2.1 Variational Autoencoder

Variational autoencoders (VAEs) [14] are widely adopted in modern large-scale image and video generation models [23] to reduce the computation of the subsequent diffusion model and facilitate efficient training and inference. Typically, a variational auto-encoder is usually composed of an encoder and a decoder; the encoder compresses the raw redundant pixel information into a compact latent representation, while the decoder reconstructs the original input from these latent features. The quality of VAE reconstruction directly establishes an upper bound for the realism and clarity achievable by the generative process, whereas the distribution of latent representations significantly impacts the convergence behavior of subsequent Diffusion Transformers (DiT).

Temporally-Causal Compression. Following MAGVIT [31], we adopt a temporally causal convolutional architecture for both the encoder and decoder, allowing joint spatial-temporal compression of images and videos within latent space. To be more specific, the model transforms the input data from the RGB pixel space with shape $(T'+1, H', W', 3)$ into a continuous latent representation with shape $(T+1, H, W, C)$, where (t, h, w, c) denotes time, height, width and channel dimensions with $r_t = \frac{T'}{T}$, $r_h = \frac{H'}{H}$, and $r_w = \frac{W'}{W}$ representing the downsample ratios along these three axes, respectively. Benefiting from the causal design, the VAE model can seamlessly process image input and output in the case of $T = T' = 0$. The overall compression ratio is given by

$$r = \frac{C \times T \times H \times W}{3 \times T' \times H' \times W'} = \frac{C}{3 \times r_t \times r_h \times r_w}. \tag{1}$$

In our practice, for the sake of training and inference efficiency and overall reconstruction and generation performance, we set $(r_t, r_h, r_w) = (4, 16, 16)$ and $C = 48$. To accommodate the higher downsampling rate and pursue better generation performance, we remove the patchification operation on the DiT side, following the strategy adopted in DCAE [3].

VAE Training. Our VAE is trained with $L1$ reconstruction loss, KL loss, LPIPS [34] perceptual loss and adversarial training loss. Adversarial training has shown to be effective in improving the quality of VAE reconstruction by enforcing finer supervision on local textures and detailed structures. Taking into account appearance and motion modeling simultaneously, we apply a hybrid discriminator with an architecture similar to that used in PatchGAN [11].

2.2 Diffusion Transformer

With the visual tokens encoded by VAE and text tokens generated by a text encoder, we employ the transformer as our diffusion backbone [20], where a fine-tuned decoder-only LLM as the text encoder. The visual tokens are then concatenated with textual tokens and fed into the transformer blocks.

Decoupled Spatial and Temporal Layers. Considering both training and inference efficiency, we build the diffusion transformer with decoupled spatial and temporal layers, where the spatial layers perform attention aggregation within each frame, while the temporal layers focus attention computation across frames. We perform window partition within each frame in the temporal layers, allowing for a global receptive field across the temporal dimension. In addition, textual tokens only participate in cross-modality interaction in spatial layers.

MMDiT Architecture. For the transformer blocks, we follow the MMDiT design in Stable Diffusion 3 [5], where a multi-modality self-attention layer is applied exclusively in spatial layers to integrate both the visual and textual tokens, whereas a self-attention layer only processes the visual tokens in temporal layers. Considering the semantic differences between visual and textual tokens, we use two separate sets of weights including adaptive layer norm, QKV projection, and MLP, for the two modalities in spatial layers. To prevent training instability, the Q and K embeddings are normalized prior to the attention matrix calculation.

Multishot MM-RoPE. In this paper, in addition to using 3D RoPE encoding for visual tokens, following Seaweed [24] and LCT [9], we add 3D Multi-modal RoPE (MM-RoPE) in the concatenated sequences by adding extra 1D positional encoding for textual tokens. The MM-RoPE also supports interleaved sequences

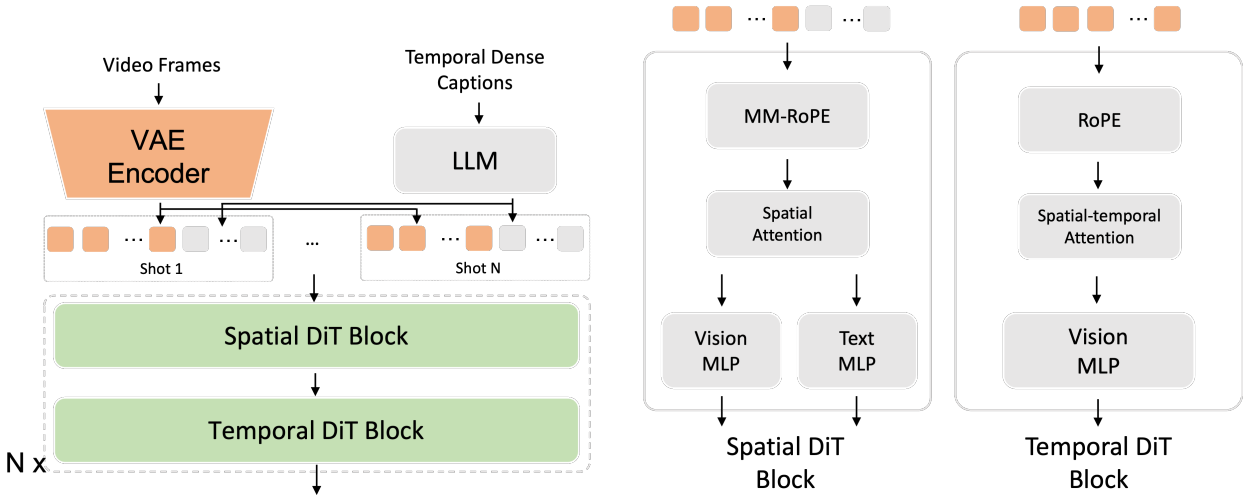


Figure 2 Our diffusion transformer architecture.

of visual tokens and textual tokens, and can be extended to training video with multiple shots, where shots are organized in the temporal order of actions and each shot has its own detailed caption.

Unified Task Formulation. To enable conditional video generation, we concatenate the noisy inputs with cleaned or zero-padded frames along the channel dimension, and use binary masks to indicate which frames are instructions to follow [7]. With this formulation, we can further unify different generation tasks such as text-to-image, text-to-video and image-to-video [4]. During the training process, we mix these tasks and adjust the proportion by controlling the conditional inputs.

2.3 Diffusion Refiner

Take into account the training and inference efficiency, we employ a cascaded diffusion framework for high-resolution (HR) video generation. The base model generates 480p videos first, which are then upscaled to 720p or 1080p high-resolution videos through a learned diffusion refiner model to enhance visual details and textures.

Refiner Model Training. To facilitate training, the diffusion refiner model is initialized from the pre-trained base model. Different from the base model, the diffusion refiner model is trained with conditioning on the low-resolution (LR) videos. Specifically, the LR video is upsampled to a high resolution first, then concatenated with the diffusion noise along the channel dimension to form the input of the diffusion transformer.

2.4 Prompt Engineering (PE)

As described in Sec 3.2, texts used in DiT are form of dense video captions. Therefore, we need to employ a large language model to convert the user prompts into corresponding caption format. To achieve this, we initialize based on Qwen2.5-14B [29] and employ two stages to implement high-quality Prompt Engineering (PE): Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL).

Supervised Fine-Tuning. In the SFT stage, we synthesize large amount of user prompts and their dense caption expression by manual annotation. We specially devide the image-to-video (i2v) and text-to-video (t2v) tasks, as they are different in user prompt styles. We then adpot a fully fine-tuning strategy to train the model on the annotated data to aquire basic rephrasing abilitity.

Reinforcement Learning. However, due to the presence of model hallucinations, the results of the first SFT stage cannot guarantee that the semantics of the rewritten results fully meet the requirements of the user prompts. Therefore, we carefully collect a dataset of pairs with correct and incorrect rephrased results to

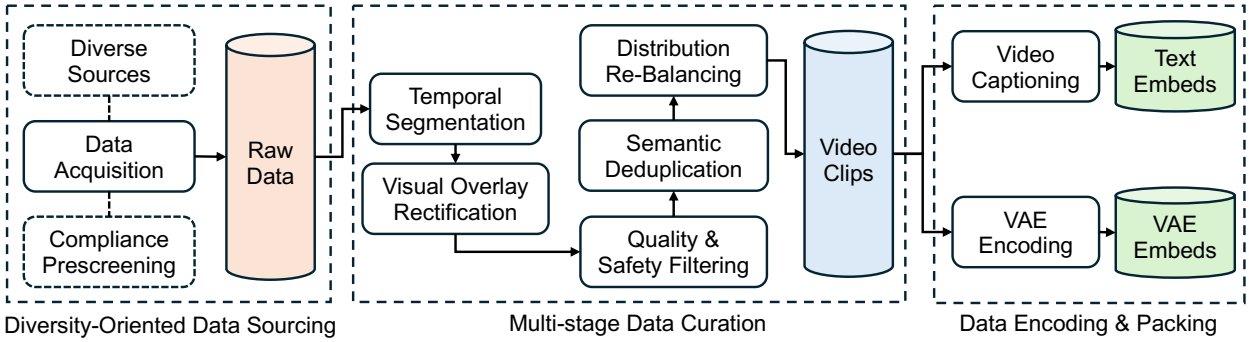


Figure 3 Our video data processing pipeline, transforming heterogeneous raw videos into a refined, feature-rich training dataset. The workflow comprises three main phases: (1) Diversity-oriented data sourcing for initial acquisition and compliance prescreening from diverse sources; (2) Multi-stage data curation refines raw data into video clips; and (3) Offline data packing where video captioning and VAE encoding are used to generate text and VAE embeddings for model training.

perform the Direct Preference Optimization (DPO) [13, 21] training. In this stage, we used the Low-Rank Adaptation (LoRA) [10] fine-tuning strategy on the SFT model.

After the above stages, our prompt engineering model has strong ability to understand user prompts and gives precise and high-quality rephrased results in video caption format, consistent with DiT training.

3 Data

The performance of video generation models is inextricably linked to the scale, diversity, and quality of the training data. While our broader training corpus incorporates both video and image datasets, with image data preparation following methodologies similar to Seedream [8], this section specifically details our rigorous approach to curating video data. We develop a systematic data processing workflow, illustrated in figure 3, to transform vast, heterogeneous raw video collections into a refined, high-quality, diverse, and safe dataset for training robust video generation models. This workflow is deployed as a robust, automated system optimized for high-throughput processing of massive data volumes.

3.1 Data Pre-Processing

At the heart of our video data curation is a multi-stage pre-processing pipeline, designed to tackle the challenges of raw video collections. Each subsequent stage systematically elevates the dataset's standard, preparing it for robust model training. The following paragraphs detail each component of this comprehensive pipeline, which ensures that only video clips meeting our stringent criteria contribute to the final dataset.

Diversity-Oriented Data Sourcing. Our video data acquisition strategy prioritizes ethically and legally sourced content from diverse public and licensed repositories. We aim to maximize coverage across critical dimensions, including clip duration, resolution, subject matter (e.g., humans, animals, objects), scene types (e.g., natural landscapes, urban environments), subject actions, genres (e.g., documentary, animation), artistic styles, camera kinematics, and cinematographic techniques. Raw video collections exhibit significant heterogeneity and often contain undesirable elements, posing key challenges that our pipeline is designed to address.

Shot-Aware Temporal Segmentation. Raw long-form videos are not suitable for direct model training. We employ automated shot boundary detection techniques by analyzing inter-frame visual dissimilarities or utilizing pre-trained detectors to identify natural scene transitions. Subsequently, videos are segmented into shorter clips, with a maximum duration of 12 seconds. Each resulting clip may contain one or multiple temporally coherent shots, preserving local narrative flow while ensuring manageable input lengths for model ingestion.

Visual Overlay Rectification. Many source videos contain extraneous visual overlays such as logos, watermarks,

subtitles, or on-screen graphics that can introduce noise or bias. Our rectification stage identifies these occlusions using a hybrid approach of heuristic rule-based systems and specialized object detection models. Frames are then adaptively cropped to maximize the retention of the primary visual content, yielding cleaner and more focused video data.

Quality and Safety Filtering. To ensure the model is trained on high-quality and ethically compliant data, we enforce rigorous filtering via visual assessment and safety screening. First, clips exhibiting visual defects such as blurriness, excessive jittering, low aesthetic quality, poor cinematographic composition, or predominantly static content are systematically identified and removed by our specialized visual quality model. Second, we rigorously exclude harmful or inappropriate material, deploying advanced classifiers to detect content pertaining to pornography, explicit violence, child exploitation, and explicit nudity, thereby ensuring ethical compliance and dataset safety.

Semantic Deduplication. To promote dataset diversity and prevent model overfitting to redundant content, we perform semantic deduplication. Video clips are represented by robust feature embeddings extracted from an internally developed video representation model, and these embeddings enable clustering of visually and semantically similar clips. Within each identified cluster of near-duplicates, only the single instance with the highest overall quality score (from the preceding quality filtering stage) is retained.

Distribution Rebalancing. Raw data often exhibits significant category imbalance across various attributes. We analyze the dataset's distribution along these dimensions by quantifying frequencies across attributes tailored to different semantic and technical perspectives, such as subject categories, scene types, dominant actions, genres, visual styles, clip duration, resolution, and motion characteristics. For over-represented head categories, downsampling is applied. Conversely, for under-represented tail categories, we increase their sampling probability during training and initiate targeted data acquisition to augment their presence, aiming for a more equitable and comprehensive representation of the visual world.

3.2 Video Captioning

Video captions largely affect the instruction-following capabilities of the video generation model. We mainly improve the quality and accuracy of captions to ensure that important content and actions can be seen and described proprerly.

Caption Style. We adopt a dense caption style integrating dynamic and static features. For dynamic features, we meticulously describe actions and camera movements of a video clip, highlighting changing elements. For static features, we elaborate on the characteristics of core characters or scenes in the video.

Caption Elements. We define specific categories dynamic and static features respectively. dynamic features cover categories of motions, subjects or scenes changing and camera movements, while static features include appearances, aesthetics, styles, etc. We collect diverse data on such categories and conduct high-quality manual annotations for training. The trained caption model can accurately describe the critical content of complex and abstract video materials.

Model Training. We train the caption model on the annotated data with Tarsier2 [32], a model with strong video understanding capabilities. The visual encoder is frozen and the language model is fully fine-tuned. We train on both Chinese and English data to acquire bilingual capabilities.

During inference, we use our PE model described in Sec 2.4 to rephrase user prompts into detail video captions, in which the format is aligned with the training captions in content and structure.

3.3 Efficient Engineering Infrastructure

Engineering Infrastructure Overview. Our engineering infrastructure for data processing is illustrated in figure 4, which consists of three layers: at the top is the unified platform layer, automating human-in-the-loop workflows, managing tasks, visualizing data, and monitoring pipelines, etc.; in the middle is the computation framework layer, which employs BMF [2] and Ray [19] for heterogeneous computing across CPU/GPU/NPU architectures and optimizes resource allocation for both stable and elastic computing; at the bottom is the underlying resources layer, which leverages cloud infrastructure from ByteCloud (internal) and Volcengine (external).

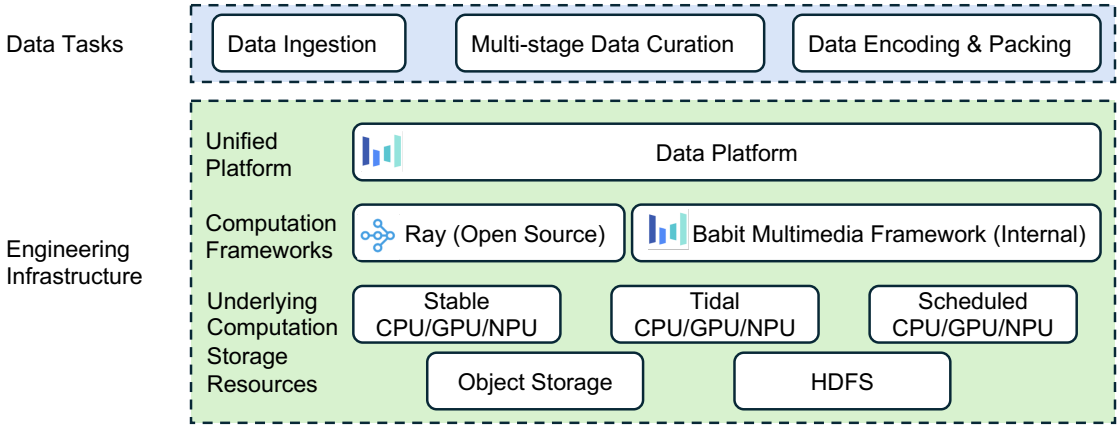


Figure 4 Overview of our engineering infrastructure for data processing.

Efficient Heterogeneous Computing. To maximize resource utilization, our frameworks dynamically allocate video operations to optimal hardware (e.g., CPU for decoding, GPU for deep model inference). Asynchronous communication between computation units is used to mitigate bottlenecks introduced by the performance gap between different types of computation hardware. To address the complexities arising from the instability of elastic computation resources, our frameworks incorporate two critical capabilities: adaptive auto-scaling to handle resource fluctuations and failure retry mechanisms for preempted tasks. Customized versions of BMF and Ray implement these optimizations, delivering near-linear scalability and extremely high throughput to efficiently process massive-scale video training data.

4 Model Training

As shown in Figure 5, we present our training and inference stages of Seedance 1.0. Our training process is divided into several substages, including pre-training, continue training (CT), supervised fine-tuning (SFT) and human feedback alignment (RLHF). Our refiner also includes pre-training, SFT and RLHF. The visualization results during different training stages are presented in Figure 6, where each stage can progressively improve the results.

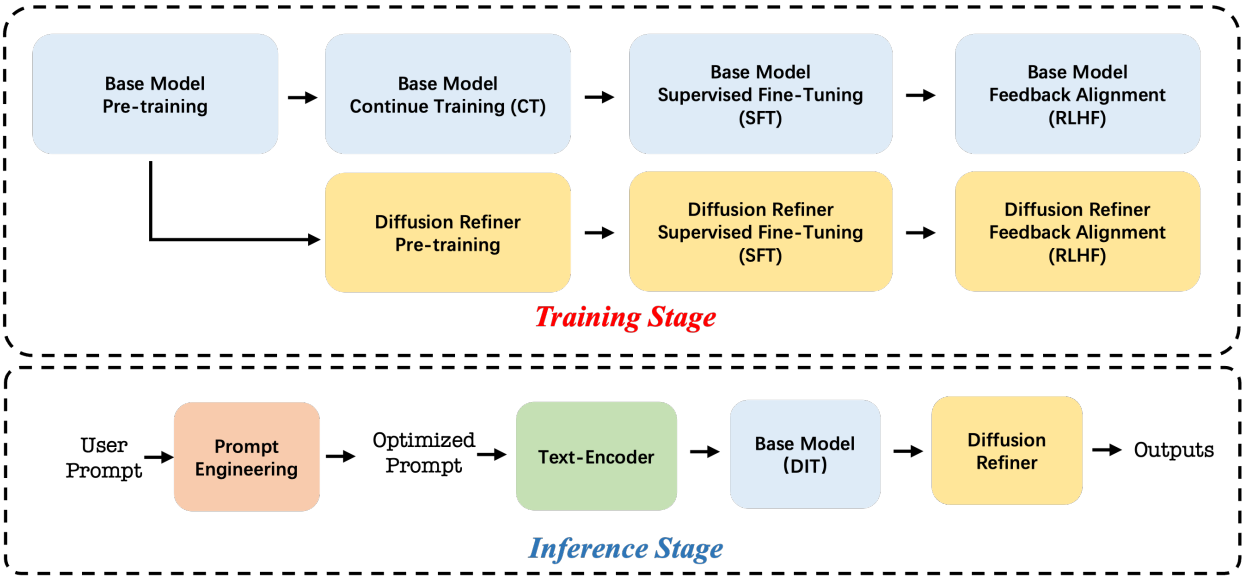


Figure 5 Overview of training and inference pipeline.

4.1 Pre-Training

Diffusion Scheduling. During training, we employ the flow matching framework with velocity prediction, and a training timestep is sampled from a logit-normal distribution. Considering that videos with higher resolution and longer duration require more noise to disrupt their signal, we then transform the training timestep with a resolution-aware shift, which increases the noise perturbation for videos with higher resolution and longer duration.

Progressive Training. To enable higher data throughput and training efficiency, we initialize the model with sufficient low-resolution text-to-image ($256px$) training and then progressively introduce video modalities with higher resolution and higher fps in following stages: (1) We conduct image-video joint training using $256px$ images and video clips from 3 to 12 seconds (12 fps). (2) In the second stage, we increase the training resolution to $640px$ while maintaining the same duration. (3) In the final stage, we train the models with 24 fps video to further improve the video smoothness. During video pre-training, we also retain a small portion of text-to-image task to maintain semantic alignment and set the proportion of the image-to-video task to 20% to activate the ability to follow visual prompts.

4.2 Continue Training (CT)

As the image-to-video task constitutes only a small fraction of pre-training, the model's potential in this area remains underexplored. To address this, we introduce the Continue Training (CT) phase focused on strengthening image-to-video generation after pre-training. In this phase, we increase the image-to-video ratio from 20% to 40% and further refine the training dataset to improve overall multitask performance.

High-Quality Data Selection. We select a subset of the pre-training data with higher aesthetic quality and richer motion dynamics by using a series of specialized evaluation models, including aesthetic scorer and motion evaluators based on optical flow. Since the first frame is always provided in the image-to-video task, we design two types of caption for training: (1) original long captions with detailed descriptions of both dynamic and static content, and (2) short captions that focus solely on motion dynamics by removing the static description corresponding to the first frame. This encourages stronger semantic alignment with the training objective.

Training Strategy. During continued training, we use slightly fewer GPUs than in the pre-training stage, while maintaining an annealed learning rate schedule. The richer motion dynamics and diverse captions enable the model to generate more natural and smoother videos. Furthermore, the higher aesthetic quality of the training data leads to significant improvements in the visual fidelity of text-to-video generation. As a result, the final model supports both text-to-video and image-to-video tasks with enhanced overall performance.

4.3 Supervised Fine-Tuning (SFT)

Following CT, we perform supervised fine-tuning (SFT) to further align the model's output with human preferences regarding visual quality and motion coherence. During this phase, the model trains on a carefully curated set of high-quality video-text pairs with manually verified captions, allowing it to generate videos with improved aesthetics and more consistent motion dynamics.

Human-Curated Dataset. Ensuring data quality and distributional balance is essential. To achieve this, we define several hundred categories based on visual style, motion type, and other key attributes. We then collect data in a targeted manner within each category, resulting in a curated dataset of high-quality video samples with accurate and meaningful captions.

Model Merging. To fully leverage high-quality data, we train separate models on curated subsets designed to capture a wide range of styles, motions, and scenarios. The resulting models are subsequently merged into a single model that integrates their respective strengths. Each model is trained with a smaller learning rate than in pre-training and utilizes a limited number of GPUs. Moreover, we apply early stopping at an effective point to prevent overfitting and maintain text controllability. The final merging step significantly improves both visual fidelity and motion quality.

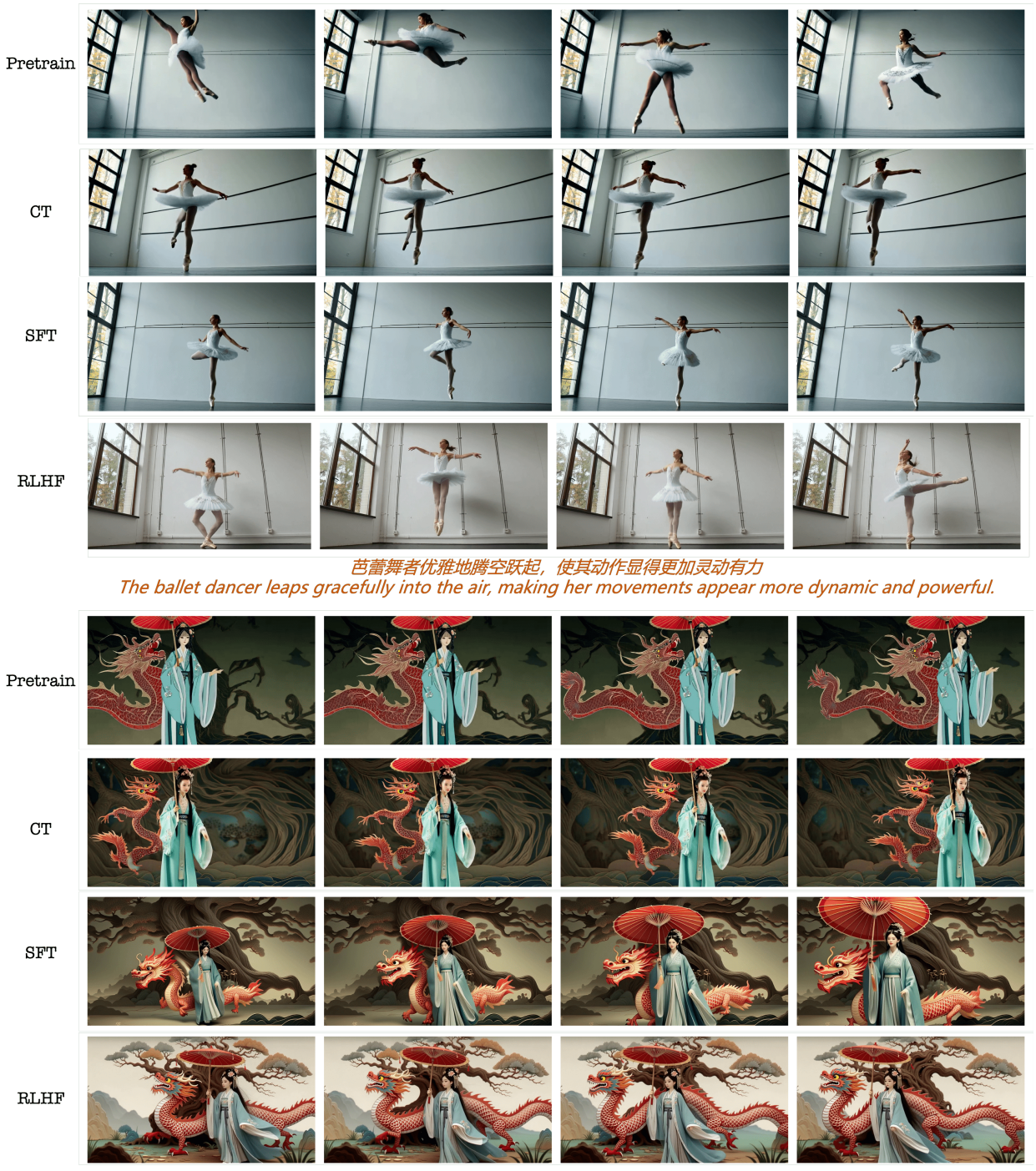


Figure 6 Visualization during different post-training stages.

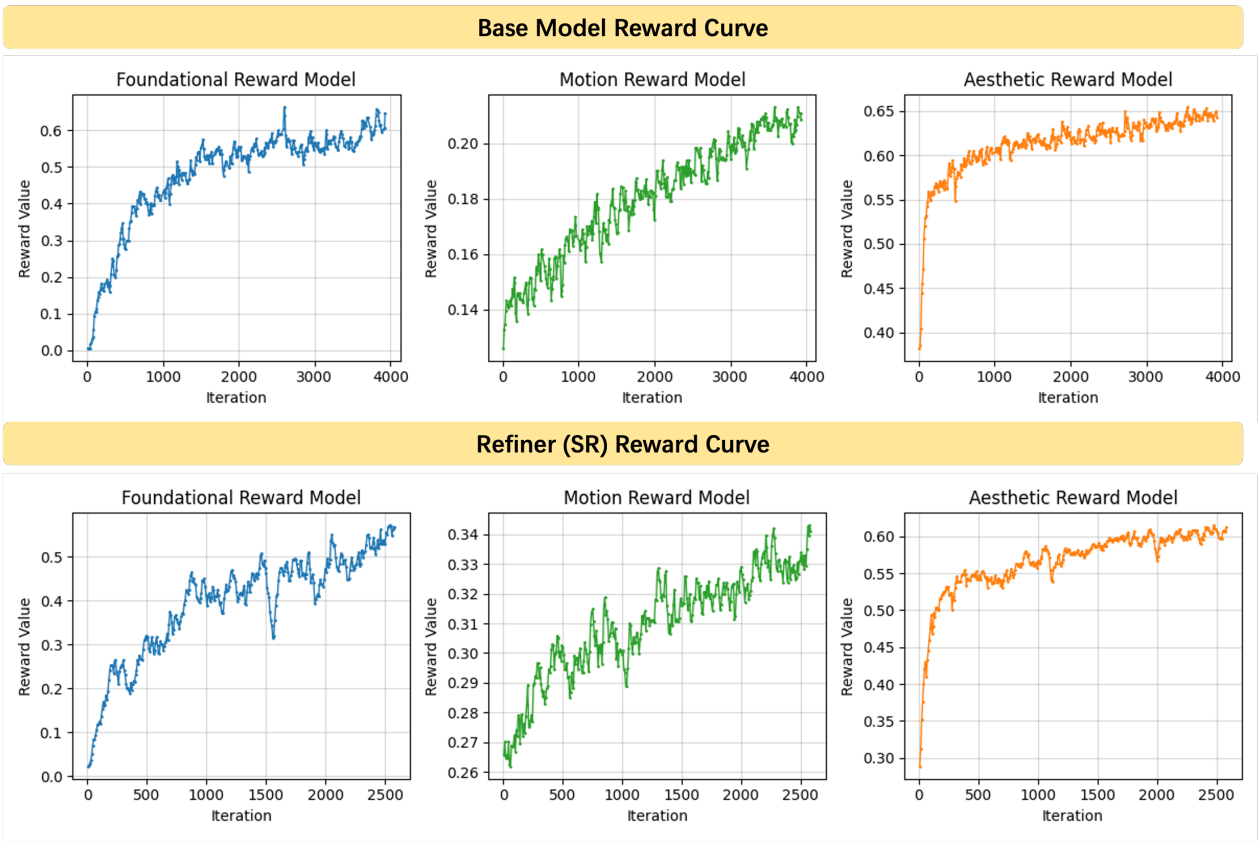


Figure 7 The reward curves show that the values across diverse reward models all exhibit a stable and consistent upward trend during the base model and Refiner RLHF process.

4.4 Human Feedback Alignment (RLHF)

4.4.1 Feedback Data Infrastructure

We collect prompts from training datasets and online users, and perform data balancing and information filtering on prompts to discard duplicate and ambiguous ones. We collect high-quality video data pairs for human preference labeling, including synthetic videos generated by different stages of our model. Experimental results demonstrate that the incorporation of multiple source visual materials can further enhance the domain capacity of the RM model, expand the preference upper bound of RM, and strengthen generalization capabilities. We adopt a multi-dimensional annotation approach in the labeling process, i.e., selecting the best and worst videos under a specific labeling dimension while ensuring that the best videos are not inferior to the worst ones in other dimensions.

4.4.2 Reward Model

To comprehensively enhance model performance, we design a sophisticated reward system comprising three specialized reward models: Foundational Reward Model, Motion Reward Model, and Aesthetic Reward Model. These dimension-specific reward models, coupled with video-tailored RLHF optimization strategies, enable comprehensive improvements in multiple aspects of the model capabilities, as illustrated in Figure 7. Foundational reward model focuses on enhancing fundamental model capabilities, such as image-text alignment and structural stability. We employ a Vision-Language Model as the architecture of this reward model. Motion reward model helps to mitigate video artifacts while enhancing motion amplitude and vividness. Given that video aesthetics primarily derive from keyframes, we design the aesthetic reward model from image-space input inspired by Seedream [6, 8], with the data source modified to use keyframes from videos.

4.4.3 Base Model Feedback Learning

Reward feedback learning [17, 18, 28, 33] have been widely used in currnet diffusion models. In Seedance 1.0, we simulate the video inference pipeline during training, directly predict x_0 (generated clean video) when the Reward Model (RM) adequately assesses video quality. The optimization strategy directly maximizes the composite rewards from multiple RMs. Comparative experiments against DPO/PPO/GRPO demonstrate that our reward maximization approach is the most efficient and effective approach, comprehensively improving text-video alignment, motion quality, and aesthetics. Furthermore, we preform multi-round iterative learning between the diffusion model and RMs. This approach raises the performance bound of the RLHF process and is more stable and controllable than dynamic update of the RM.

4.4.4 Super-Resolution RLHF Framework

As shown in Figure 8, we also apply RLHF on our diffusion refiner, which can be regarded as a diffusion-based conditional generative model. During training, low-resolution VAE latent space representations serve as conditional inputs to the super-resolution model, while the generated high-resolution videos are evaluated by multiple Reward Models. We directly maximize a linear combination of these reward signals. Notably, our approach applies RLHF directly to the accelerated refiner model, effectively enhancing motion quality and visual fidelity in low-NFE scenarios while maintaining computational efficiency.

5 Inference Optimizations

5.1 Model Acceleration

DiT Optimizations. To accelerate DiT inference, we adopt diffusion distillation techniques to reduce the number of function evaluations (NFE) required for generation. We incorporate the Trajectory Segmented Consistency Distillation (TSCD) technique, originally introduced in HyperSD[22], which partitions the denoising trajectory into multiple segments and enforces consistency between predicted and target states across these segments. This allows the student model to learn an accurate approximation of the diffusion process with fewer steps. Using TSCD, our DiT model performs competitively with 4x acceleration, offering a strong balance between speed and fidelity. To push acceleration further, we incorporate Score Distillation from RayFlow[25], which aligns the student model's predicted noise (i.e., score function) with that of the teacher using expected noise consistency. This approach supports trajectory-level optimization for each sample, enabling more stable and adaptive sampling even at low NFEs. It effectively improves generalization and reduces artifacts during fast generation. To improve visual quality, we extend the adversarial training strategy from APT[16] to a multi-step distillation setting, incorporating human preference data for supervision. A learned discriminator guides the student model toward outputs favored by human judgments, effectively mitigating artifacts from aggressive acceleration and enhancing perceptual realism.

Through the proposed distillation pipeline, our final distilled model achieves comparable results to the original model across four expert-evaluated dimensions: prompt alignment, motion quality, visual fidelity, and consistency with the source image.

VAE Optimizations. In video generation tasks, the decoding process from latent space to pixel space incurs significant computational cost. We profiled the VAE decoder and found that stages closer to the pixel space dominate the latency. By narrowing the channel widths in these stages, we design a thin VAE decoder. Retraining it with a fixed pre-trained encoder, we achieve a 2× speedup with no loss in visual quality of the end-to-end video generation.

5.2 Inference Infrastructure

High-Performance Kernel. Extensive kernel fusion efforts have been conducted on the model's core modules, resulting in a cumulative 15% improvement in the model's inference throughput.

Quantization and Sparse. Building on the Seedream [8] technical solution, we have implemented fine-grained mixed-precision quantization tailored for Attention and Gemm operations. Moreover, our exploration revealed that the sparse attributes of DiTs exhibit hierarchical and blockified structures across and within various

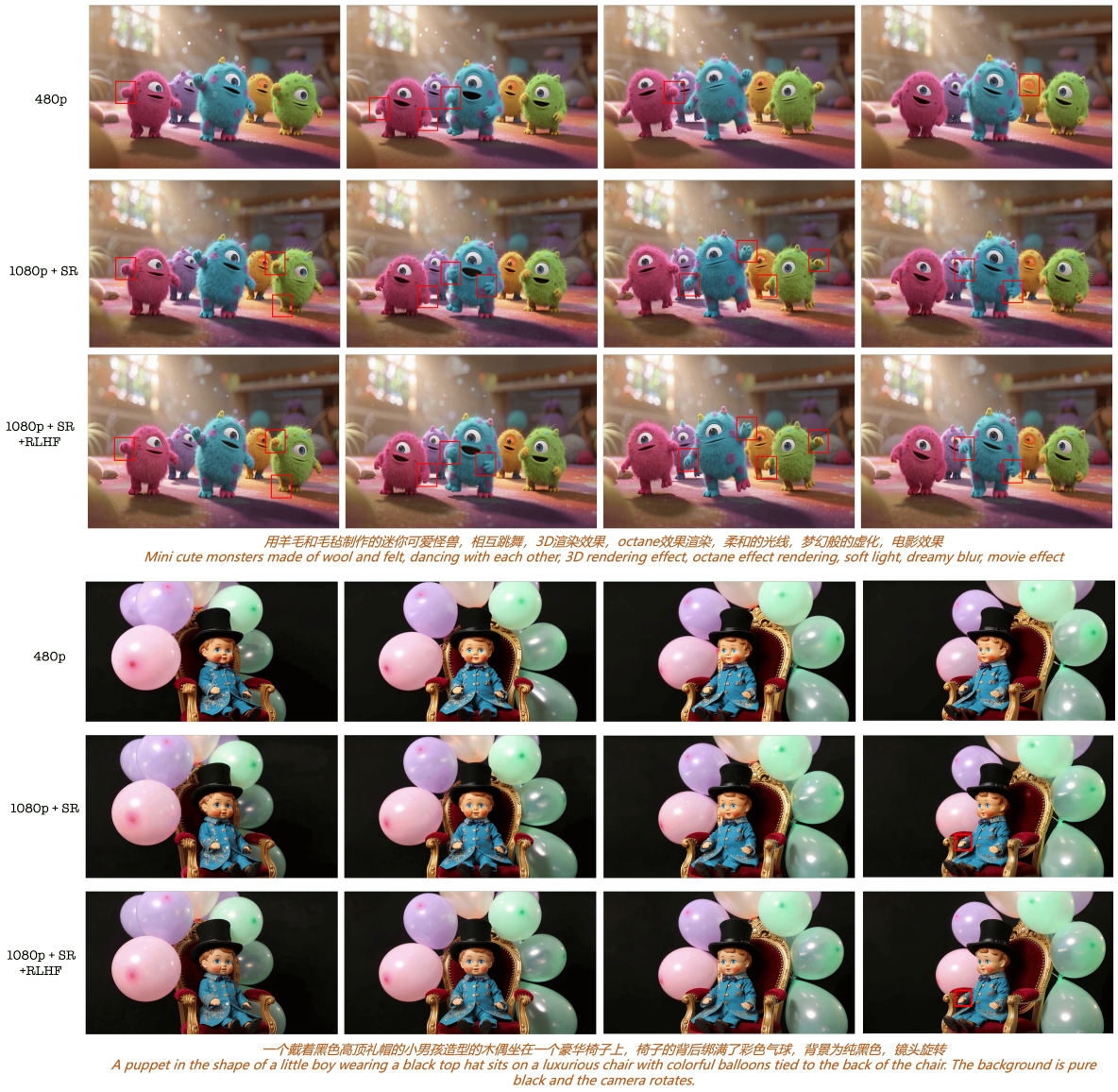


Figure 8 Visualization during different resolutions and RLHF process.

modalities. Expanding on the methodology established by AdaSpa [27], we have introduced a streamlined tuning solution focused on minimizing search stage overhead. Additionally, we have successfully integrated our optimized fine-grained Attention Quantization approach into this scheme. Numerous efforts have been dedicated to mitigating the impact of full quantization and sparsity on the quality of pixel-level generation. We have achieved an optimal balance between performance and efficiency.

Parallelism Strategy. In order to decrease the allocated massive memory due to the long sequence in video generation schema. A customized adaptive hybrid parallel strategy has been proposed to effectively split the sequences. This approach integrates the concept of context parallelism to optimize communication processes, resulting in a reduction of communication overhead to a quarter of the level observed in Ulysses [12]. Simultaneously, we have further reduced end-to-end communication overhead by introducing FP8 communication.

Async Offloading Strategy. Due to the extensive computational demands of attention coupled with the large model size. We developed an automated and adaptive AsyncOffloading strategy. We successfully solved the large model deployment problem on memory-limited devices with a performance drop of less than 2%.

Hybrid Parallelism for Distributed VAE. Moreover, to address the issue of high GPU memory consumption due to the VAE-Decoder, we implemented an adaptive hybrid parallel strategy. This method partitions the input data along the spatial and temporal dimensions simultaneously and employs efficient collective communication for Conv3D computation. Thus, we further improved parallel scaling performance.

Pipeline Optimizations. We adopted kernel fusion, quantization, parallelization, continuous batching, prefix caching, and other common techniques to improve the overall throughput of the prompt engineering effectively. Furthermore, to tackle the issue of low encoding efficiency in long videos, we have implemented video encoding acceleration solutions.

These innovations have effectively boosted the E2E efficiency of the whole inference pipeline.

6 Training Infrastructure

6.1 Pre-Training Optimization

To support efficient large-scale pre-training of long-context video models on thousands of GPUs, we have designed a highly optimized training infrastructure. Our system focuses on maximizing hardware efficiency, scalability, and robustness. It integrates high-performance kernel fusion, a hybrid parallelism strategy, multi-level activation checkpointing (MLAC), runtime-aware workload balancing, and multi-level fault tolerance. These components work together to ensure stable, high-throughput training under diverse workloads and hardware scales.

High-Performance Kernel. To fully utilize GPU hardware resources, we combined `torch.compile` with handcrafted CUDA kernels for performance-critical operators. We identified memory-bound operations and fuse them into single CUDA kernels to minimize redundant memory access, such as rotary position encoding (RoPE) and normalization. These fused kernels store intermediate results in registers or shared memory, significantly improving arithmetic intensity and reducing global memory traffic by over 90%.

Parallelism Strategy. We adopted a hybrid parallelism strategy combining data parallelism and sequence parallelism to efficiently train long-context models on thousands of GPUs. Specifically, we employed Hybrid Sharded Data Parallelism (HSDP) [35] for memory-efficient weight sharding and mitigating performance degradation observed when scaling to over thousands of GPUs. For sequence parallelism, we followed the Ulysses [12] approach, sharding tokens across GPUs along the sequence and head dimensions to enable parallel processing of long video samples.

Multi-Levels Activation Checkpointing. Multi-Level Activation Checkpointing (MLAC) [24] policy is employed to reduce GPU memory pressure under negligible recomputation overhead during backpropagation. MLAC implements optimized asynchronous caching and prefetching mechanisms to maximize the overlap between memory transfers and forward/backward computation. We leveraged MLAC to prioritize offloading output tensors of the operators (ops) with the highest recomputation cost during model training, e.g., attention and FC2 layer in MLP module. Furthermore, MLAC was applied to offload input tensors of the activation checkpointing module to attain zero activation occupancy in GPU memory, which allows us to lower the degree of sequence parallelism and thereby reduce communication overhead.

Workload Balance. Large-scale video pre-training often involves heterogeneous data types (e.g., long vs. short videos, varying resolution), which introduces significant computational imbalance across GPUs. To address this, we applied a runtime-aware workload balancing strategy [24], leveraging an additional all-to-all communication step to distribute workload evenly across GPUs. This balancing strategy is performed within each batch to maintain data consistency, and is asynchronously precomputed in the background to avoid stalling the main training loop. Our approach significantly reduced inter-GPU idle time and improves overall training throughput.

Fault Tolerance. In large-scale training jobs running on thousands of GPUs over extended periods, transient failures are inevitable. To ensure robustness, we integrated fault tolerance at multiple levels. First, we implemented periodic checkpointing of both model and optimizer states, with full support for FSDP-sharded weights. The states of the dataloader were also saved to ensure bitwise-exact resumption. Second, we

conducted thorough machine health checks before launching each job to eliminate potential stragglers and faulty nodes. Third, we reduced model initialization overhead to maximize effective training time. For example, we utilized PyTorch's meta tensor initialization to directly load model parameters, eliminating the time typically spent on standard initialization. Combined, these strategies enhanced training reliability and minimize the impact of hardware or software failures during prolonged distributed runs.

6.2 Post-Training Optimization

Post-training primarily consists of three phases: supervised fine-tuning, reinforcement learning, and distillation. During this stage, it is essential not only to optimize training efficiency but also to minimize GPU memory consumption (e.g., reducing peak memory usage and fragmentation) and enhance overall memory utilization. The suboptimal GPU memory utilization observed in the post-training stage primarily stems from three factors:

- **Memory Contention**. During the reinforcement learning and distillation phases, GPU memory is sequentially and dynamically shared among various components, including the Text Encoder, DiT, VAE, reward models, and their corresponding activation tensors.
- **Complex Training Modes**. The coexistence of trainable and frozen model components complicates memory management and introduces additional optimization challenges.
- **Diverse Workloads**. The concurrent presence of both long and short video sequences creates variable memory demands, making traditional static memory optimization methods ineffective.

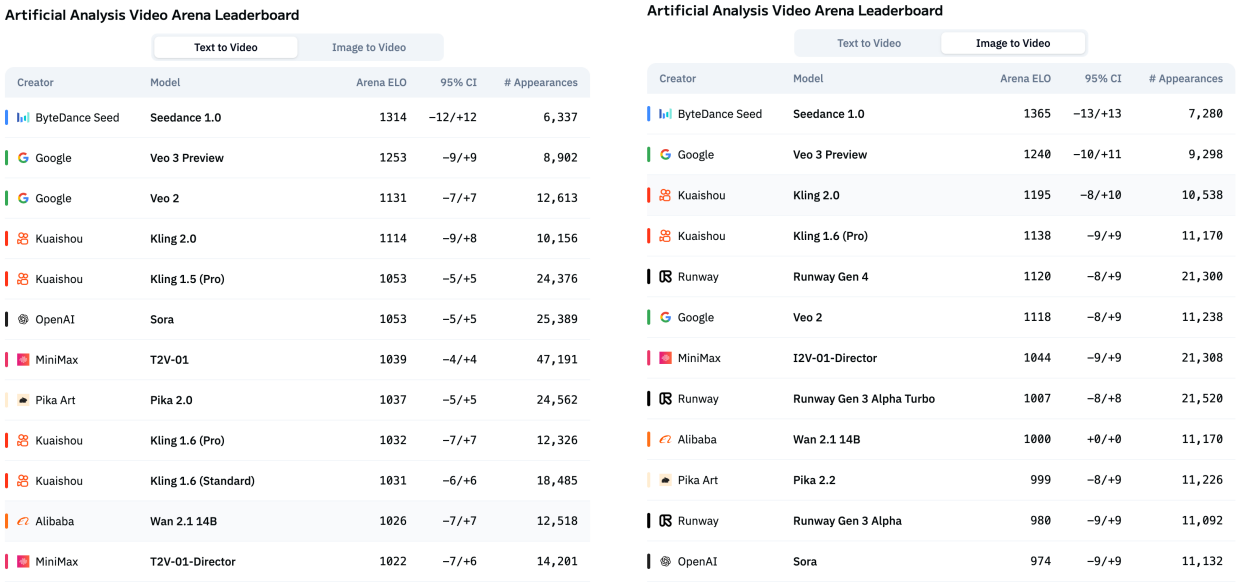
To effectively address these challenges, we have developed a dynamic memory management framework that incorporates CPU offloading and recomputation techniques. Additionally, we adopted the parallelization strategies previously used during pre-training, leveraging FSDP and sequence parallelism to enable efficient multi-node scaling.

- **Memory Optimization**. To ensure simplicity and ease of use, we utilized PyTorch hooks to implement CPU offloading, thereby minimizing intrusive modifications to user code. Through detailed profiling and modeling, we identified optimal CPU offloading and recomputation strategies. In addition, we applied localized static memory planning to mitigate memory fragmentation caused by frequent allocation and free of tensors with varying sizes.
- **Parallelism Strategy**. To maximize hardware utilization, we configured different degrees of sequence parallelism across different models based on their computational characteristics. Additionally, we set `TORCH_NCCL_AVOID_RECORD_STREAMS=1` to eliminate delayed memory release issues. Additionally, we manually managed the `free_event_queue` to address the problem of delayed parameter release in FSDP when parameters are frozen. Furthermore, we utilized `register_post_backward_reshard_only_hook` to adjust the order of memory allocation and release during backward computation under the frozen mode.

These optimizations ensure stable and efficient post-training performance, even in complex scenarios involving multiple model components and diverse video workloads.

7 Model Performance

This section provides a comprehensive evaluation of Seedance 1.0, structured as follows. In Section 7.1, we first present results from an external public evaluation platform, where Seedance 1.0 tops the leaderboards in both text-to-video and image-to-video. Section 7.2 details the internal evaluation, covering benchmark design, absolute scoring, and comparative analysis using the Good-Same-Bad (GSB) metric. The subsequent subsections highlight Seedance 1.0's strengths in multi-shot transitions and multi-style generation. The overall results are presented in Figure 1.



Artificial Analysis Video Arena Leaderboard

Text to Video | Image to Video

Creator	Model	Arena ELO	95% CI	# Appearances
ByteDance Seed	Seedance 1.0	1314	−12/+12	6,337
Google	Veo 3 Preview	1253	−9/+9	8,902
Google	Veo 2	1131	−7/+7	12,613
Kuaishou	Kling 2.0	1114	−9/+8	10,156
Kuaishou	Kling 1.5 (Pro)	1053	−5/+5	24,376
OpenAI	Sora	1053	−5/+5	25,389
MiniMax	T2V-01	1039	−4/+4	47,191
Pika Art	Pika 2.0	1037	−5/+5	24,562
Kuaishou	Kling 1.6 (Pro)	1032	−7/+7	12,326
Kuaishou	Kling 1.6 (Standard)	1031	−6/+6	18,485
Alibaba	Wan 2.1 14B	1026	−7/+7	12,518
MiniMax	T2V-01-Director	1022	−7/+6	14,201

Artificial Analysis Video Arena Leaderboard

Text to Video | Image to Video

Creator	Model	Arena ELO	95% CI	# Appearances
ByteDance Seed	Seedance 1.0	1365	−13/+13	7,280
Google	Veo 3 Preview	1240	−10/+11	9,298
Kuaishou	Kling 2.0	1195	−8/+10	10,538
Kuaishou	Kling 1.6 (Pro)	1138	−9/+9	11,170
Runway	Runway Gen 4	1120	−8/+9	21,300
Google	Veo 2	1118	−8/+9	11,238
MiniMax	I2V-01-Director	1044	−9/+9	21,308
Runway	Runway Gen 3 Alpha Turbo	1007	−8/+8	21,520
Alibaba	Wan 2.1 14B	1000	+0/+0	11,170
Pika Art	Pika 2.2	999	−8/+9	11,226
Runway	Runway Gen 3 Alpha	980	−9/+9	11,092
OpenAI	Sora	974	−9/+9	11,132

Figure 9 Results from Artificial Analysis Arena. Seedance 1.0 achieves the top position on both the text-to-video and image-to-video leaderboards.

7.1 Artificial Analysis Arena

Artificial Analysis [1] has emerged as a widely recognized and trusted benchmarking platform, particularly in the domains of image and video generation. It offers an open arena in which various generative models are evaluated and scored by the public. Leveraging a large corpus of comparison results, the platform calculates Elo scores to reflect user preferences across different models. The Artificial Analysis Video Arena Leaderboard comprises two distinct tracks: text-to-video and image-to-video. Seedance 1.0 has participated in both categories. Some notable external competitors include Veo 3, Kling 2.0, Runway Gen4, OpenAI Sora, and Wan 2.1.

Seedance 1.0 tops both the text-to-video and image-to-video leaderboards, demonstrating a substantial performance advantage over competing models. In particular, it outperforms the second- and third-best models, Veo 3 and Kling 2.0, by over 100 points in the image-to-video task. Notably, Seedance 1.0 attains state-of-the-art results across both tasks using a single unified model, whereas prior models typically excelled in one domain while underperforming in the other. The subsequent sections provide a detailed analysis of Seedance 1.0's advantages in each scenario.

7.2 Comprehensive Evaluation

Besides overall user preferences, a comprehensive benchmark is equally important for the evaluation of visual generation models, as it enables a more holistic assessment of model capabilities. We developed SeedVideoBench-1.0, a comprehensive benchmark for video generation, comprising 300 prompts each for T2V and I2V. We then collaborated with film director experts to co-develop evaluation criteria and conducted a detailed manual expert evaluation.

7.2.1 SeedVideoBench 1.0

To comprehensively evaluate video generation models across diverse scenarios, we proposed SeedVideoBench-1.0, a benchmark designed through systematic analysis of real-world user prompts. This benchmark encompasses a wide range of application scenarios, including special effects, e-commerce, and professional-generated content (PGC). Additionally, a detailed taxonomy has been developed to assess model capabilities. The following section demonstrates the classification of main label categories, using text-to-video as an example.

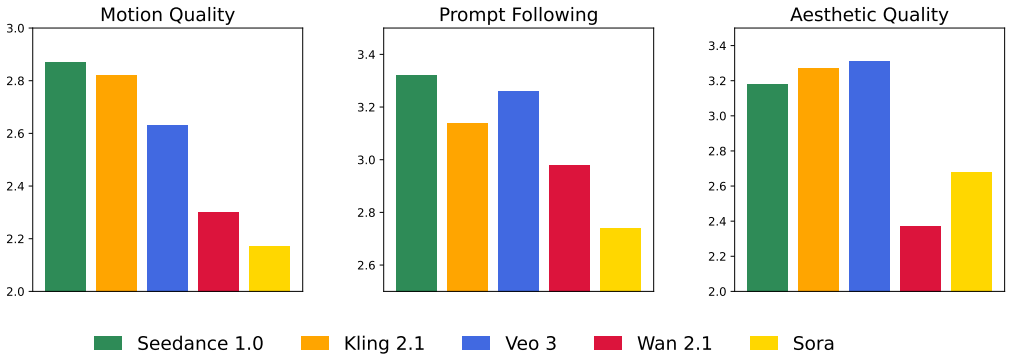


Figure 10 Absolute Evaluation for Text-to-Video task.

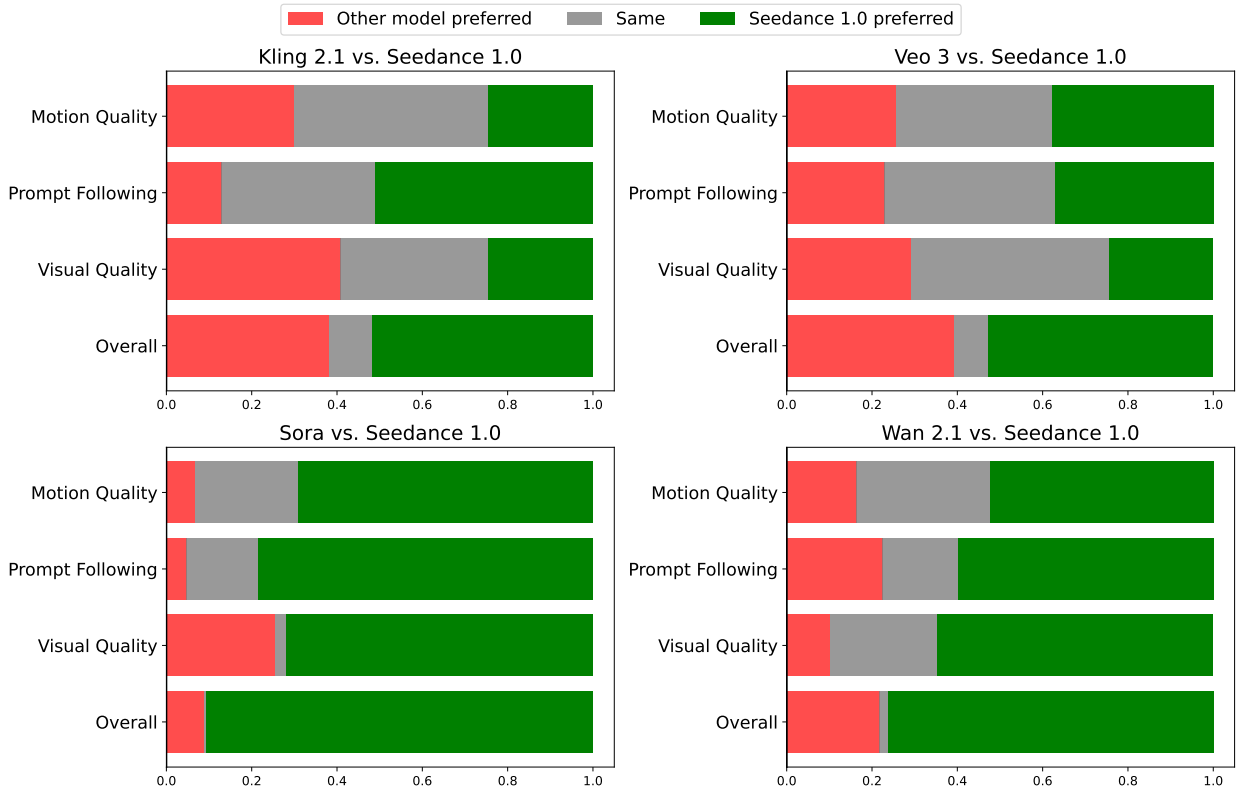


Figure 11 GSB Evaluation for Text-to-Video Task.

- *Subject* It is essential to first evaluate the model's ability to accurately generate primary entities, including humans, animals, natural scenes, consumer goods, and some virtual subjects.
- *Subject Description* The focus is on models' ability to produce accurate representations of primary subjects. It includes subject quantity, entity attributes (e.g. appearance characteristics of human subjects, object properties of physical items), and spatial positioning.
- *Action* Action simulation and generation represent fundamental capabilities of video generation models, indicative of their proficiency in capturing real-world dynamics and underlying physical laws. This category assesses motion-related actions across multiple categories, including human activities, multi-entity

interactions, animal locomotion, sports movements, natural phenomena (e.g., weather events, biological processes), physical principles (e.g., gravity, fluid dynamics), and creative or imaginative motion patterns.

- *Action Description* This category provides a finer-grained analysis of action generation, focusing on action number, movement direction, temporal sequencing, motion intensity, and expression of emotional states.
- *Camera* The camera language component reflects a distinctive dimension of artistic expression in video generation, encompassing camera movements, shooting angles, shot size definition and variation, as well as transitions between multiple shots. SeedVideoBench-1.0 integrates a range of professional camera movements, including circular tracking shots, dolly-in shots, Hitchcock zooms, lateral pans, and follow shots.
- *Aesthetic description* Aesthetics evaluation is an essential component in assessing visual generation models. This part encompasses style consistency, compositional atmosphere, lighting and shadow dynamics, and other factors governing the overall aesthetic quality of the generated videos.

The taxonomy for image-to-video is similar, with the addition of a labeling system for the first frame. For both text-to-video and image-to-video tasks, we construct 300 prompts each, uniformly distributed across the aforementioned categories. The quantity of prompts per category is designed to ensure sufficient discriminative and statistical confidence in the evaluation.

7.2.2 Video Evaluation Metrics

In collaboration with film directors, we developed a set of specialized evaluation metrics for generated videos, enabling assessment from a professional perspective. Unlike public preference evaluations, which often emphasize aesthetic appeal while neglecting fine-grained distinctions in model capabilities, this framework is structured around four core dimensions.

- *Motion Quality* Motion Quality is the first intuitive impression that generated videos bring to users. It includes multiple aspects such as structural accuracy, motion plausibility, motion stability, and motion vividness. Structural accuracy focuses on detecting structural anomalies in generated content, such as extra limbs, truncation, unnatural bending, or inhuman postures. Motion plausibility involves physical plausibility in trajectory and speed, adherence to physical laws and common sense, and the identification of unnaturally static subjects or those with insufficient movement amplitude. Separately, motion stability is evaluated to detect artifacts caused by subject or background dynamics, while motion vividness addresses the coherence and realism of action sequences, including macro-structural integrity and the aesthetic quality of camera motion.
- *Prompt Following* Prompt Following represents a foundational capability of generative models, reflecting their ability to produce content aligned with human intent. This evaluation focuses on multiple dimensions, including action responsiveness, subject description fidelity, stylistic conformity, incorporation of auxiliary entities, temporal alignment of motion, camera behavior, and environmental depiction accuracy.
- *Aesthetic Quality* Evaluation of aesthetic appeal and visual quality in generated video emphasizes visual texture, perceptibility of AI sense, material detail fidelity, and the artistic expression of aesthetic intent.
- *Preservation* Original image preservation, specific to image-to-video tasks, is assessed across multiple dimensions, including subject consistency, stylistic coherence, material fidelity, visual content alignment, and consistency in color and lighting.

7.2.3 Human Evaluation

Leveraging SeedVideoBench 1.0, we conducted a comprehensive comparative evaluation of Seedance 1.0 against several leading video generation models across two tasks: text-to-video and image-to-video generation. For the text-to-video task, comparative models include Kling 2.1(Master), Veo 3, Wan 2.1, and Sora; for the image-to-video task, Sora is replaced by Runway Gen4. Two evaluation protocols are adopted: Absolute Score and the Good-Same-Bad (GSB) comparison metric. The Absolute Score employs a five-point Likert scale (where 1 indicates extreme dissatisfaction and 5 signifies utmost satisfaction), facilitating unified performance

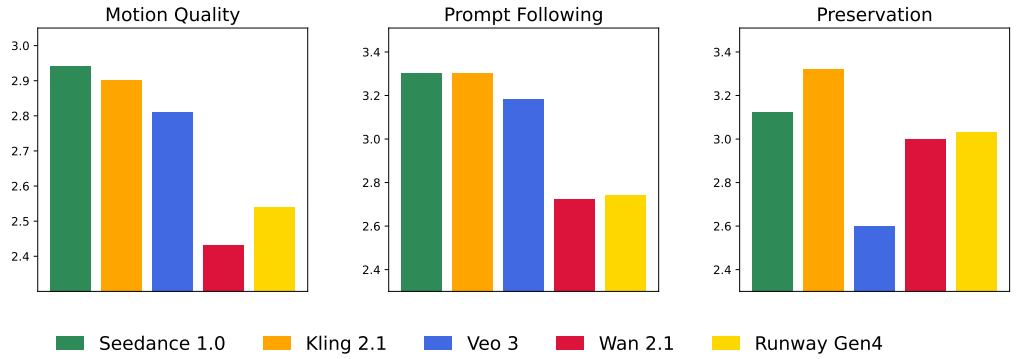


Figure 12 Absolute Evaluation for Image-to-Video task.

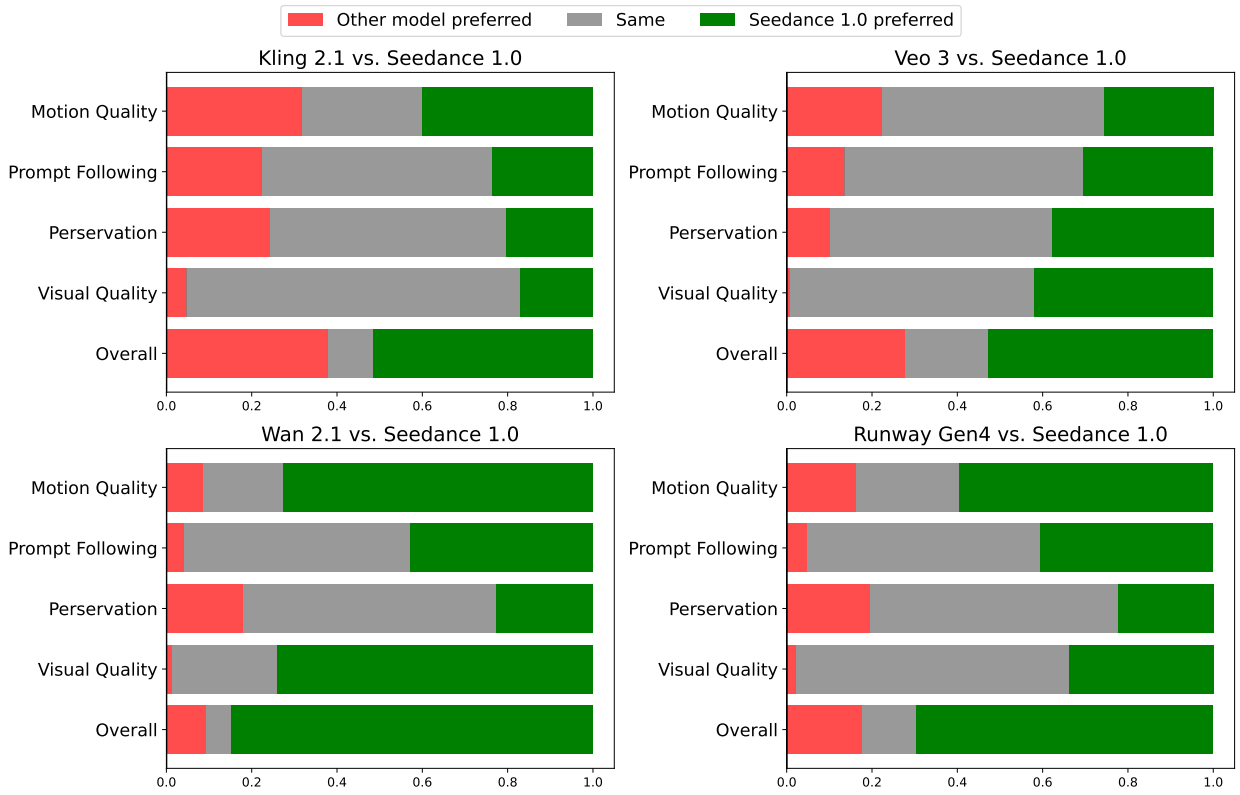


Figure 13 GSB Evaluation for Image-to-Video task.

comparison across models. The GSB metric conducts pairwise comparisons to assess relative video quality, enabling fine-grained differentiation between model outputs.

Figures 10 and 11 show the absolute scores and GSB results for video generation models in the text-to-video task. Seedance 1.0, Kling 2.1, and Veo 3 substantially outperform other models. While Kling 2.1 demonstrates strong motion quality and visual fidelity, its limited prompt-following capability negatively impacts its overall effectiveness. In text-to-video generation, precise instruction adherence is critical to the adoption of generated content. Seedance 1.0 and Veo 3 exhibit superior prompt-following capability, driving their higher rankings on the Artificial Analysis leaderboard. Veo 3 excels at generating realistic videos, but its comparatively weaker motion quality constrains its capacity for complex video synthesis.

Prompt: 电影感，暖色调，中远景拍摄，一个美丽的女人坐在酒吧门口哭泣，她一头短发，穿着红色长裙，背后是一片模糊的霓虹玻璃。镜头切换，特写她脚边的一地烟头和空酒瓶。镜头切换，一个穿西装的男人走到她身旁拍了拍她的肩膀

Figure 14 Comparison of Multi-Shot Generation. Top: Seedance 1.0; Middle: Kling 2.1; Bottom: Veo 3.

Figures 12 and 13 present the absolute scores and GSB results for the image-to-video task. Seedance 1.0 and Kling 2.1 exhibit strong overall performance in this scenario. Adding image input as a condition introduces challenges in preserving character and background. Veo 3 performs relatively weak in this regard, occasionally altering lighting conditions, object textures, and other visual elements of the reference image. Additionally, it suffers from some quality degradation issues such as oily appearance or blurred details, which substantially affect its overall effectiveness. Kling 2.1 excels in motion quality, producing natural and coherent dynamics suitable for complex scenarios, though it occasionally experiences detail breakdown. Seedance 1.0 matches Kling 2.1's motion quality while offering superior prompt-following capability in scenarios involving complex shot transitions or detailed instruction prompts, resulting in more favorable overall performance.

Prompt:从观众视角切入场内灯光聚焦，转为拳手挥拳慢动作，接着是对手反应的超近特写，最后切到裁判吹哨一刻的动静对比。

Prompt: 镜头从废墟中前行机器人脚步切入，切至头部光学镜头扫描特写，再转为它视角中的城市轮廓扫描图像，最终定格在墙上涂鸦'HUMANITY?'。

Prompt:清晨，一位少年骑着自行车穿过欧洲老城区。镜头从石板路上的车轮切入，切换为街边行人的仰视视角，再是他穿梭巷道的第三人称航拍视角，最后以他驶入阳光洒满的广场作为结束画面。

Figure 15 Multi-Shot Generation for Seedance 1.0.

7.3 Multi-Shot Generation

Seedance 1.0 demonstrates the capability to generate multiple consecutive shots from a single prompt, while ensuring subject continuity and stylistic coherence across frames. This enables the model to handle complex narrative techniques commonly used in cinematic storytelling. Specifically, Seedance 1.0 facilitates the construction of shot-reverse shot sequences for dialogic interaction, as well as the use of cut-in and cut-away shots to enrich narrative pacing and contextual layering. Furthermore, it supports match cuts and action cuts, enabling seamless transitions and preserving visual continuity. These competencies highlight Seedance's proficiency in cinematic shot composition and temporal coherence, offering enhanced creative control and narrative expressiveness for video content generation. Figure 14 presents an example of continuous shot transitions generated by Seedance 1.0, which exhibits more coherent and fluid cinematic storytelling compared to other models.

Prompt: 像素风，像素化的纽约曼哈顿，像素直升机在高楼间穿梭

Prompt: 微缩景观，羊绒绒羊毛毡，超级特写，一只小狐狸奔跑在金黄色的麦浪中，小王子坐在小山坡上看着它，阳光洒在麦田里，泛起波光粼粼

Prompt: [动物新闻网] 一只系着领带的长颈鹿站在用编织藤蔓做成的新闻台前。它身后的画面展示着正在迁徙的斑马，还有滚动字幕实时更新情况。一只巨嘴鸟气象播报员扇动着翅膀飞进画面，用鸟喙轻敲悬浮屏幕，预报明天水坑的情况。

Figure 16 Multi-Style Generation for Seedance 1.0.

7.4 Multi-Style Alignment

Seedance 1.0 exhibits strong generalization across a broad spectrum of visual styles. In text-to-video (T2V) tasks, Seedance 1.0 enables direct generation of fine-grained stylistic videos, while in image-to-video (I2V) tasks, it reliably preserves visual characteristics of the reference image. The model supports a wide range of real-world cinematic styles, including black-and-white silent films, classic Hong Kong cinema, and retro Hollywood aesthetics, as well as animated and fantasy-oriented styles such as Japanese anime, cyberpunk futurism, and ink-wash animation. This multi-style adaptability facilitates seamless transitions between realism and fantasy without the need for extensive task-specific tuning. As a result, Seedance 1.0 offers exceptional versatility and controllability, making it well-suited for professional filmmaking and AIGC creation.

7.5 Visualization

We present several visual outcomes by Seedance 1.0 in Figure 14,15,16. For additional examples, please refer to the official website for an enhanced viewing experience.

8 Conclusion

We have introduced Seedance 1.0, a native bilingual video generation foundation model that unifies multiple generation paradigms (such as text-to-video and image-to-video) and excels in instruction following, motion stability, and visual quality. We presented our technical improvements in dataset construction, efficient architecture design with training paradigm, post-training optimization, and inference acceleration, which are integrated effectively to achieve a high-performance model with fast inference. It demonstrates excellent capabilities in handling complex scenarios, multi-shot generation, and long-range temporal coherence, all while delivering fast and photorealistic generation experiences.

References

[1] artificialanalysis.ai. artificialanalysis. https://artificialanalysis.ai/text-to-video/arena?tab=leaderboard, 2025.

[2] ByteDance. bmf. https://babitmf.github.io/, 2024.

[3] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. arXiv preprint arXiv:2410.10733, 2024.

[4] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. arXiv e-prints, pages arXiv–2305, 2023.

[5] 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.

[6] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346, 2025.

[7] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In European Conference on Computer Vision, pages 205–224. Springer, 2024.

[8] Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. arXiv preprint arXiv:2503.07703, 2025.

[9] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. arXiv preprint arXiv:2503.10589, 2025.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[12] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. arXiv preprint arXiv:2309.14509, 2023.

[13] Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian GE, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, et al. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. arXiv preprint arXiv:2412.15156, 2024.

[14] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. CoRR, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

[15] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.

[16] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. arXiv preprint arXiv:2501.08316, 2025.

[17] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470, 2025.

[18] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. arXiv preprint arXiv:2501.13918, 2025.

[19] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging {AI} applications. In 13th USENIX symposium on operating systems design and implementation (OSDI 18), pages 561–577, 2018.

[20] William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4195–4205, 2023.

[21] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36:53728–53741, 2023.

[22] Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. Advances in Neural Information Processing Systems, 37:117340–117362, 2025.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, pages 10684–10695, 2022.

[24] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. arXiv preprint arXiv:2504.08685, 2025.

[25] Huiyang Shao, Xin Xia, Yuhong Yang, Yuxi Ren, Xing Wang, and Xuefeng Xiao. Rayflow: Instance-aware diffusion acceleration via adaptive flow trajectories. arXiv preprint arXiv:2503.07699, 2025.

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

[27] Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. arXiv preprint arXiv:2502.21079, 2025.

[28] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818, 2025.

[29] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[30] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072, 2024.

[31] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737, 2023.

[32] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. arXiv preprint arXiv:2501.07888, 2025.

[33] Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. Onlinevpo: Align video diffusion model with online video-centric preference optimization. arXiv preprint arXiv:2412.15159, 2024.

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

[35] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. arXiv preprint arXiv:2304.11277, 2023.

Appendix

A Contributions and Acknowledgments

All contributors of Seedance are listed in alphabetical order by their last names.

Core Contributors

Yu Gao
Haoyuan Guo
Tuyen Hoang
Weilin Huang
Lu Jiang
Fangyuan Kong
Huixia Li
Jiashi Li
Liang Li
Xiaojie Li
Xunsong Li
Yifu Li
Shanchuan Lin
Zhijie Lin
Jiawei Liu
Shu Liu
Xiaonan Nie
Zhiwu Qing
Yuxi Ren
Li Sun
Zhi Tian
Rui Wang
Sen Wang
Guoqiang Wei
Guohong Wu
Jie Wu
Ruiqi Xia
Fei Xiao
Xuefeng Xiao
Jiangqiao Yan
Ceyuan Yang
Jianchao Yang
Runkai Yang
Tao Yang
Yihang Yang
Zilyu Ye
Xuejiao Zeng
Yan Zeng
Heng Zhang
Yang Zhao
Xiaozheng Zheng
Peihao Zhu
Jiaxin Zou
Feilong Zuo

Contributors

Sheng Bi
Hao Chen
Haoshen Chen
Haoxin Chen
Xiaoya Chen
Feng Cheng
Xuyan Chi
Xiaojing Dong
Junliang Fan
Jing Fang
Liangke Gui
Qiushan Guo
Bibo He
Ruoqing Hu
Siqi Jiang
Ashley Kim
Gen Li
Yiying Li
Haibin Lin
Feng Ling
Gaohong Liu
Zuxi Liu
Zhibei Ma
Yanghua Peng
Lei Shi
Zuquan Song
Renfei Sun
Qinlong Wang
Xuanda Wang
Xun Wang
Ye Wang
Meng Wei
Yawei Wen
Ruolan Wu
Xiaohu Wu
Yonghui Wu
Xin Xia
Tingshuai Yan
Zhouqike Yang
Ziyan Yang
Linxiao Yuan
Zhonghua Zhai
Manlin Zhang
Xinyan Zhang

Xinyu Zhang
Zixiang Zhang
Qi Zhao
Rui Zhu
Wenjia Zhu