ASDSV: Multimodal Generation Made Efficient with Approximate Speculative Diffusion and Speculative Verification

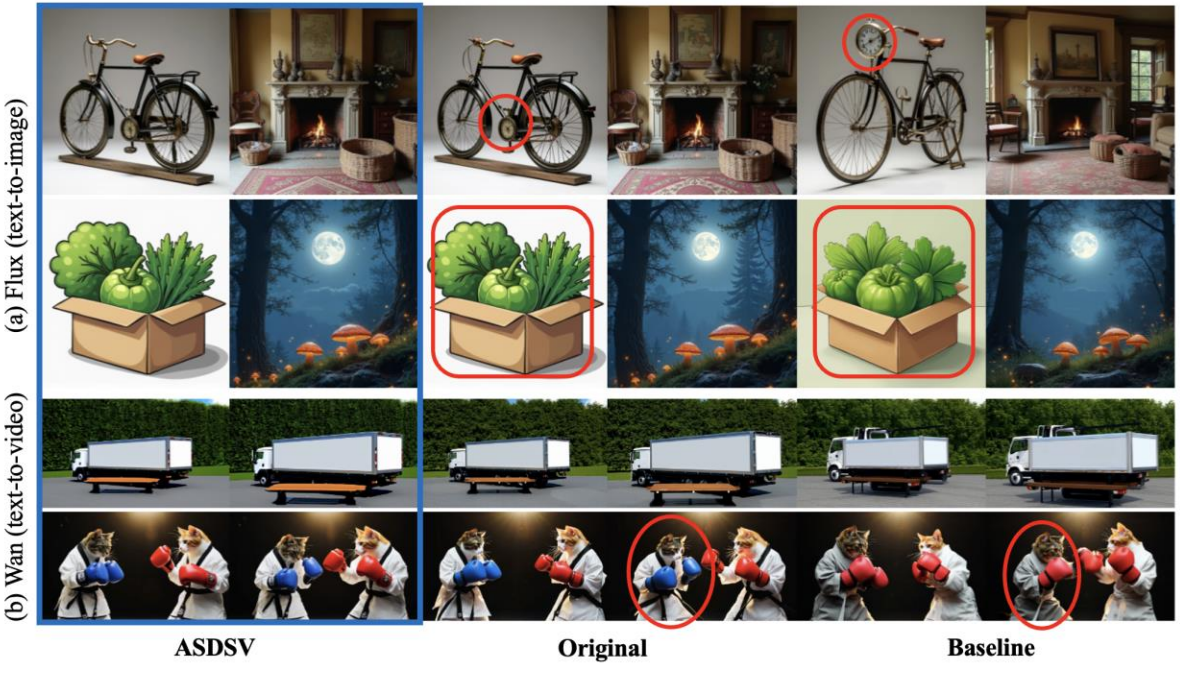
Kaijun Zhou, Xingyu Yan, Xingda Wei, Xijun Li, Jinyu Gu✉

Presenter: Kaijun Zhou

39-th Conference on Neural Information Processing Systems (NeurIPS 2025)

Motivation

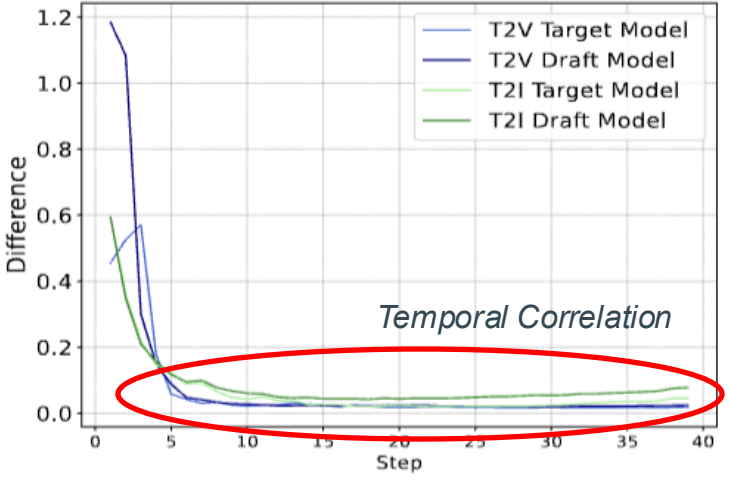
1. **High Latency**: Each inference requires tens to hundreds of sequential denoising steps and each step is computationally expensive (especially high resolution).
2. **Speculative Decoding** (SD): SD has emerged as a powerful paradigm for latency reduction in language model. It typically employs a lightweight draft model to predict tokens.

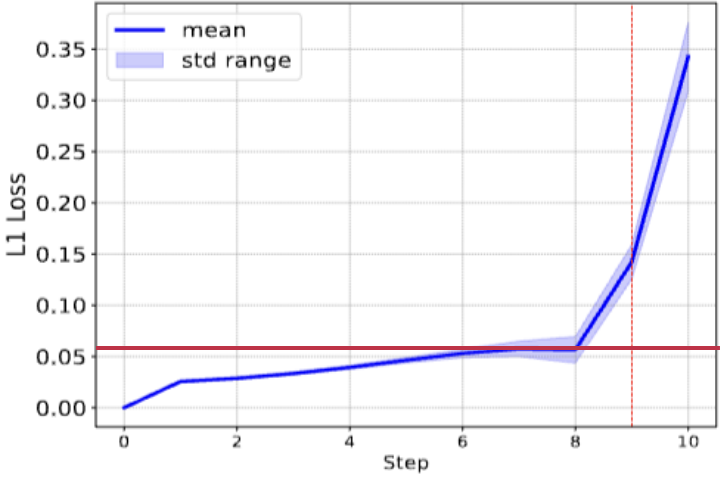


3. **Limitations of Existing Acceleration Methods:** Both *Distillation* and *Architectural optimizations* usually require extra training and are tightly coupled with model designs, while *Caching-based approaches* may lead to a noticeable degradation in generation quality (left).

Observation

1. **Availability of "off-the-shelf" draft models**: either as official variants released in different sizes or as community-compressed versions, facilitating the use of speculative diffusion (e.g., Wan2.1-1.3B for Wan2.1-14B and Stable-Diffusion 3.5's medium and large variants).



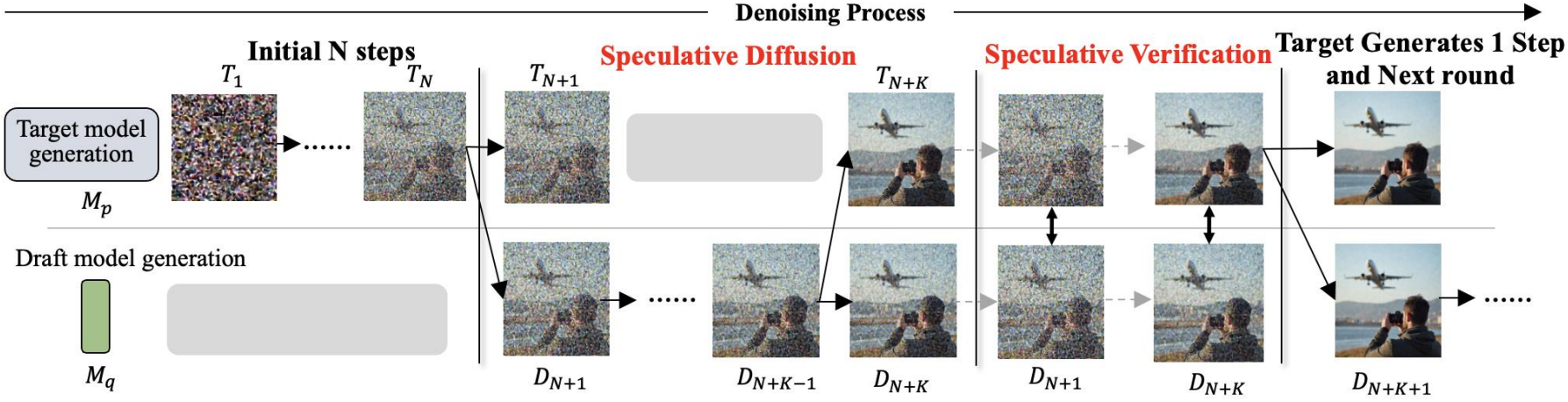


2. **Strong temporal correlation**: The left shows the output trends between draft and target models. This refers to the fact that the magnitude of change of the draft and the target model's output between consecutive steps is highly similar (Right: specifically <0.05 at early verification-steps). D_i and T_i denote the predicted outputs from the draft and target models at step i, respectively. Formalized as: i.e., $||D_i - D_{i-1}||_1 \approx ||T_i - T_{i-1}||_1$

Method #1: *ASDSV*

Workflow:

1. For each round of ASD, the draft model sequentially samples *K* steps, while the target model only executes for the first step and the last step.
2. The following SV process will calculate two differences between D_{N+1} and T_{N+1} and between D_{N+k} and T_{N+k} to determine if they are smaller than a threshold.



ASD — *Temporal Correlation (similar output trends)*

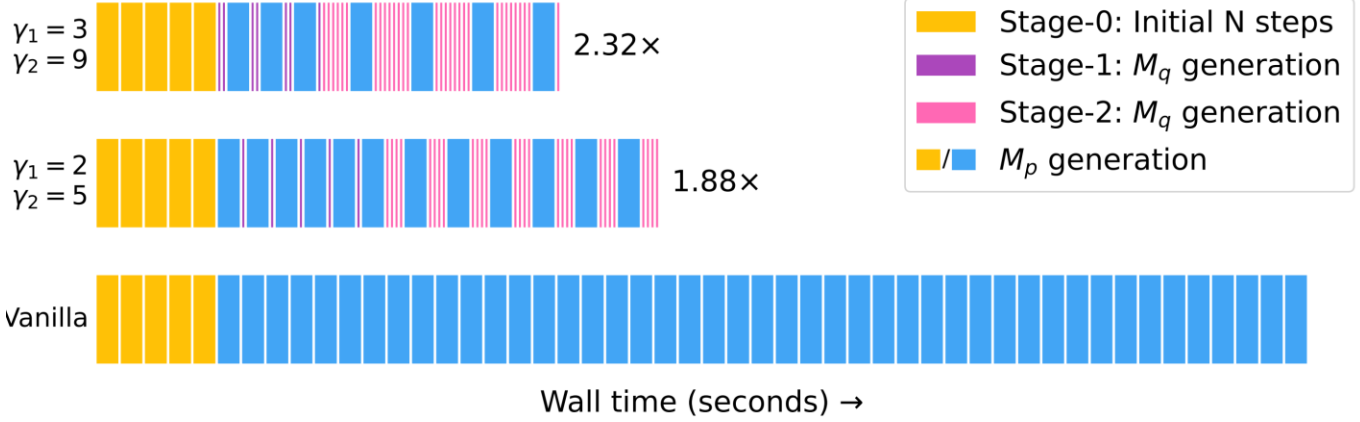
SV — *Speculatively Verify The First & Last steps*

} *speculatively concluded that all intermediate sample steps are verified*

Method #2: *Multi-Stage Speculative Strategy*

During the denoising process, the differences between adjacent steps are **larger in the early stage and smaller in the later stage**. This characteristic requires fine-grained control to balance generation speed and quality.

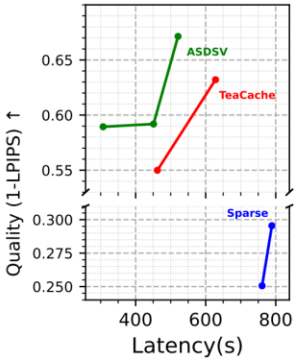
So **ASDSV** employs a three-stage strategy, the following is an example trace:
1. Stage-0: the initial N steps;
2. Stage-1: the subsequent warmup stage, employ a smaller γ_1 as K value to preserve a high success rate of verification during the early stage;
3. Stage-2: the main generation stage, employ a larger γ_2 as K value to achieve a higher acceleration ratio.

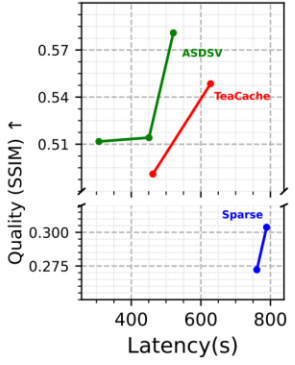


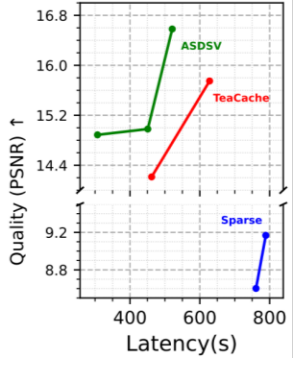
Experiment Results

Achieves up to 1.77X-3.01X speedup in e2e inference with a minimal 0.3%-0.4% drop in VBench score using Wan2.1 model.

Method	Efficiency			Visual Quality			
	FLOPs (P) ↓	Latency (s) ↓	Speedup ↑	VBench ↑	LPIPS ↓	PNSR ↑	SSIM ↑
Wan2.1 (81 frames, 832×480)							
Wan2.1 ($T = 50$)	168.1	924	1×	82.69	-	-	-
Sparse-fast	138.4	761	1.21×	78.97	0.75	8.6	0.27
Sparse-slow	153.1	789	1.17×	81.91	0.70	9.2	0.30
TeaCache-fast	84.2	462	2×	81.83	0.45	14.2	0.49
TeaCache-slow	114.4	628	1.47×	82.17	0.37	15.75	0.55
ASDSV-fast	**52.6**	**307**	**3.01×**	**82.36**	0.41	14.89	0.51
ASDSV-slow	74.4	521	1.77×	82.29	**0.33**	**16.58**	**0.58**







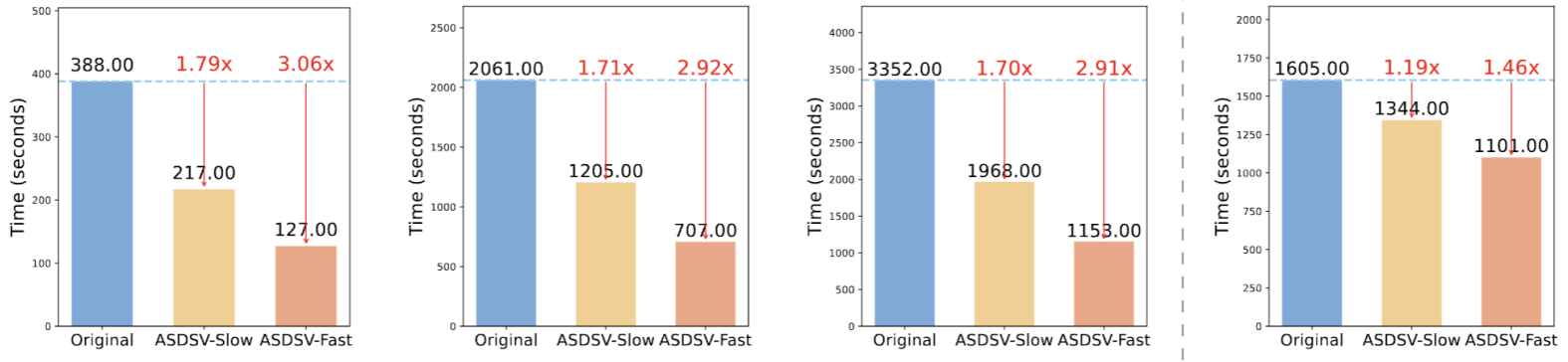
Experiment Results

1. Supports both image and video generation on different resolutions and lengths.
2. Maintains robust performance over multiple seeds.

Method	Efficiency			Visual Quality				
	FLOPs (T) ↓	Latency (s) ↓	Speedup ↑	FID ↓	CLIP ↑	LPIPS ↓	IS ↑	PSNR ↑
FLUX.1-dev (512×512)								
FLUX.1-dev ($T = 50$)	1082.5	9.18	1×	-	31.14	-	26.32	-
TeaCache-fast	330.4	**4.25**	2.16×	8.16	**31.34**	0.30	25.66	28.89
TeaCache-slow	566.8	5.73	1.6×	4.33	31.20	0.14	26.7	31.67
ASDSV-fast	**206.3**	6.25	1.47×	4.11	31.15	0.11	26.66	32.95
ASDSV-slow	292.3	7.3	1.26×	**3.90**	31.15	**0.10**	26.52	**33.14**

Images Metric	Ratio (Original / ASDSV)
CLIP	100.04% ± 0.10%
IS	103.84% ± 3.7%

Videos Metric	Ratio (ASDSV / Teacache)
LPIPS↓	88.02% ± 2.85%
PSNR↑	106.10% ± 1.25%
SSIM↑	104.95% ± 1.84%



(a) 480P, 41 frames (b) 480P, 141 frames (c) 720P, 81 frames (d) 1024×1024, 1 frame

Text-to-Image Visual Results

Text-to-Video Visual Results

Original | ASDSV-slow | TeaCache-slow | Sparse-slow

Thank you!

Kaijun Zhou
zhoukaijun@sjtu.edu.cn

Jinyu Gu
gujinyu@sjtu.edu.cn

School of Computer Science, Shanghai Jiao Tong University