SPECULATIVE DECODING FOR AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video diffusion has emerged as a promising paradigm for streaming video synthesis, yet accelerating inference remains a critical challenge. While speculative decoding is the dominant acceleration strategy for large language models (LLMs), adapting it to video generation remains non-trivial, as continuous spatiotemporal video tensors lack discrete token probability distributions for exact rejection sampling. To bridge this gap, we introduce SDVG, a training-free framework that brings speculative decoding to autoregressive video diffusion via reward-guided block verification. Instead of discrete logits, SDVG employs a lightweight image-quality router with *worst-frame score aggregation* to evaluate candidate blocks proposed by a compact draft model. Accepted drafts are directly committed to the target model's key-value (KV) cache, while rejected blocks trigger target regeneration. By enforcing mandatory target-based generation on the initial anchor block to fix global scene composition, a single calibrated threshold provides a simple, continuous control knob over the quality–speed trade-off. We further introduce SDVG-hybrid, which integrates step-level trajectory stitching on rejected blocks to reuse draft compute and incorporates an on-demand temporal cascade to minimize verification overhead. Evaluated on $1{,}003$ MovieGenVideoBench prompts (832×480), standard SDVG achieves a $1.59\times$ speedup while preserving 98.1% of target visual quality (VisionReward), reaching up to $2.05\times$ acceleration at 95.9% retention ($+17.4\%$ over draft-only generation). Furthermore, SDVG-hybrid advances the Pareto frontier to a $1.88\times$ speedup at 98.0% quality retention and $1.96\times$ at 96.9% retention. Requiring no retraining or architectural modifications, SDVG serves as a practical, plug-and-play accelerator for modern autoregressive video pipelines.

1 INTRODUCTION

Autoregressive video generation has recently emerged as a compelling paradigm for efficient, streaming video synthesis. Unlike conventional video diffusion models that generate all frames jointly (Brooks et al., 2024; Polyak et al., 2024), autoregressive approaches produce video content block by block, conditioning each new block on previously generated context through a shared key-value (KV) cache—mirroring the autoregressive framework of large language models (LLMs). This design naturally enables streaming generation, allowing frames to be displayed progressively as they are synthesized rather than waiting for the entire sequence to complete. Building on this paradigm, recent frameworks such as Self-Forcing (Huang et al., 2025) train causal video diffusion transformers with self-generated conditioning to mitigate exposure bias, enabling real-time streaming video output on a single GPU.

Despite these structural advantages, state-of-the-art autoregressive video models heavily rely on massive architectures (10B+ parameters), incurring severe computational overhead. For instance, frontier open-source 14B autoregressive models (Millon, 2025) require enterprise-grade hardware (e.g., NVIDIA B200 GPUs) to attain real-time throughput. Conversely, compact 1B-scale models (Wan et al., 2025) operate at less than a quarter of the computational cost, albeit with a noticeable compromise in visual fidelity. This stark trade-off leads to our central research question:

Can we capture the speed of small models while retaining the generation quality of large ones?

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



Figure 1: **Qualitative comparison on MovieGenVideoBench.** Each cell shows four uniformly sampled frames. Rows compare draft-only, SDVG, SDVG-hybrid, and target-only; columns show Big Sur and Tokyo street prompts. SDVG (τ=-0.7) closely matches target quality at **1.59**$\times$ speedup; SDVG-hybrid (k=3, τ=-1.0, cascade) reaches **1.96**$\times$.

Speculative decoding in large language models (LLMs) (Leviathan et al., 2023; Chen et al., 2023) offers a compelling paradigm: a compact draft model proposes candidate tokens, while a large target model verifies them only when necessary. The block-by-block architecture of autoregressive video generation aligns naturally with this strategy. Because each generated block serves as a self-contained spatiotemporal unit before being committed to the key-value (KV) cache, it provides a natural boundary for per-block verification and dynamic model routing.

However, enabling seamless cooperation between large and small models in video synthesis remains non-trivial. Existing collaborative inference paradigms fall short in autoregressive contexts: step-level stitching methods (e.g., T-Stitch (Pan et al., 2024), SRDiffusion (Cheng et al., 2025), and HybridStitch (Sun et al., 2026)) split the denoising trajectory across timesteps rather than spatial/temporal blocks, ignoring the causal autoregressive structure. On the other hand, system-level routing schemes like MoDM (Xia et al., 2025) rely on global cache hits and fail to offer per-block quality guarantees. Consequently, these approaches yield suboptimal performance for autoregressive video generation.

Crucially, a fundamental domain gap prevents the direct application of LLM speculative decoding to video: while classical speculative decoding verifies drafts via exact token probability distributions (Leviathan et al., 2023), video blocks are continuous, high-dimensional spatiotemporal tensors lacking discrete logits. Thus, standard token-level acceptance criteria are strictly inapplicable, leaving *speculative decoding for autoregressive video* an open challenge.

In this work, we propose **Speculative Decoding for Autoregressive Video Generation** (**SDVG**), a training-free, plug-and-play framework that requires no architectural modifications to either the draft or target models. Our primary contributions are summarized as follows:

- **Pioneering Speculative Decoding for Autoregressive Video**: To the best of our knowledge, SDVG is the **first** training-free speculative decoding framework tailored for autoregressive video diffusion. By dynamically routing video blocks via reward-guided quality metrics, SDVG achieves a $1.59\times$ speedup while preserving 98.1% of the target model's visual quality—demonstrating that plain reward routing suffices without complex trajectory engineering.
- **Video-Specific Routing Design**: We identify three critical design choices essential for effective continuous-domain verification: (1) a calibration-free quality threshold acting as a robust speed–quality control knob, (2) mandatory target-based first-block generation to anchor scene composition, and (3) a worst-frame quality scoring strategy that exposes transient single-frame artifacts obscured by block-level spatial averaging.
- **Pushing Acceleration Limits via SDVG-Hybrid**: We further extend our framework into SDVG-hybrid by synergizing block-level speculation with step-level denoising trajectory stitching. By delegating initial denoising steps of rejected blocks to the draft model and incorporating an on-

demand evaluation cascade, SDVG-hybrid achieves up to a $1.96\times$ speedup with 96.9% quality retention, significantly outperforming aggressive thresholding on SDVG alone.

2 Background and Related Work

Video Generation and Acceleration. Diffusion-based video models have evolved from early pixel-space formulations (Ho et al., 2022) to large latent transformer architectures (Brooks et al., 2024; Polyak et al., 2024). Inference acceleration in this domain has primarily relied on step distillation (Yin et al., 2024; Zhang et al., 2025c; Wang et al., 2024) and GPU kernel optimization (Zhang et al., 2025a;b). These efficiency-focused techniques operate orthogonally to model architecture routing and can be seamlessly combined with SDVG.

Autoregressive Video Generation. Autoregressive video models synthesize future video blocks causally based on previously generated context. Traditional autoregressive training relies on ground-truth historical frames (teacher forcing), creating a fundamental train-test discrepancy that induces severe exposure bias during inference. Recent frameworks, such as Self-Forcing (Huang et al., 2025), mitigate this issue by training causal video diffusion transformers with self-generated conditioning, enabling stable real-time streaming video generation.

Collaborative Multi-Model Inference. Collaborative inference leverages models of varying computational capacities to reduce synthesis overhead. For instance, T-Stitch (Pan et al., 2024), SRDiffusion (Cheng et al., 2025), and HybridStitch (Sun et al., 2026) split the denoising trajectory between compact and large models at pre-defined noise levels. System-level approaches like MoDM (Xia et al., 2025) route whole requests based on global cache hits, reducing average latency. However, these methods rely on content-agnostic timestep partitioning or static caching rules, making them ill-suited for dynamic, block-wise autoregressive video generation.

Speculative Decoding. Speculative decoding (Leviathan et al., 2023; Chen et al., 2023) pairs a compact draft model with a large target model, where the drafter proposes token sequences that are verified in parallel by the target to strictly preserve the target distribution. Recent work like RSD (Liao et al., 2025) extends speculative verification to reasoning steps via a Process Reward Model (PRM). Inspired by this paradigm, SDVG brings speculative verification to autoregressive video. Because continuous spatiotemporal video tensors lack discrete logit distributions, exact rejection sampling is mathematically inapplicable; we bridge this gap by replacing discrete verification with an image quality model acting as a proxy reward for block-level routing.

3 Method

This section introduces SDVG, a training-free speculative decoding framework tailored for autoregressive video diffusion. We first present the core block-level speculation mechanism (Section 3.1), detailing its inference workflow, worst-frame quality aggregation, and anchor-block strategy. We then extend this foundation to SDVG-hybrid (Section 3.2), which integrates step-level trajectory stitching with proxy-based routing on rejected blocks to push the limits of computational acceleration.

3.1 SDVG: Block-Level Speculation

Given a text prompt p, a compact draft model $\mathcal{D}$, and a large target model $\mathcal{T}$, our goal is to define a routing policy $\pi : \mathbb{R} \to \{0, 1\}$ that maps a per-block quality score $q_b \in \mathbb{R}$ to a binary decision—accept (1) or reject (0)—thereby optimizing the throughput–quality trade-off during streaming inference.

Inference Workflow. For each autoregressive block b, the draft model $\mathcal{D}$ executes S denoising steps to propose a candidate latent block $\hat{\mathbf{x}}_b$. This candidate is subsequently mapped to the pixel domain via a VAE decoder and evaluated by a lightweight router. Serving as the continuous-domain surrogate for token verification in classical speculative decoding, the router evaluates the decoded frames and outputs a scalar quality score q_b. If $q_b \geq \tau$, the candidate is *accepted*: $\hat{\mathbf{x}}_b$ is committed to the target model's key-value (KV) cache $\mathbf{K}^{\mathcal{T}}$, and the decoded frames are immediately streamed to output. Conversely, if $q_b < \tau$, the draft is *rejected*; the target model $\mathcal{T}$ regenerates the block

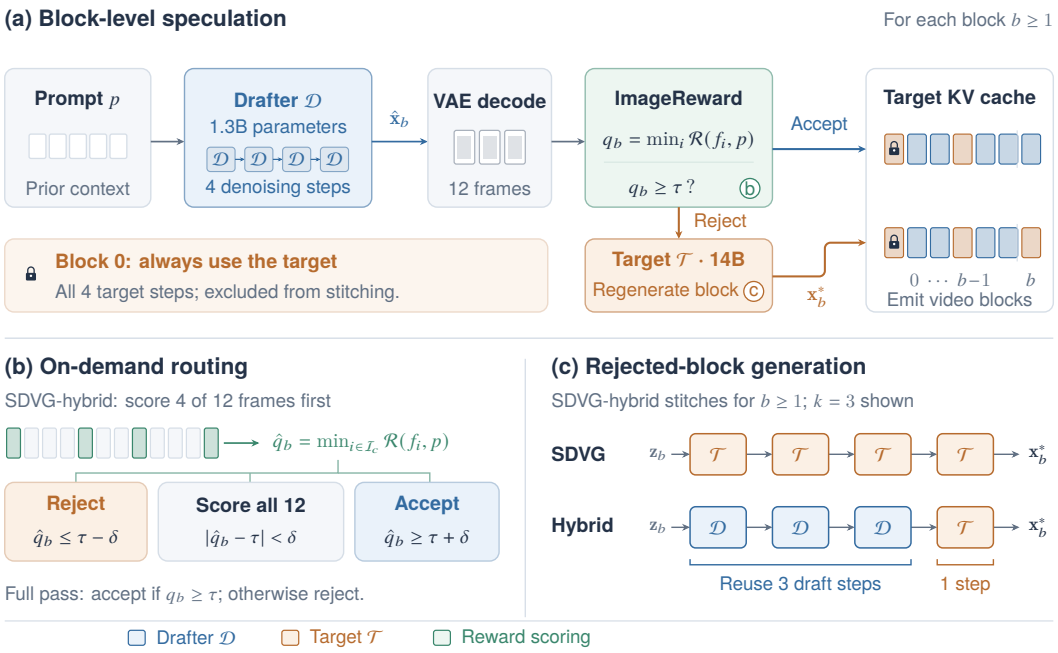

Figure 2: **Inference pipeline.** For each video block, the 1.3B drafter generates a candidate in 4 denoising steps. The draft is VAE-decoded and scored by ImageReward with min-frame aggregation ($q_b = \min_i \mathcal{R}(f_i, p)$). If $q_b \geq \tau$, the draft is accepted (blue) and committed to the target KV cache; otherwise the 14B target regenerates the block (orange). Block 0 is always force-rejected to anchor scene composition. SDVG-hybrid modifies two points of this flow (panels b and c): the router scores only 4 of the 12 decoded frames and runs the full pass solely when the proxy lands in the grey band, and on rejected blocks the target resumes from step k of the schedule rather than from the initial noise, the earlier steps being taken from the drafter's trajectory. Block 0 is excluded from stitching.

from scratch, updates $\mathbf{K}^{\mathcal{T}}$, and emits its verified frames. The decision threshold τ is a fixed scalar calibrated offline to balance inference speed and visual fidelity.

Worst-Frame Quality Aggregation. Because speculative verification is executed per block, the aggregate score q_b must reflect the visual integrity of the entire temporal segment. Standard spatial or temporal mean-pooling across frames can easily conceal transient perceptual artifacts—such as localized structural distortions or single-frame motion blurs—that remain highly perceptible to human viewers. To prevent artifact propagation, we adopt a conservative *worst-frame* scoring mechanism that takes the minimum reward across all F decoded frames within block b:

$$q_b = \min_{i=1\ldots F} \mathcal{R}\left(\mathbf{f}_i^{(b)}, p\right), \tag{1}$$

where $\mathcal{R}(\mathbf{f}, p)$ denotes the frame-level quality score computed by a lightweight image reward model (e.g., ImageReward) given prompt p.

Mandatory Target Generation for Anchor Block ($b = 0$). At the onset of generation, the initial block $b = 0$ operates without prior context and serves as the global visual anchor—establishing scene composition, spatial layout, and stylistic attributes inherited by all subsequent blocks via the shared KV cache $\mathbf{K}^{\mathcal{T}}$. Accepting an imperfect draft at $b = 0$ risks compounding structural degradation throughout the entire video sequence. To anchor high-quality synthesis, we enforce a strict policy: block $b = 0$ is always generated by the target model $\mathcal{T}$, bypassing draft evaluation entirely.

3.2 SDVG-Hybrid: Step-Level Stitching with On-Demand Routing

While standard SDVG operates purely at the block level—relegating rejected blocks entirely to the target model $\mathcal{T}$—SDVG-hybrid decouples the spatial-temporal block axis from the denoising

timestep axis. Upon rejection, rather than discarding candidate latents and re-evaluating from scratch, SDVG-hybrid delegates only the necessary terminal denoising steps to $\mathcal{T}$. Reusing intermediate draft compute allows the router to adopt finer-grained control and a more aggressive acceptance policy.

Step-Level Trajectory Stitching. In standard SDVG, a rejected block triggers a full target resampling across all S denoising steps, discarding the intermediate latents computed during candidate generation. To eliminate this compute redundancy, SDVG-hybrid retains the initial k denoising steps executed by $\mathcal{D}$ during evaluation, invoking $\mathcal{T}$ only for the remaining $S-k$ steps (Pan et al., 2024):

$$\hat{\mathbf{x}}_b^{(k)} = \mathcal{D}^k(\mathbf{z}_b), \qquad \mathbf{x}_b^* = \mathcal{T}^{S-k}\left(\hat{\mathbf{x}}_b^{(k)}\right), \tag{2}$$

where $\mathbf{z}_b$ denotes initial Gaussian noise, $\mathcal{D}^k$ represents k denoising steps by the draft model, and $\mathcal{T}^{S-k}$ represents the remaining $S-k$ refinement steps executed by the target model. Because $\hat{\mathbf{x}}_b^{(k)}$ is already obtained during the candidate scoring pass, this step-level stitching adds zero computational overhead while leveraging $\mathcal{T}$ where its capacity matters most (i.e., fine-grained detail generation at low noise levels). As with standard SDVG, the initial anchor block ($b=0$) remains fully generated by $\mathcal{T}$ to preserve global visual coherence.

On-Demand Cascade Routing. In practice, decoding every candidate frame via VAE and scoring it through the reward model introduces non-negligible verification overhead. To minimize this bottleneck, we introduce a hierarchical, on-demand scoring mechanism. We first extract a sparse subset of $F_c = 4$ temporally uniform proxy frames to compute a preliminary score $\hat{q}_b$. The router then conditionally determines whether to fall back to the full F-frame evaluation based on a confidence margin δ:

$$\text{Decision} = \begin{cases} \text{Accept}, & \text{if } \hat{q}_b \geq \tau + \delta, \\ \text{Reject}, & \text{if } \hat{q}_b \leq \tau - \delta, \\ \text{Evaluate } q_b \text{ on all } F \text{ frames}, & \text{if } \hat{q}_b \in (\tau - \delta,\, \tau + \delta). \end{cases} \tag{3}$$

When the proxy evaluation yields a decisive result ($\hat{q}_b \notin (\tau-\delta, \tau+\delta)$), full F-frame decoding and scoring are completely bypassed, reducing verification cost to F_c/F of the full scoring pass. Empirically, the vast majority of blocks fall outside this ambiguous interval (e.g., only 9.1% land within $(\tau-\delta, \tau+\delta)$ at $\tau=-1.0$), resulting in significant operational speedups.

4 Experiments

In this section, we empirically evaluate the performance and efficiency of SDVG and SDVG-hybrid. We first outline the base models, hardware infrastructure, and evaluation protocols (Section 4.1). We then analyze the Pareto trade-offs between video quality and generation speed against boundary baselines (Section 4.2). Finally, we conduct thorough ablation studies to isolate the individual contributions of key architectural components, including worst-frame quality aggregation, anchor block generation, and on-demand cascade routing (Section 4.3).

4.1 Experimental Setup

We pair a Krea Realtime Video 14B target (Millon, 2025) with a Wan2.1-T2V-1.3B Self-Forcing draft and use ImageReward (Xu et al., 2023) for routing. On 1,003 MovieGenVideoBench prompts (Polyak et al., 2024), each method generates nine video blocks at 832×480 using two NVIDIA RTX A6000 GPUs. We report VisionReward, EvalCrafter, latency per video, and speedup over target-only; draft-only and target-only bound the quality–speed range. Appendix A provides the denoising schedule, decoding protocol, hardware configuration, baseline definitions, and reproducibility details.

4.2 Main Results

We systematically evaluate SDVG and SDVG-hybrid across various quality thresholds τ, benchmarking them against the boundary *Draft-only* ($\mathcal{D}$-only) and *Target-only* ($\mathcal{T}$-only) baselines. Table 1 summarizes the overall quality–speed trade-offs on MovieGenVideoBench alongside fine-grained EvalCrafter metrics, while Figure 3 illustrates the corresponding Pareto frontiers.

Table 1: **Main results on MovieGenVideoBench.** SDVG and SDVG-hybrid across the routing threshold. VR is VisionReward, higher being better, and Ret. is the same quantity expressed as a percentage of target-only. Every SDVG-hybrid row stitches k=3 denoising steps of a rejected block and resolves most routing decisions from a four-frame proxy before falling back on all twelve, leaving block 0 to the target, so the rows differ only in the threshold. The lower panel reports EvalCrafter's raw metrics for the same settings, with the arrows following the direction that benchmark's fitted aggregation rewards (for IS, action score and flow score, lower is better) and with *warp* denoting the optical-flow warping error, *BLEU* the BLIP caption BLEU, and *detect* the object detection rate.

Method	τ	VR ↑	Ret. (%)	Time (s)	Speedup	Accept
Target-only	—	0.0788	100.0	97.0	1.00×	—
SDVG	−0.7	0.0773	98.10	60.9	1.59×	73.1%
SDVG	−1.0	0.0764	96.95	57.2	1.69×	78.0%
SDVG	−1.5	0.0757	96.07	51.6	1.88×	83.4%
SDVG	−2.0	0.0756	95.94	47.4	2.05×	87.5%
SDVG-hybrid	−0.7	0.0772	97.97	51.6	1.88×	72.8%
SDVG-hybrid	−1.0	0.0764	96.95	49.5	1.96×	77.5%
SDVG-hybrid	−1.5	0.0759	96.32	48.2	2.01×	83.2%
SDVG-hybrid	−2.0	0.0751	95.30	46.1	2.11×	87.4%
Draft-only	—	0.0644	81.73	25.7	3.77×	—

Method	τ	VQA_T ↑	IS ↓	warp ↓	action ↓	flow ↓	BLEU ↑	detect ↑
Target-only	—	93.33	18.62	0.0066	45.18	4.99	21.66	78.09
SDVG	−0.7	93.11	18.99	0.0085	45.38	5.16	21.92	77.85
SDVG	−1.0	93.08	19.17	0.0088	45.29	5.18	21.89	78.44
SDVG	−1.5	92.81	19.15	0.0090	45.28	5.31	21.74	78.54
SDVG	−2.0	93.04	19.20	0.0092	45.45	5.35	21.73	78.70
SDVG-hybrid	−0.7	92.94	19.13	0.0089	45.49	5.39	21.87	78.19
SDVG-hybrid	−1.0	92.94	19.01	0.0091	45.38	5.38	21.88	77.88
SDVG-hybrid	−1.5	93.13	19.12	0.0092	45.47	5.37	21.86	77.90
SDVG-hybrid	−2.0	92.98	19.08	0.0092	45.48	5.36	21.78	77.82
Draft-only	—	92.75	19.52	0.0102	46.95	5.90	21.00	72.57

Quality–Speed Trade-Off. Varying the minimum-frame threshold τ from -0.7 to -2.0 traces a smooth Pareto frontier between the two baseline models (Figure 3). Under conservative thresholding ($\tau = -0.7$), SDVG retains 98.10% of target-only VisionReward (0.0773 vs. 0.0788) while delivering a $\mathbf{1.59}\times$ speedup. Relaxing the threshold to $\tau = -1.0$ accelerates inference to $\mathbf{1.69}\times$ with 96.95% quality retention. Even under aggressive thresholding ($\tau = -2.0$), SDVG achieves a $\mathbf{2.05}\times$ speedup while preserving 95.94% of target quality—outperforming the *Draft-only* baseline by $+17.4\%$ in VisionReward (0.0756 vs. 0.0644). Latency gains exhibit diminishing returns beyond $\tau = -1.5$: because most quality-critical blocks naturally clear this threshold, further relaxation yields marginal speed improvements at the expense of visual fidelity.

Quality–Acceptance Rate Relationship. Analyzing visual quality relative to block acceptance rates reveals a highly favorable regime at conservative thresholds. Increasing the acceptance rate from 73.1% ($\tau = -0.7$) to 78.0% ($\tau = -1.0$) results in only a minimal 1.2% relative decline in VisionReward ($0.0773 \rightarrow 0.0764$). This confirms that the router effectively admits borderline draft candidates whose visual quality closely matches target-generated outputs. Beyond 78% acceptance, quality degradation accelerates—dropping to 0.0757 at 83.4% acceptance and 0.0756 at 87.5%—as sub-optimal draft blocks are increasingly admitted. Consequently, the threshold interval $\tau \in [-1.0, -0.7]$ (corresponding to 73%–78% block acceptance) represents the optimal operating regime for balancing quality and efficiency.

Pareto Extension via SDVG-Hybrid. Integrating step-level trajectory stitching ($k = 3$) and on-demand cascade routing into SDVG-hybrid significantly expands the Pareto frontier. At $\tau = -1.0$,

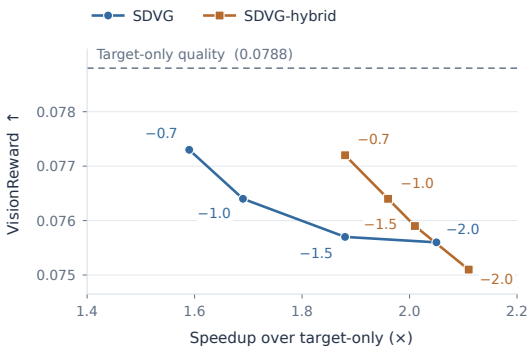

Figure 3: **Quality–speed Pareto curve.** VisionReward against speedup over target-only for the threshold sweeps of SDVG and SDVG-hybrid; the dashed line is target-only. The SDVG-hybrid curve sits above the SDVG curve through $2.0\times$, with the gap widest at the conservative end: hybrid $\tau{=}{-}0.7$ matches SDVG $\tau{=}{-}0.7$ quality 9.3 s sooner.

SDVG-hybrid achieves a $\mathbf{1.96}\times$ speedup (49.5 s per video) while preserving 96.9% of target quality (0.0764 VisionReward). This operating point strictly dominates standard SDVG at $\tau = -1.5$ across both speed and quality dimensions. Furthermore, compared to standard SDVG at the same threshold ($\tau = -1.0$), SDVG-hybrid boosts acceleration from $1.69\times$ to $1.96\times$ with zero degradation in visual quality. This improvement stems from two factors: trajectory stitching reuses intermediate draft latents instead of replaying full target denoising, while cascade routing eliminates redundant VAE decoding and reward evaluation overheads. The relative advantage of the hybrid model narrows as generation approaches $2.0\times$ acceleration, where permissive thresholds allow draft quality limits to dominate runtime savings. We disentangle these individual mechanisms in Section 4.3.

4.3 Ablation Studies

We conduct comprehensive ablation studies to isolate the individual contributions of key components in SDVG and SDVG-hybrid. Table 2 evaluates the foundational design choices of block-level speculation—specifically the choice of routing signal and frame-level score aggregation. Table 3 systematically deconstructs SDVG-hybrid across stitching depths k, anchor-block protection policies, and cascade thresholds. All experimental configurations enforce mandatory target generation for the anchor block ($b = 0$) unless explicitly noted otherwise.

Efficacy of Reward-Guided Routing. To verify that the ImageReward signal provides meaningful quality guidance, we replace the reward router with random accept/reject decisions matched to an identical acceptance rate (70.0%). As shown in Table 2, random routing causes VisionReward to drop sharply to 0.0706—a 0.0067 decrease compared to reward-guided SDVG at $\tau = -0.7$. Lacking visual awareness, the random router frequently admits artifact-laden drafts while rejecting high-quality candidates at equal rates, completely negating the benefit of selective regeneration.

To delineate the utility bounds of our policy, we evaluate a baseline that enforces anchor block protection ($b = 0$) while unconditionally accepting all subsequent draft blocks without verification. This naive policy achieves 0.0757 VisionReward at $2.21\times$ speedup. Applying random routing on top of this policy yields 0.0771, indicating that while protecting the anchor block provides a crucial structural foundation, dynamic reward-guided routing is indispensable for recovering near-target visual quality.

Worst-Frame vs. Mean-Frame Aggregation. Standard SDVG evaluates candidate blocks via worst-frame aggregation (Eq. 1), flagging a block as soon as a single frame exhibits visual degradation. As reported in Table 2, mean-pooling variants consistently underperform worst-frame aggregation at comparable acceptance rates (0.0755 at 78.4% acceptance for mean-pooling vs. 0.0773 at 73.1% for worst-frame at $\tau = -0.7$). This disparity arises because temporal mean-pooling across F frames absorbs localized structural distortions and single-frame motion artifacts, allowing corrupted blocks to pass verification undetected.

Deconstructing SDVG-Hybrid. Table 3 provides a detailed breakdown of SDVG-hybrid, systematically isolating the impact of stitching depth k, anchor-block protection, and on-demand cascade verification across multiple routing thresholds.

Table 2: **Ablation study on MovieGenVideoBench.** 1,003 prompts, 832×480 resolution, 9 blocks/video. VR = VisionReward (higher is better). Time = average wall-clock latency per video. Acceptance rate is computed across all 9 blocks, with the always-rejected anchor block ($b = 0$) counted as a rejection. "—" indicates not applicable.

Method	VR ↑	Time (s)	Speedup	Accept
Target-only	0.0788	97.0	1.00×	—
SDVG (τ=−0.7, ours)	**0.0773**	60.9	1.59×	73.1%
Avg-frame scoring (replacing min-frame)				
Avg-frame τ=−0.2	0.0767	63.2	1.54×	70.2%
Avg-frame τ=−0.5	0.0759	59.0	1.64×	75.3%
Avg-frame τ=−0.7	0.0755	56.9	1.71×	78.4%
Routing signal ablation				
Force-reject + random routing	0.0771	58.2	1.67×	70.3%
Force-reject only, no verification	0.0757	43.8	2.21×	100.0%
Random routing	0.0706	60.2	1.61×	70.0%
Draft-only	0.0644	25.7	3.77×	—

Table 3: **SDVG-hybrid ablations on MovieGenVideoBench.** 1,003 prompts, 832×480 resolution, 9 blocks/video, worst-frame ImageReward scoring. k denotes the number of denoising steps delegated to the draft model on rejected blocks. Retention denotes VisionReward relative to target-only (0.0788). The default rows represent full configurations, while lower panels vary one factor at a time.

Method	Cascade	VR ↑	Time (s)	Speedup	Retention
Target-only	—	0.0788	97.0	1.00×	100.0%
SDVG-hybrid (k=3, τ=−0.7)	✓	**0.0772**	**51.6**	**1.88×**	**98.0%**
SDVG-hybrid (k=3, τ=−1.0)	✓	**0.0764**	**49.5**	**1.96×**	**96.9%**
Stitching depth (τ=−0.7, block 0 protected)					
k=2	×	0.0756	56.2	1.73×	95.9%
k=3	×	**0.0766**	54.3	1.79×	**97.2%**
k=4	×	0.0752	52.5	1.85×	95.5%
Stitching block 0 as well (τ=−0.7)					
k=2	×	0.0688	58.9	1.65×	87.3%
k=3	×	0.0677	51.6	1.88×	85.9%
k=4	×	0.0646	46.7	2.08×	82.0%
Varying the threshold (k=3, block 0 protected)					
τ=−0.7	×	0.0766	54.3	1.79×	97.2%
τ=−0.7	✓	0.0772	51.6	1.88×	98.0%
τ=−1.0	×	0.0763	51.7	1.88×	96.9%
τ=−1.0	✓	0.0764	49.5	1.96×	96.9%
τ=−1.5	×	0.0760	49.1	1.97×	96.4%
τ=−1.5	✓	0.0759	48.2	2.01×	96.3%
τ=−2.0	×	0.0758	49.3	1.97×	96.2%
τ=−2.0	✓	0.0751	46.1	2.11×	95.3%

Impact of Stitching Depth (k). Delegating the initial k denoising steps of a rejected block to the draft model introduces zero computational overhead, as those forward passes are executed during candidate evaluation. Crucially, the partially denoised latent state provides a more informative initialization than pure Gaussian noise. We find that $k = 3$ represents the optimal trade-off: it preserves 97.2% of target visual quality (0.0766 VisionReward) while accelerating inference to 1.79×. Smaller stitching depths ($k = 2$) leave excessive refinement work to the target model, reducing speed gains. Conversely, setting $k = 4$ delegates the entire 4-step denoising process to the

draft model, leaving the target model zero opportunity for refinement on rejected blocks and causing visual quality to drop to 0.0752.

Criticality of Protecting Anchor Block ($b = 0$). Extending step-level stitching to the anchor block ($b = 0$) induces severe visual degradation. At $k = 3$, stitching block 0 reduces VisionReward from 0.0766 to 0.0677, with quality deteriorating further as k increases. Because block 0 operates without prior context, it establishes the global visual composition and scene structure for the entire video stream. Allowing a truncated draft trajectory to define this anchor block introduces irreversible structural errors that compound through the shared KV cache across all subsequent blocks.

Efficiency of On-Demand Cascade Verification. Evaluating a 4-frame temporal proxy rather than all 12 decoded frames significantly reduces verification overhead without compromising decision accuracy. At $\tau = -1.0$, only 9.1% of candidate blocks land within the ambiguity margin $(\tau-\delta, \tau+\delta)$ and require full 12-frame evaluation. This sparse sampling reduces latency by 2.2 s per video (boosting speedup from $1.88\times$ to $1.96\times$) with zero loss in VisionReward (0.0764). The proportion of ambiguous blocks shrinks under more permissive thresholds (dropping from 11.5% at $\tau = -0.7$ to 3.6% at $\tau = -2.0$), making cascade savings most effective under conservative operating regimes. Across all evaluated thresholds, cascade routing yields latency reductions between 0.9 s and 3.2 s per video with negligible variations in visual quality (≤ 0.0007 VisionReward).

5 Limitations

Distributional Shift in Continuous Verification. Unlike classical LLM speculative decoding, which guarantees exact target distribution matching via discrete rejection sampling, `SDVG` operates in continuous space and permits a controlled distributional shift toward the draft model. While stricter acceptance thresholds systematically mitigate this shift, the trade-off remains bounded by the intrinsic visual quality gap between the draft and target architectures.

Proxy Scope of Frame-Level Reward Models. Our router relies on ImageReward, which evaluates individual frames independently in the spatial domain. Consequently, the scoring mechanism lacks explicit awareness of temporal consistency and complex motion dynamics across adjacent frames. Developing dedicated video-block quality estimators that explicitly capture spatiotemporal coherence represents a promising direction for refining block-level routing decisions.

Draft Compute Overhead on Rejections. When candidate blocks are rejected (or during mandatory target generation for the anchor block $b = 0$), the drafter's denoising steps and VAE decoding passes represent unrecovered computational overhead. Incorporating asynchronous speculative execution or lightweight VAE decoding pipelines could further amortize this latency penalty in practical serving systems.

6 Conclusion

We presented `SDVG`, a training-free speculative decoding framework that accelerates autoregressive video diffusion through dynamic, reward-guided model routing. By pairing a lightweight image quality router with worst-frame score aggregation and anchor-block protection, `SDVG` enables streaming synthesis while preserving visual fidelity. On MovieGenVideoBench, standard `SDVG` retains 98.1% of target-only quality at a $1.59\times$ speedup, and achieves up to $2.05\times$ speedup while preserving 95.9% of target quality. Building on this foundation, `SDVG`-hybrid integrates step-level trajectory stitching and on-demand cascade verification—reusing intermediate draft latents on rejected blocks and sparse frame evaluation to further eliminate compute redundancy. Together, these extensions advance the Pareto frontier to a $1.88\times$ speedup at 98.0% quality retention and a $1.96\times$ speedup at 96.9% retention. Because `SDVG` requires no retrainings or architectural modifications, it offers a practical, plug-and-play solution for adaptive inference-time compute allocation in modern video generation pipelines.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and to aid and polish the writing. We did not use generative AI tools to produce, alter, or select any experimental result reported here. Additionally, we used generative AI tools for research execution, including the code that generates our figures and runs the analyses. All AI-assisted work was reviewed and verified by the authors: every generated passage was checked against our own results and claims, and the code was tested against independently computed statistics. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023. URL https://arxiv.org/abs/2302.01318.

Shenggan Cheng, Yuanxin Wei, Lansong Diao, Yong Liu, Bujiao Chen, Lianghua Huang, Yu Liu, Wenyuan Yu, Jiangsu Du, Wei Lin, and Yang You. Srdiffusion: Accelerate video diffusion inference via sketching-rendering cooperation, 2025. URL https://arxiv.org/abs/2505.19151.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. URL https://arxiv.org/abs/2204.03458.

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

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding, 2023. URL https://arxiv.org/abs/2211.17192.

Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. Reward-guided speculative decoding for efficient llm reasoning, 2025. URL https://arxiv.org/abs/2501.19324.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2024. URL https://arxiv.org/abs/2310.11440.

Erwann Millon. Krea realtime 14b: Real-time video generation, 2025. URL https://github.com/krea-ai/realtime-video.

Zizheng Pan, Bohan Zhuang, De-An Huang, Weili Nie, Zhiding Yu, Chaowei Xiao, Jianfei Cai, and Anima Anandkumar. T-stitch: Accelerating sampling in pre-trained diffusion models with trajectory stitching, 2024. URL https://arxiv.org/abs/2402.14167.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Desen Sun, Jason Hon, Jintao Zhang, and Sihang Liu. Hybridstitch: Pixel and timestep level model stitching for diffusion acceleration. *arXiv preprint arXiv:2603.07815*, 2026.

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, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng

Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024.

Yuchen Xia, Divyam Sharma, Yichao Yuan, Souvik Kundu, and Nishil Talati. Modm: Efficient serving for image generation via mixture-of-diffusion models, 2025. URL https://arxiv.org/abs/2503.11972.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2304.05977.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation, 2026. URL https://arxiv.org/abs/2412.21059.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation, 2024. URL https://arxiv.org/abs/2311.18828.

Jintao Zhang, Jia Wei, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Haoxu Wang, Kai Jiang, Jun Zhu, and Jianfei Chen. Sageattention3: Microscaling fp4 attention for inference and an exploration of 8-bit training. *arXiv preprint arXiv:2505.11594*, 2025a.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Haocheng Xi, Jun Zhu, Jianfei Chen, et al. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In *Forty-second International Conference on Machine Learning*, 2025b.

Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Turbodiffusion: Accelerating video diffusion models by 100-200 times. *arXiv preprint arXiv:2512.16093*, 2025c.

A Experimental Setup Details

Base Models and Router Architecture. We evaluate our framework on a pair of autoregressive video diffusion models built on the Wan2.1 backbone (Wan et al., 2025). The *target model* $\mathcal{T}$ is Krea Realtime Video 14B (Millon, 2025), distilled from Wan2.1-T2V-14B via Self-Forcing (Huang et al., 2025). The *draft model* $\mathcal{D}$ is the Wan2.1-T2V-1.3B Self-Forcing model. Both architectures share a causal attention backbone with key-value (KV) caching driven by 3D RoPE positional embeddings, executing $S = 4$ denoising steps per block with the schedule $\mathbf{t} = [1000, 937, 833, 625, 0]$. For candidate verification, the router utilizes ImageReward (Xu et al., 2023), an off-the-shelf text-to-image reward model, to score decoded draft frames in the pixel domain.

Generation Protocol. Each synthesized video consists of $B = 9$ autoregressive blocks, corresponding to 27 latent frames in total (3 latent frames per block). The causal VAE decoder maps these to 9 pixel frames for the initial anchor block ($b = 0$) and 12 pixel frames for each subsequent block at 832×480 resolution. To guarantee strict reproducibility, all experimental runs use a fixed random seed (42).

Hardware and System Implementation. All experiments are conducted on two NVIDIA RTX A6000 GPUs (48 GB VRAM each). GPU 0 hosts the diffusion transformers ($\mathcal{T}$ and $\mathcal{D}$), while GPU 1 hosts the text encoder (UMT5-XXL), causal VAE decoder, and ImageReward router. To eliminate verification bottlenecks, we employ asynchronous CUDA streams that overlap cross-device tensor transfers and reward computation with GPU 0's denoising process, effectively hiding scoring latency.

Evaluation Benchmarks and Metrics. We construct our evaluation set from MovieGen-VideoBench (Polyak et al., 2024), comprising 1,003 text prompts spanning diverse visual categories (e.g., landscapes, animals, human activities, and cinematic sequences). Visual synthesis quality is evaluated using VisionReward (Xu et al., 2026) and EvalCrafter (Liu et al., 2024). System efficiency is measured by wall-clock latency per video (excluding model loading and warmup), from which relative throughput speedup over target-only generation is derived.

Baselines. We benchmark `SDVG` and `SDVG`-hybrid against two boundary baselines: (1) *Draft-only* ($\mathcal{D}$-only), where all blocks are generated exclusively by the 1.3B draft model (representing maximum speed but lower visual quality); and (2) *Target-only* ($\mathcal{T}$-only), where all blocks are generated by the 14B target model (representing upper-bound visual quality at full computational cost). `SDVG` and `SDVG`-hybrid operate along the Pareto frontier between these extremes by sweeping the quality acceptance threshold τ, the number of stitched denoising steps k, and the cascade margin δ.

B Additional Qualitative Results

Figures 4 and 5 repeat the comparison of Figure 1 on four further prompts drawn from the same evaluation. Each row is one method—draft-only, SDVG ($\tau{=}-0.7$), SDVG-hybrid ($k{=}3$, $\tau{=}-1.0$, cascade, block 0 left to the target) and target-only—and each cell shows four uniformly sampled frames of a single generated video, so a row can be read left to right as the video plays.

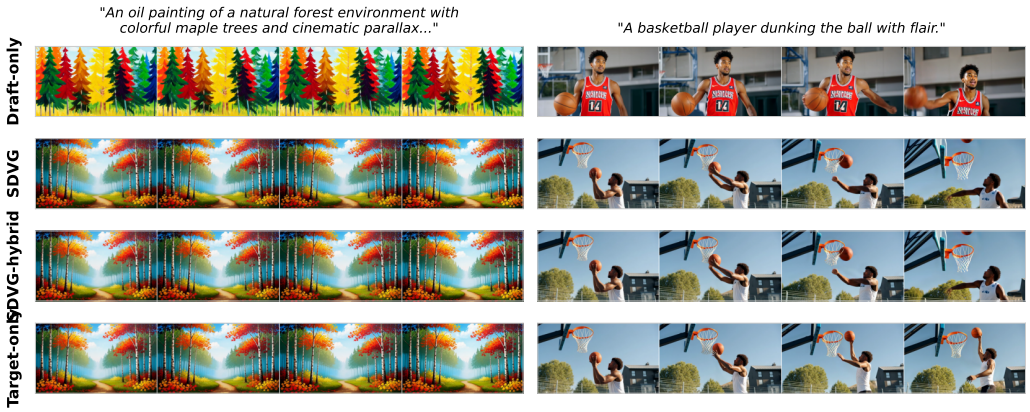

Figure 4: Qualitative comparison on two further MovieGenVideoBench prompts. The draft's maple forest drifts in composition and saturation while both SDVG variants hold the target's palette, and on the basketball clip the draft loses the hoop and the player's limbs entirely.

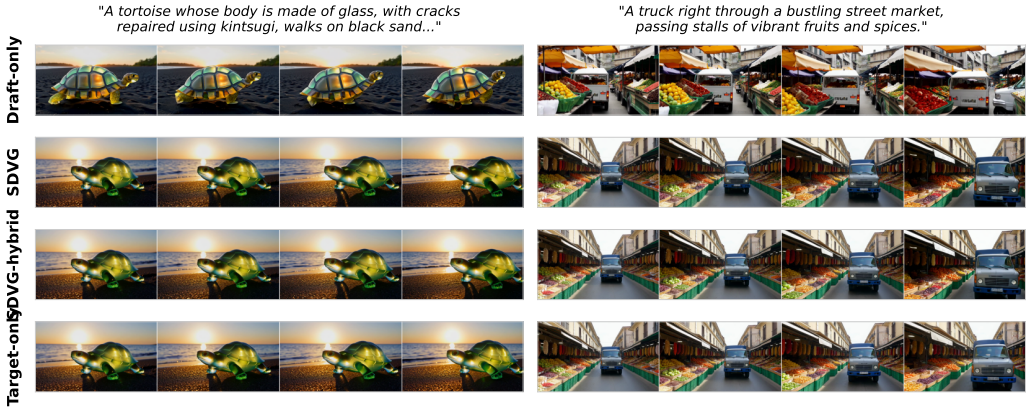

Figure 5: Qualitative comparison on two further MovieGenVideoBench prompts. These two stress fine detail: the kintsugi cracks on the glass tortoise and the density of the market stalls both degrade in the draft and recover in the regenerated blocks.