A Additional Experiments

A.1 Results of Draft Model Comparison

In autoregressive language models, draft models (M_q) are typically selected from existing off-the-shelf smaller transformers with the same architecture as the target model (M_p). For example, T5-XXL (11B) as M_p and T5-large (800M) as M_q. Following this practice, we evaluate the performance of different draft models on text-to-image generation. Flux.1-dev is the state-of-the-art text-to-image model, and Flux-lite (distilled model) and Flux-SVD (quantized model) are two efficient draft variants of Flux.1-dev.

Tab 1 illustrates that SDSV consistently outperforms both draft models, which demonstrates the effectiveness and broad applicability of SDSV across diverse compression techniques. The quantized model Flux-SVD shows the best overall performance and is chosen as the draft model for Flux.1-dev in our experiments. FID, LPIPS and PSNR are reference-based metrics for measuring the similarity between generated results and Flux.1-dev original outputs. CLIP score measures the semantic alignment between the input prompt and the generated images and IS measures the diversity of the generated images.

Table 1: Performance of FLUX.1-dev compared with different draft models. *SDSV w/ Flux-lite* and *SDSV w/ Flux-SVD* are SDSV that use Flux.1-dev as the target model and Flux-lite and Flux-SVD as draft model, respectively.

Model	**FLUX.1-dev**				
Score	**FID** ↓	**CLIP** ↑	**LPIPS** ↓	**IS** ↑	**PSNR** ↑
Flux.1-dev	—	31.14	—	26.32	—
Flux-lite	9.51	31.07	0.40	25.59	29.73
Flux-SVD	5.52	**31.18**	0.19	26.26	30.53
SDSV w/ Flux-lite	6.62	31.20	0.22	25.68	30.49
SDSV w/ Flux-SVD	**4.97**	**31.24**	**0.11**	**26.36**	**32.69**

Additionally, we improved the draft model baseline by combined the Stage-0 into the draft model diffusion process. SDSV-slow and SDSV-fast are the two variants of SDSV with different settings, detailed in Experiments Settings of the main paper. Specifically, we use the target model to generate the first $N = 5$ steps as the draft model's initial steps (the same as the SDSV-fast), and then use the draft model to generate the remaining steps. The evaluation metric results are shown in Table 2. The visual results are shown in Figure 1. SDSV outperforms the Improved-Draft baseline in both visual quality metric (VBench score) and similarity metrics (LPIPS, PSNR, and SSIM).

Table 2: Efficiency and visual quality comparison of SDSV and improved draft model baseline on text-to-video generation.

Method	**Efficiency**			**Visual Quality**			
	FLOPs (P) ↓	**Latency (s)** ↓	**Speedup** ↑	**VBench** ↑	**LPIPS** ↓	**PNSR** ↑	**SSIM** ↑
Improved-Draft	28.0	256	3.67×	82.23	0.42	14.70	0.50
SDSV-fast	52.6	307	3.01×	**82.36**	0.41	14.89	0.51
SDSV-slow	74.4	521	1.77×	82.29	**0.33**	**16.58**	**0.58**

A.2 Analysis of Multi-Stage Speculative Strategy

SDSV employs a multi-stage speculative strategy to balance the trade-off between generation speedup and performance: first N initial steps (Stage-0) combined with the subsequent Stage-1 serve as the warmup period for the diffusion process, establishing the foundation for high-quality generation, followed by the main generation stage as Stage-2. A key characteristic of the warmup period is the relatively large variation between outputs of adjacent steps, unlike the highly similar outputs observed between consecutive steps in Stage-2. To ensure higher quality model outputs, modifications to the architecture are typically avoided during the warmup period.

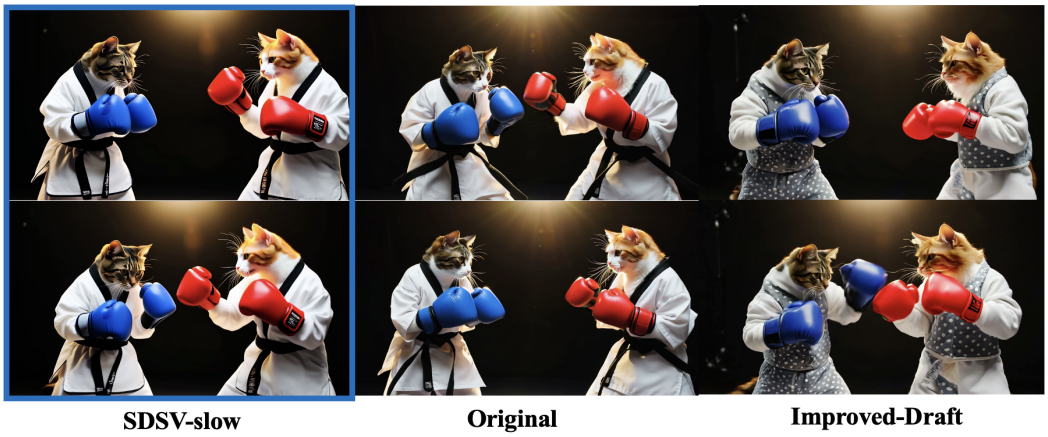

Figure 1: Visual comparisons of improved draft model baseline on text-to-video generation.

We evaluated two warmup approaches: the 2 steps method and a more conservative fixed ratio of 20% of total diffusion steps. After completing their respective warmup phases, they employ the same speculative strategies as SDSV in Stage-2. Our experimental results in Figure 2 reveal a clear visual difference. Configuration (a) with minimal warmup steps significantly reduces target model invocations, improving speed but compromising generation quality. In contrast, configuration (c) with larger warmup steps maintains higher quality output but with reduced acceleration. To balance this trade-off, we divided the warmup phase into two distinct components (Stage-0 and Stage-1): first allowing the target model to establish critical structural elements, then employing small speculative steps to accelerate the remaining warmup procedure. As demonstrated in configuration (d), which uses 2 steps for Stage-0 followed by speculative steps with $K = 3$ for Stage-1, then employing the same $K = 9$ speculative steps for Stage-2 as configurations (a) and (c). This multi-stage approach achieves nearly 2.1× acceleration over the vanilla diffusion process while preserving better visual quality compared to configuration (a).

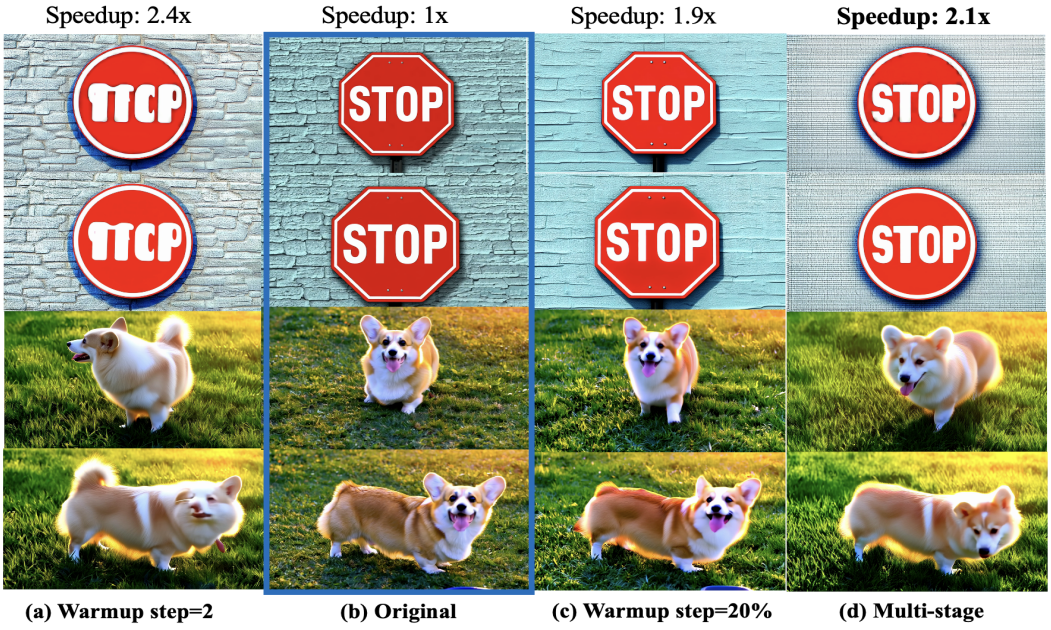

Figure 2: Comparison of different warmup and stage strategies: (a) minimal warmup with 2 steps followed by SDSV Stage-2, (b) vanilla diffusion process, (c) conservative warmup with 20% of total steps (10 steps) followed by SDSV Stage-2, (d) proposed multi-stage strategy with 2 steps for Stage-0 and 8 steps for Stage-1 (20% total diffusion steps as warmup), followed by the same SDSV Stage-2.

B Social Impact

The acceleration of diffusion models provided by SDSV reduces computational resources and latency, improving real-time applicability of state-of-the-art diffusion models while promoting environmental sustainability through reduced energy consumption. However, it is important to note that SDSV focuses primarily on efficiency gains and does not address inherent challenges such as privacy, bias, and fairness in the underlying diffusion models.

C Additional Visualization

We provide comprehensive visual comparisons between SDSV and baseline acceleration methods for both text-to-image and text-to-video generation, as shown in Figure 3 and Figure 4. Extensive results demonstrate the superior visual fidelity of SDSV across different generation scenarios.

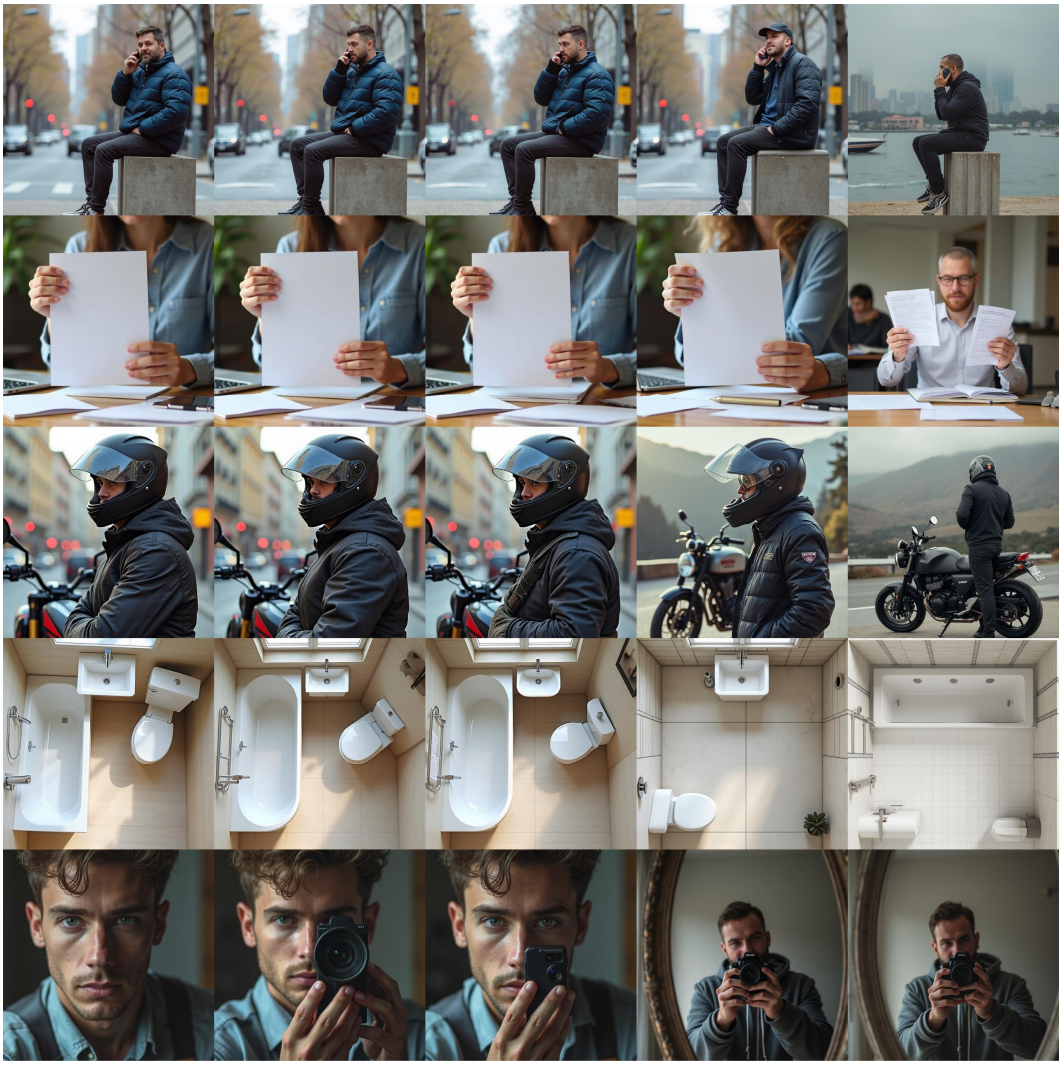

Figure 3: Comparison of different accelerating methods on text-to-image generation using Flux.1-dev at 512x512 resolution.

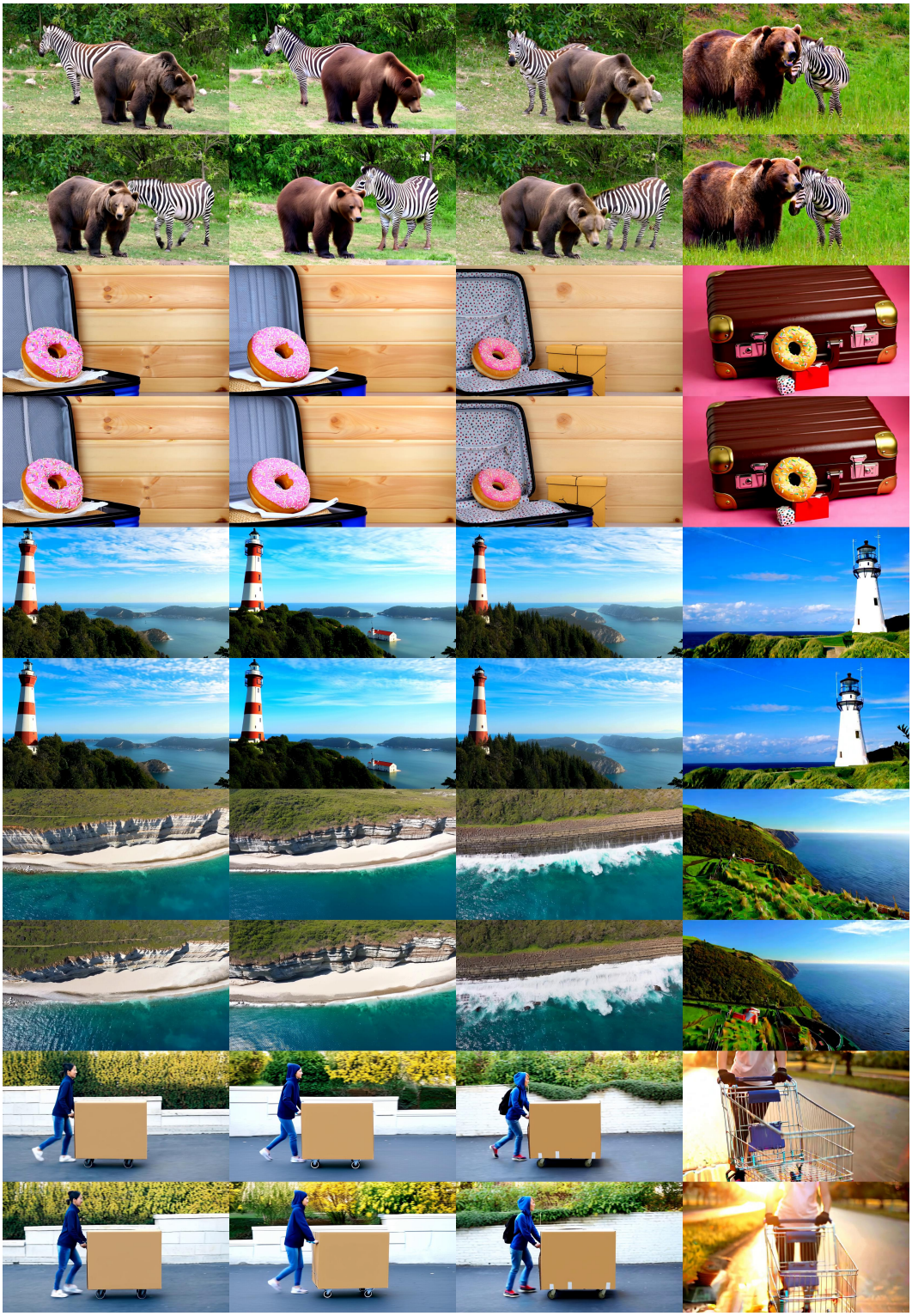

Figure 4: Comparison of different accelerating methods on text-to-video generation using Wan2.1 at 480P resolution with 81 frames.