VVS: Accelerating Speculative Decoding for Visual Autoregressive Generation via Partial Verification Skipping

Supplementary Material

1. Implement Details

We present the pseudocode of the VVS framework in Algorithm 2 to further illustrate our design.

Algorithm 2 VVS with Partial Verification Skipping

Require: $\wp$: text prompt; $\mathcal{M}_T$: target model; $\mathcal{M}_D$: drafter model; L: max generated length; V_{last}: whether last step was verified
Ensure: Generated token sequence $\mathcal{S}$ for decoding to image

Initialize feature cache $\mathcal{C}$ and prefill $\wp$ into $\mathcal{M}_T$
$\mathbf{a} \leftarrow \mathcal{M}_D(\wp)$ ▷ initial accepted candidate
$V_{\text{last}} \leftarrow$ False
for $t = 0 \dots L-1$ **do**
  $f_t \leftarrow \text{Retrieve}(\mathcal{C}, \mathbf{a})$ ▷ retrieve features for drafting
  $C_t \leftarrow \mathcal{M}_D(f_t, \mathbf{a})$ ▷ draft candidates
  $\mathbf{skip} \leftarrow \text{Decision}(C_t, t, V_{\text{last}})$ ▷ Algorithm 1
  if $\neg$ **skip then**
    $\mathbf{logits}_t, \mathbf{h}_t \leftarrow \mathcal{M}_T(C_t)$
    $\mathbf{a} \leftarrow \text{Verify}(\mathbf{logits}_t, C_t)$ ▷ verification
    $\mathcal{C} \leftarrow \text{UpdateCache}(\mathbf{h}_t)$ ▷ cache new features
    $V_{\text{last}} \leftarrow$ False
  else
    $\mathbf{a} \leftarrow \text{UniformSample}(C_t)$ ▷ verfication-free
    $V_{\text{last}} \leftarrow$ True
  end if
  $\mathcal{S} \leftarrow \mathcal{S} \circ \mathbf{a}$ ▷ concatenate output
end for
return $\mathcal{S}$

2. Supplementary Results of Drafting Stage Analysis

In Tab. 4, the results demonstrate that after substituting the verification outcomes, both the acceleration performance and generation quality of SD remain highly stable. Tab. 5 and Tab. 6 further illustrate the impact of various staleness features, highlighting their reusability.

3. Prompts used in Qualitative Experiment

- A vast desert landscape under a starry sky, with a single tent illuminated by a warm campfire.
- A futuristic glass train speeding through a cherry-blossom forest at sunset, petals swirling in its slipstream.
- A rustic farmhouse kitchen with fresh sunflowers in a mason jar, warm afternoon light streaming through open windows.

Table 4. Verification redundancy experiment. r represents the proportion of verified results that are replaced for all iterations. We replace the verified tokens of the target model with the same number of tokens from the token path with the highest confidence score.

r	TPF	FID (↓)	CLIP Score (↑)	Precision (↑)	Recall (↑)
0%	2.27	26.63	0.3223	0.5370	0.5962
25%	2.27	26.36	0.3218	0.5344	0.6118
33%	2.27	26.23	0.3225	0.5334	0.6120
50%	2.27	26.55	0.3225	0.5364	0.6128
100%	2.25	26.19	0.3220	0.5502	0.6158

Table 5. Feature reusability experiment results. s denotes the extra staleness introduced for cached features; $s = -1$ indicates using fresh features for each step; $s = 0$ indicates using the most recent features cached from prior steps; $s = +i$ indicates using features with additional staleness i compared with $s = 0$.

s	τ	FID (↓)	CLIP Score (↑)	Precision (↑)	Recall (↑)
-1	2.27	26.53	0.3217	0.5328	0.6126
0	1.66	24.13	0.3215	0.5354	0.6685
+1	1.57	23.80	0.3210	0.5352	0.6990
+2	1.52	23.96	0.3210	0.5260	0.6850
+3	1.49	23.89	0.3206	0.5324	0.6950

Table 6. Feature blending experiment results. s denotes the extra staleness for cached features; s_1: staleness of type 1 features, s_2: staleness of type 2 features; $s = -1$ indicates using fresh features for each step; $s = 0$ indicates using the most recent features cached from prior steps; MAL: mean accept length.

s_1	s_2	MAL	FID (↓)	CLIP Score (↑)	Precision (↑)	Recall (↑)
-1	-1	2.27	26.53	0.3217	0.5328	0.6126
-1	0	1.93	25.17	0.3223	0.5402	0.6516
-1	1	1.89	25.14	0.3218	0.5336	0.6588
-1	5	1.86	24.65	0.3215	0.5220	0.6722
0	0	1.66	24.13	0.3215	0.5354	0.6685

- A close-up of a wolf with intense, focused eyes and thick gray fur, staring directly at the camera, set against a blurred forest background.
- A surreal landscape with giant, floating islands connected by glowing energy bridges, all under a purple sky with two moons.
- A young man with short curly hair wearing a denim jacket, looking thoughtfully into the distance under city lights.
- An astronaut standing on the desolate surface of Mars,

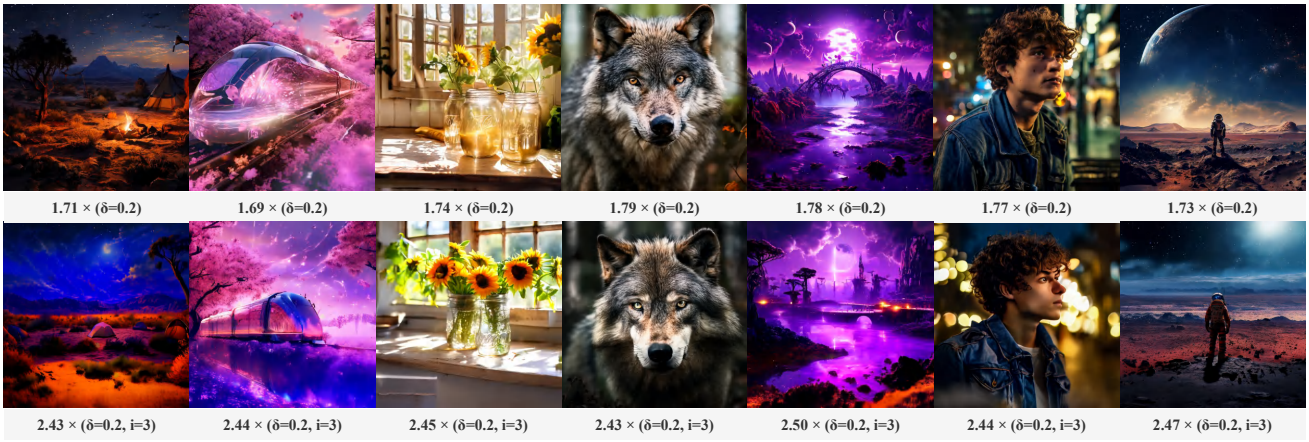
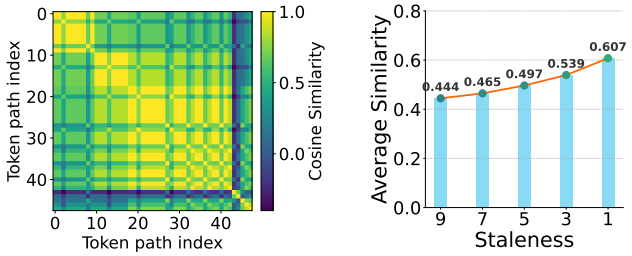


Figure 9. Qualitative comparison of results between the acceptance-relax-based SD framework (upper) and ours (lower) on the Lumina-mGPT model. δ denotes the probability threshold for relaxed acceptance, while i is the interval between two verification-free steps.



(a) (b)

Figure 10. For the Lumina-mGPT model: (a) Token path similarity within a candidate token tree. (b) Token-level feature similarity during the SD process. (c) Qualitative results and speedup.

gazing at the small, distant blue Earth in the black, star-filled sky.

4. Additional Experiments on Generalization

We further validated our VVS framework on the Lumina-mGPT model. Fig. 9 offers a visual demonstration of the resulting image quality, using the same prompts as in Sec. 3. The token path similarity is also relatively high and feature similarity also exists during the SD process, as shown in 10. Under the same relaxation threshold $\delta = 0.2$, VVS markedly cuts the target model's forward passes while preserving generation fidelity. This confirms the broad applicability of our approach across models. Due to the interchangeability of visual tokens, we believe VVS can be leveraged to accelerate more visual AR models.

5. System Overhead

Computation Overhead: After applying the downsampling strategy (§ 4.4 & Alg. 1) and `torch.jit` optimization, the similarity s computation accounts for $< 10\%$ of AR decoding for Lumina-mGPT-7B (6s vs. 70s on a single RTX4090), and this overhead can be fully covered by the latency reduction of VVS. The depth of drafted token tree is constant and independent of the scale of target model, this overhead ratio diminishes as the model grows, demonstrating excellent scalability. Moreover, VVS-U supports rule-based scheduling with negligible overhead ($< 1\%$).

Memory Overhead: the constraint on consecutive skips allows VVS maintain only minimal cached features ($n = \max(L_i) = 5$, which is a constant). For LlamaGen, with the latest 5 features of 1280 dimensions stored in BF16, the cache footprint is under 100 kB, which is negligible on modern GPUs sporting 24 GB or more of memory.

6. Discussion of Quantitative Results

Employing the relaxed acceptance of SD for image AR generation introduces a lossy approximation. However, this trade-off is somehow favorable due to: **i)** stochastic sampling nature: visual AR generation typically operates with a high temperature, implying that the model inherently explores a diverse distribution rather than a deterministic path. **ii)** visual token interchangeability: As noted in GSD and LANTERN, multiple diverse tokens can convey equally valid visual semantics. Therefore, while the relaxation introduces uncertainty δ (noise), this noise primarily results in switching between semantically similar visual tokens (e.g., slight variations in texture) rather than catastrophic semantic errors. Thus, SD schemes can even surpass the vanilla AR baseline on certain metrics. By omitting verification, VVS amplifies noise and, in certain configurations, leads to lower recall and other performance metrics.