Supplementary Materials

A. Proof of proposition 1

$$
\begin{aligned}
\alpha_{p,q} &= \mathbb{E}_{q(x)}\left[\min\left(1, \frac{p(x)}{q(x)}\right)\right] \\
&= \sum_x q(x) \min\left(1, \frac{p(x)}{q(x)}\right) \\
&= \sum_x \begin{cases} p(x), & \text{if } p(x) < q(x) \\ q(x), & \text{otherwise} \end{cases} \\
&= \sum_x \min(p(x), q(x)) \\
&= \sum_x \frac{p(x) + q(x) - |p(x) - q(x)|}{2} \\
&= \frac{1}{2}\left(\sum_x p(x) + \sum_x q(x) - \sum_x |p(x) - q(x)|\right) \\
&= \frac{1}{2}\left(1 + 1 - \sum_x |p(x) - q(x)|\right) \\
&= 1 - \frac{1}{2}\sum_x |p(x) - q(x)| \\
&= 1 - TV(p, q)
\end{aligned}
$$

B. Full algorithm of GSD

In this section, we provide detailed pseudocode for the implementation of GSD in Algorithm 1 and 2. Since GSD is built upon SJD, it operates by simply replacing the VERIFY(·) part with our VERIFY_GSD(·). For a more detailed implementation, please refer to the attached code.

Algorithm 1 Grouped Speculative Decoding

Require: Speculative Length L, maximum seq length N, expert model p_θ, initial context $X_{0:n_0}$, Group size G, Embedding distance matrix M_d, thresholds d, δ

$k \leftarrow L, n \leftarrow n_0$
while $n < N$ **do**
  $q_{L-k:L}, \hat{X}_{L-k:L} \sim \texttt{Rand-init}(\cdot)$
  parallel for $j = 0$ to L ▷ Parallel Verify
    $p_j \leftarrow p_\theta(\cdot \mid [X_{0:n}, \hat{X}_{0:j}])$
  end for
  $(\hat{X}_{0:k},\ k) \leftarrow \text{VERIFY\_GSD}(\hat{X}_{0:L}, p_{0:L},\ q_{0:L}, G)$
  $X_{n:n+k-1} \leftarrow \hat{X}_{0:k}$ ▷ Accept.
  $q_{0:L-k} \leftarrow p_{k:L}, \hat{X}_{0:L-k} \leftarrow \hat{X}_{k:L}$ ▷ Draft update
  $n \leftarrow n + k$
end while
return X

Algorithm 2 VERIFY_GSD(X,p,q,G)

Require: Draft $\hat{X}_{0:L}$, Verifier : $p_{0:L}(\cdot)$, Drafter : $q_{0:L}(\cdot)$, Group size G, Embedding distance matrix M_d, thresholds d, δ

for $k = 0$ to L **do**
  $p\_sort_{vals}, p\_sort_{idx} \leftarrow \texttt{sort}(p_k)$
  $idx \leftarrow \texttt{find-idx}(p\_sort_{vals}, p_k(\hat{X}_k))$

  $C_{idxs} \leftarrow p\_sort_{idx}[idx - G//2 : idx + G//2]$
  $Cvals \leftarrow p_k[C_{idxs}]$

  for cv, ci in $[Cvals, C_{idxs}]$ **do**
    if $|cv - p_k(\hat{X}_k)| > \delta$ then C_{idxs}.pop(ci)
    if $M_d[\hat{X}_k, ci]$ >d then C_{idxs}.pop(ci)
  end for

  $p'_C \leftarrow \texttt{sum}(p_k[C_{idxs}])$
  $q'_C \leftarrow \texttt{sum}(q_k[C_{idxs}])$
  if not $r \sim \mathcal{U}[0,1] \leq \min\left(1, \frac{p'_C}{q'_C}\right)$ **then**
    $x \sim [p_k - q_k]_+$, $\hat{X}_k \leftarrow x$, **break**.
  end if
end for
return $\hat{X}_{0:k}, k$

C. Additional Results

In this section, we present additional experiments expanding upon the visualizations discussed in the main text.

Top-1 probabilities In Fig. 1, we illustrate the visualization of Top-1 probabilities across a wider variety of images. As shown, regardless of the prompts, many images exhibit numerous tokens with low Top-1 probability distributions.

Visual quality comparison In Fig. 3, we visually illustrate the differences in generation quality among various methods compared in Table 1. As shown in the figure, our GSD achieves approximately a 4× speed-up while maintaining generation quality comparable to lossless methods such as vanilla AR and SJD. In contrast, the naive lossy method also achieves acceleration but significantly degrades generation quality.

GSD generation performance Fig. 2 presents further qualitative results of our method when accelerated by an average factor of 3.6. As demonstrated in the figure, our GSD significantly accelerates AR image decoding while maintaining generation quality across diverse prompts.

D. Prompts on Qualititave Experiment

In Figure. 9 on main paper, the prompts for each images are as follows :

- *Rusty robot on a skateboard in the hallway of domitory, photography, 4k, realistic, detailed, bright*
- *Origami astronaut, waliking in the cloud, bright background, realistic, 4k, photography, bright color*
- *photography, realistic, White cute fluffy dog, skyblue background, very intricate, very detailed, realistic., bright*
- *color photo, photography, Face of a young man, very detailed, realistic. sharp, film grain, high contrast*
- *animation art work, cute, cat character, bright color pallette*

ICCV
#8426

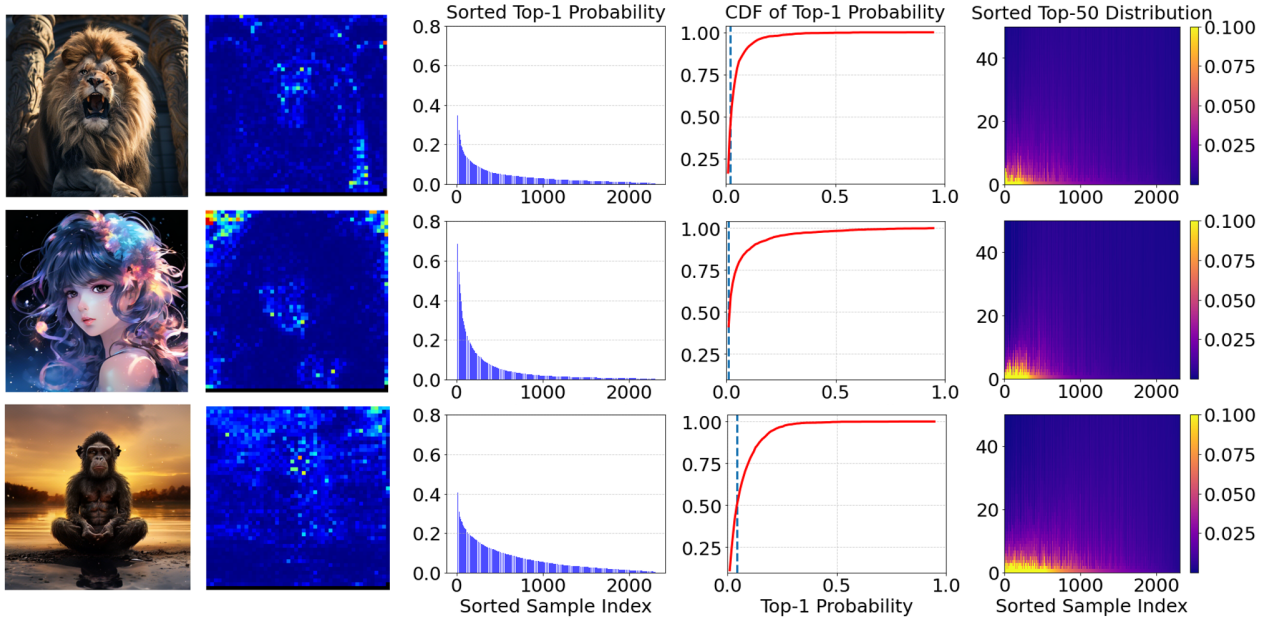

Figure 1. Additional p(x) visualization.

Figure 2. Qualtitiave experiment on various prompt. Our GSD shows on average 3.6x NFE acceleration while maintaing image quality

ICCV
#8426

Figure 3. Qualitative comparison between methods in Table 1 of the main paper