LISA: Locality-Informed Speculative Decoding for Accelerating Autoregressive Image Generation

Ying Li[1*], Siyong Jian[1*], Zhaode Wang[2], Zhiwen Chen[2], Chengfei Lv[2], and Huan Wang[1†]

[1] Westlake University, China
[2] Alibaba Group, China

Project Page: https://neuraliying.github.io/LISA

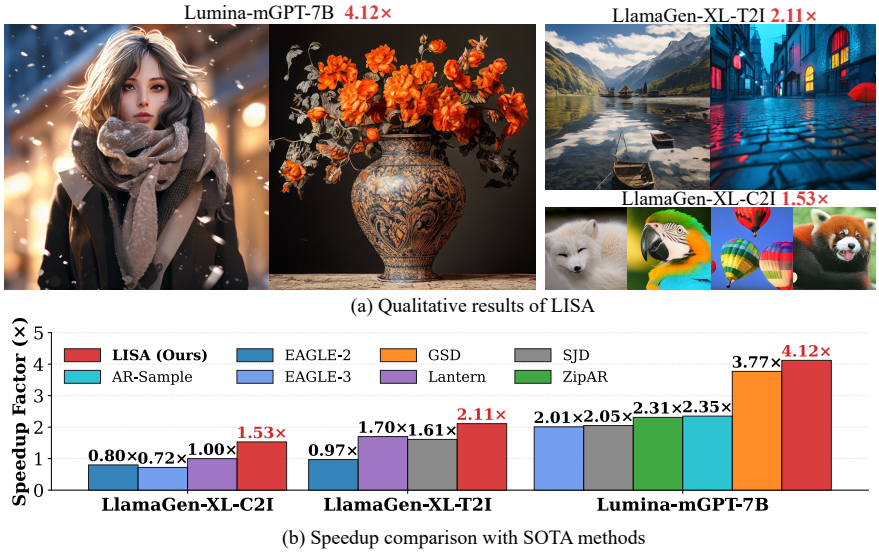


Fig. 1: Qualitative results and speedup performance of LISA. (a) Qualitative results across different models and tasks. LISA maintains high structural fidelity and aesthetic quality while significantly reducing inference latency (e.g., achieving **4.12**× acceleration on Lumina-mGPT-7B). (b) Speedup performance compared with SOTA methods. LISA is evaluated across text-to-image (T2I) and class-to-image (C2I) benchmarks, demonstrating speedup gains across tasks and model architectures.

Abstract. Autoregressive (AR) models deliver high-quality image generation but suffer from severe inference latency due to sequential decoding. While speculative decoding (SD) successfully accelerates large language models, we demonstrate that state-of-the-art methods like EAGLE-3 provide limited gains for visual AR models. We attribute this to the intrinsic nature of visual tokens: unlike text, their probability distributions are flat and non-discriminative, which hinders drafter–target alignment and causes verification to collapse under sampling. To bridge this

* These authors contributed equally.
† Corresponding author: Huan Wang (wanghuan@westlake.edu.cn).

gap, we propose **LISA**, a **L**ocality-**I**nformed **S**peculative framework for **A**utoregressive image generation. LISA overcomes alignment barriers via two key components: (i) **Locality-Informed Distillation**, which aligns the drafter by prioritizing supervision on structured uncertainty; and (ii) **Geometry-Aware Soft Verification**, which leverages embedding proximity and target confidence to enable more permissive yet safe token acceptance. Experiments across T2I and C2I tasks demonstrate that LISA improves the speed–quality trade-off, achieving up to **4.12**× acceleration and outperforming recent competitive methods.

Keywords: Autoregressive Image Generation · Speculative Decoding · Locality Geometry · Distribution Alignment

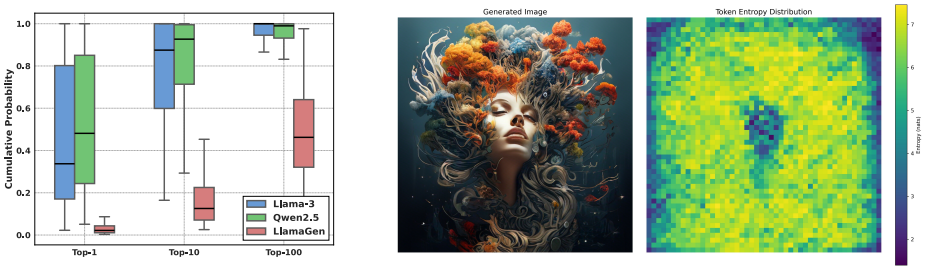
1 Introduction

Autoregressive (AR) image generation has emerged as a powerful paradigm [12, 16, 28], leveraging the architectural advances and scaling laws of large language models (LLMs) [4, 10, 30]. By predicting visual tokens sequentially, these models produce high-fidelity images and support unified multimodal workflows [38]. However, sequential decoding incurs substantial inference overhead [13, 35, 43], limiting practical deployment despite recent progress. **Speculative decoding** accelerates AR models by drafting multiple tokens and verifying them in parallel with the target model [15]. It achieves substantial speedup in LLMs without degrading output quality, and is widely adopted in recent high-performing systems [18,19,41]. However, applying speculative decoding to AR image generation remains challenging: visual tokens are **flatter and non-discriminative**, leading to **low acceptance rates** and unreliable verification [11].

We first investigate the distributional differences between visual and language tokens that hinder speculative decoding. As shown in Fig. 2(a), while language tokens exhibit "spiky" distributions with mass concentrated on a few candidates, visual tokens—quantized from continuous signals—exhibit much flatter profiles. Fig. 2(b) further reveals that prediction entropy is spatially heterogeneous, peaking at complex textures while remaining low in high-prior regions. Crucially, high-quality generation requires stochastic sampling (e.g., $T>0$, Top-K), which exacerbates drafter–target *misalignment* by further dispersing probability mass. This non-discriminative, spatially varying nature, compounded by sampling stochasticity, **challenges the core assumption** of strict token-level alignment in standard speculative decoding. These observations lead to a fundamental question that defines our study:

Key Challenge: *How can we mitigate the drafter–target misalignment inherent in visual autoregressive models to achieve inference acceleration?*

To address this challenge, we introduce **LISA**, a speculative decoding framework tailored to the empirically observed distributional and geometric properties of visual tokens. LISA integrates two complementary components: (i) a *locality-informed distillation* strategy that strengthens drafter–target alignment by exploiting the local geometry of the visual codebook; and (ii) a *geometry-aware soft*



(a) Dispersed token probability distribution (b) Spatial uncertainty map

Fig. 2: The non-discriminative and spatially heterogeneous nature of visual tokens. (a) **Dispersed token probability distribution.** Evaluated on 100 prompts from LongGenBench [23] for LLMs [6, 37] and 100 random classes for LlamaGen [28], visual tokens exhibit a flatter cumulative distribution across Top-N values, indicating higher intrinsic uncertainty. (b) **Spatial uncertainty.** For a sample from Lumina-mGPT, entropy is highly heterogeneous; uncertainty peaks at complex textures and boundaries while remaining lower in regions with strong priors like the central face.

verification (GASV) mechanism that adaptively modulates acceptance based on geometric distance and target uncertainty. Our method achieves encouraging results compared with recent competitive visual speculative decoding approaches, while jointly improving drafter alignment during training and verification during inference. Unlike prior works such as SJD [31], ZipAR [9], and GSD [27], which primarily improve inference-time decoding or verification, our approach additionally improves drafter–target alignment during training to enhance decoding efficiency under sampling.

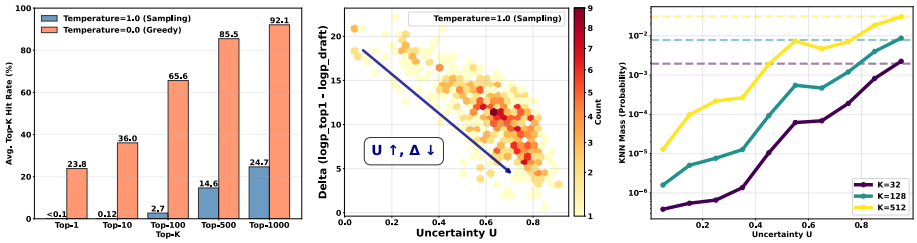
Our main contributions are summarized as follows:

- **Locality-Informed Distillation:** We propose a distillation method that leverages the geometry structure of visual tokens to improve drafter–target alignment in high-entropy regions and reduce prediction discrepancies.
- **Geometry-Aware Soft Verification:** We design GASV, a mechanism that replaces rigid pointwise verification with geometric and probabilistic cues, improving token acceptance under stochastic sampling.
- **Competitive Performance:** Experiments on various models demonstrate that LISA achieves up to **4.12**× speedup, improving over recent leading methods while maintaining competitive generation quality.

2 Related Work

2.1 Autoregressive Image Generation

Autoregressive (AR) models have emerged as a dominant paradigm due to their ability to capture complex joint distributions through conditional next-token or next-scale prediction [25, 32]. In this paradigm, raster-scan models like LlamaGen [28] and Janus-Pro [4] perform sequential prediction on flattened 2D tokens,



(a) Acceptance rate comparison (b) Uncertainty correlation (c) U-KNN mass relationship

Fig. 3: Diagnostic analysis of the misalignment problem in visual speculative decoding. (a) **Acceptance Rate Collapse:** Performance of EAGLE-3 [19] collapses in sampling mode, indicating that existing verification criteria are incompatible with the diffuse nature of visual token distributions. (b) **Uncertainty-Δ Correlation:** Heatmap of entropy U vs. log-probability gap Δ. The negative correlation reveals that higher target uncertainty leads to smaller probability discrepancies between the drafter's top-1 and the target's top-1 candidates. (c) **U-KNN mass relationship:** The target's probability mass within the drafter's geometric neighborhood (M_K) grows monotonically with U. This spatial concentration justifies shifting from rigid identity matching to locality-informed soft verification.

while multi-scale approaches such as VAR [32] and Infinity [8] generate images across progressive resolutions. Central to both is the discretization of continuous latent spaces into a categorical codebook [7,26], which enables the use of powerful Transformer backbones for scalable synthesis. However, the strictly sequential nature of these models—whether token-by-token or scale-by-scale—incurs substantial inference latency [2, 17, 20, 29, 33, 35]. As model parameters continue to scale up to improve generation quality [30, 39], the growing computational cost and memory overhead of autoregressive decoding have become a primary bottleneck for **efficient** practical deployment in real-world scenarios [36].

2.2 Speculative Decoding

Speculative Decoding (SD) [3, 15] accelerates inference by using a lightweight drafter to propose candidate tokens, which are verified by a larger target model in a single forward pass [18,19,41]. In image generation, LANTERN [11] adopts a relaxed acceptance rule for multi-modal distributions, while Medusa [1] utilizes a tree-based drafting scheme to improve stability under sampling. However, these methods primarily modify the verification or drafting architecture, failing to address the root cause of low acceptance: the intrinsic *misalignment* between drafter and target distributions. Our work distinguishes itself by being the first to systematically bridge this gap at both training and inference stages.

A closely related line of research involves parallel or cluster-based decoding strategies, such as SJD [31], ZipAR [9], and GSD [27]. These approaches leverage Jacobi-like iterations or grouped verification to allow the target model to process multiple positions or clusters in parallel, often exploring randomized

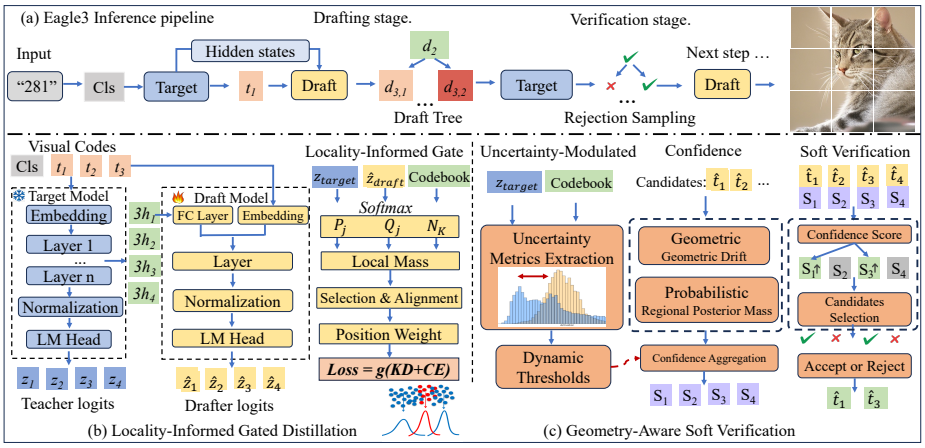


Fig. 4: Overview of the proposed framework. (a) **Inference Pipeline:** Speculative decoding on *LlamaGen-XL-C2I* where the drafter expands a candidate tree, followed by rejection sampling for verification under sampling ($T > 0$). (b) **Locality-Informed Distillation:** The training stage identifies structured local concentrations using geometric locality. We compute alignment-aware scores to prioritize misaligned regions with high target density, followed by a progressive soft gating mechanism to obtain position weights and weighted distillation on the joint support. (c) **Geometry-Aware Soft Verification:** The verification stage modulates acceptance boundaries through a structured pipeline. Target-based uncertainty is first extracted to obtain *dynamic thresholds*. In parallel, we evaluate *geometric and probabilistic confidence* for each candidate to derive a unified score. This mechanism enables stable and efficient speculative decoding under sampling by re-modulating the target distribution.

or parallelized sampling to increase throughput [34, 40]. Unlike these methods that optimize the target model's internal execution or verification logic, LISA focuses on the classic **drafter–target paradigm**. By mitigating drafter–target misalignment in visual autoregressive models, our method alleviates a key bottleneck that persists even for advanced drafters such as EAGLE-3 [19], which has been integrated into production stacks including SGLang [42] and vLLM [14].

3 Method

In this section, we present LISA, a framework for accelerating autoregressive image generation through speculative decoding. We first characterize the limitations of EAGLE-3 through an empirical analysis in Fig. 3. Building on these observations, LISA combines locality-informed distillation during training, described in Sec. 3.2, with geometry-aware soft verification during inference, described in Sec. 3.3. The overall framework is illustrated in Fig. 4.

3.1 Motivation: The Misalignment Problem

We analyze speculative decoding for autoregressive image generation using LlamaGen-XL-C2I as the target model and an EAGLE-3-based drafter. We compare acceptance behavior under greedy decoding and stochastic sampling, and report both acceptance statistics and distributional geometry diagnostics in Fig. 3.

Our analysis yields three observations that characterize the **distributional misalignment** between the drafter and target model.

1. *Acceptance collapse under stochastic sampling.* As shown in Fig. 3(a), the drafter attains a high acceptance rate under greedy decoding, whereas its acceptance rate drops substantially under stochastic sampling. In this setting, probabilistic verification is performed over the draft candidates, while the target distribution assigns probability mass to a much broader set of tokens. Consequently, the probability assigned to any individual draft token can be small, leading to frequent rejection.
2. *Uncertainty implies small preference gaps.* Fig. 3(b) illustrates the relationship between the normalized entropy U and the target log-probability margin

$$\Delta = \log p_t(\text{top1}) - \log p_t(\text{draft\_top1}). \tag{1}$$

   The density map reveals a strong negative correlation: as the target uncertainty U increases, the preference gap Δ between the target top-1 token and the drafter top-1 token decreases. As illustrated in Fig. 2(b), high-uncertainty positions frequently occur around complex textures, object boundaries, and other visually ambiguous regions.
3. *Local probability concentration.* Fig. 3(c) shows that the target KNN probability mass M_K around the drafter's prediction increases with uncertainty U across different values of K, indicating that probability mass becomes increasingly concentrated in local neighborhoods under high uncertainty.

Implications. Visual token distributions can remain locally concentrated despite being flat and multimodal, making pointwise verification overly restrictive. This is consistent with the redundancy of VQ codebooks, where perceptually similar patches map to nearby codebook entries [11]. LISA therefore combines **locality-informed distillation** for training-side alignment with **geometry-aware soft verification** for adaptive inference-time acceptance.

3.2 Locality-Informed Distillation

As shown in Fig. 4(b), locality-informed distillation improves drafter–target alignment by reweighting supervision according to local token-space structure. The framework contains three components: locality-based position selection, progressive soft gating, and weighted distillation over truncated supports.

Locality-informed selection. We instantiate the locality signal with a simple geometric construction. More generally, the framework only requires a per-position score that is high when the target is locally concentrated but the drafter remains misaligned.

Let $p_t(\cdot \mid b,t)$ and $p_s(\cdot \mid b,t)$ denote the target and drafter distributions at position (b,t) over vocabulary $\mathcal{V} = \{1,\dots,V\}$. Let $\mathbf{e}_v$ be an L2-normalized token embedding in a geometry-bearing space, such as a VQ codebook.

We summarize the target's top-K_c region using a probability-weighted centroid:

$$\mathbf{c}_{b,t} = \frac{\sum_{k=1}^{K_c} \tilde{p}_t^{(k)} \mathbf{e}_{i_k}}{\left\| \sum_{k=1}^{K_c} \tilde{p}_t^{(k)} \mathbf{e}_{i_k} \right\|_2}, \tag{2}$$

where $(i_1,\dots,i_{K_c})$ are the target top-K_c tokens and $\tilde{p}_t^{(k)}$ are their renormalized probabilities.

The corresponding local neighborhood is

$$\mathcal{N}_{b,t} = \{v \in \mathcal{V} : d_{\cos}(\mathbf{c}_{b,t}, \mathbf{e}_v) \le r\}, \tag{3}$$

where $d_{\cos}(\mathbf{u},\mathbf{v}) = 1 - \langle \mathbf{u},\mathbf{v} \rangle$ and r is the neighborhood radius.

We measure target concentration within this neighborhood as

$$\alpha_{b,t}^{t} = \sum_{v \in \mathcal{N}_{b,t}} p_t(v \mid b,t). \tag{4}$$

Drafter membership is represented by

$$a_{b,t} = \mathbb{I}\Big[\arg\max_v p_s(v \mid b,t) \in \mathcal{N}_{b,t}\Big], \tag{5}$$

where a soft membership function may replace the hard indicator.

We then define the locality-mismatch score

$$s_{b,t} = \alpha_{b,t}^{t}(1 - a_{b,t}), \tag{6}$$

which emphasizes positions with concentrated target mass but insufficient drafter alignment. Positions above a scheduled batch-wise quantile are selected, producing a mask $m_{b,t} \in \{0,1\}$; the selected proportion is annealed during training.

Progressive soft gating. Because hard selection may overemphasize difficult positions early in training, we modulate supervision using local alignment progress:

$$g_{\text{align}}(b,t) = \exp\Big(- \mathrm{KL}\big(\tilde{p}_t^{(\mathcal{K})}(\cdot \mid b,t) \parallel \tilde{p}_s^{(\mathcal{K})}(\cdot \mid b,t)\big)\Big), \tag{7}$$

where $\mathcal{K}(b,t)$ is the union of top-K tokens from both distributions, with probabilities renormalized over this support.

Combined with a scheduled veto factor $w_{\text{veto}}(b,t)$, the final position weight is

$$w_{b,t} = w_{\text{veto}}(b,t)\, \tilde{g}_{b,t}, \tag{8}$$

where selected positions use $\tilde{g}_{b,t} = g_{\text{align}}(b,t)$, while unselected positions retain a reduced baseline weight to preserve gradient flow.

Distillation objectives. Using $w_{b,t}$, we define a truncated cross-entropy loss over the target nucleus set $\mathcal{S}_t(b,t)$:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{b,t} w_{b,t} \sum_{v \in \mathcal{S}_t(b,t)} \tilde{p}_t(v \mid b,t) \log p_s(v \mid b,t). \tag{9}$$

We further match the target and drafter distributions over the joint support:

$$\mathcal{L}_{\mathrm{KD}} = \tau_s^2 \sum_{b,t} w_{b,t} \mathrm{KL}\Big(\tilde{p}_t^{(\mathcal{K})}(\cdot \mid b,t) \parallel \tilde{p}_s^{(\mathcal{K})}(\cdot \mid b,t)\Big). \tag{10}$$

The final objective is

$$\mathcal{L}_{\mathrm{tot}} = \lambda_{\mathrm{CE}} \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{KD}} \mathcal{L}_{\mathrm{KD}}. \tag{11}$$

3.3 Geometry-Aware Soft Verification

We propose **Geometry-Aware Soft Verification (GASV)** (Fig. 4(c)), which performs locality-aware verification in the target's whitened output-embedding space. Its acceptance thresholds are dynamically modulated by target uncertainty.

Uncertainty-modulated thresholds. Given target logits $\ell \in \mathbb{R}^V$ and $p_t = \mathrm{softmax}(\ell)$, let $\ell_{(1)}$ and $\ell_{(2)}$ denote the largest and second-largest logits. We characterize uncertainty using normalized entropy

$$U = \frac{-\sum_v p_t(v) \log p_t(v)}{\log V}, \tag{12}$$

and normalized margin

$$m_{\mathrm{norm}} = \sigma(\ell_{(1)} - \ell_{(2)}). \tag{13}$$

These statistics modulate the base operating thresholds κ and λ:

$$\begin{aligned} \kappa_{\mathrm{eff}} &= f_\kappa(\kappa;\, U,\, m_{\mathrm{norm}}), \\ \lambda_{\mathrm{eff}} &= f_\lambda(\lambda;\, U,\, m_{\mathrm{norm}}), \end{aligned} \tag{14}$$

where f_κ decreases with U and increases with m_{norm}, while f_λ follows the opposite trend.

Geometric and regional confidence. For the target top-1 token i and a candidate token v, GASV computes two signals. The first measures geometric drift in the whitened output-embedding space:

$$s_{\mathrm{geo}}(v) = \frac{\ell_i - \ell_v}{\|(\mathbf{w}_i - \mathbf{w}_v) \odot \mathrm{invstd}\|_2 + \varepsilon}. \tag{15}$$

This drift is compared with a precomputed cluster-specific budget β_c.

The second signal estimates regional posterior support. We partition the vocabulary into K clusters and define

$$\pi_c = \log \sum_{u \in \mathcal{R}(c)} \exp(\ell_u), \tag{16}$$

where $\mathcal{R}(c)$ is a representative token set for cluster c.

Let c_i and c_v denote the cluster assignments of i and v. We map the two signals to soft confidence scores:

$$g_{\text{geo}}(v) = \sigma(\theta_{\text{geo}} \left(\kappa_{\text{eff}} \beta_{c_i} - s_{\text{geo}}(v)\right)), \tag{17}$$

and

$$g_{\text{clu}}(v) = \sigma(\theta_{\text{clu}} \left(\pi_{c_v} - \pi_{c_i} - \log \lambda_{\text{eff}}\right)). \tag{18}$$

Here, θ_{geo} and θ_{clu} control the sharpness of the two soft boundaries.

Soft verification. We combine the two confidence scores through an OR-style gate:

$$g(v) = \max\left\{g_{\text{geo}}(v), g_{\text{clu}}(v)\right\}. \tag{19}$$

A candidate can therefore be supported by either geometric consistency or regional posterior mass.

Let $p_{\text{raw}}(v)$ denote the target probability after the standard sampling transformations. GASV reweights it as

$$p_{\text{soft}}(v) = \frac{p_{\text{raw}}(v) g(v)}{\sum_u p_{\text{raw}}(u) g(u)}. \tag{20}$$

This yields the normalized distribution used for verification.

Offline calibration. Following prior analyses of visual-token geometry [11], we perform one-time calibration for each target model. We precompute the vocabulary partition, representative sets $\mathcal{R}(c)$, geometry budgets β_c, and whitening statistics by: (i) whitening and clustering target prediction-space token representations; (ii) collecting cluster-wise distributions of $s_{\text{geo}}(v)$ from target rollouts; and (iii) estimating β_c using a high empirical quantile, with the 95th percentile used by default. These assets remain fixed, while (κ, λ) control the speed–quality trade-off. Only U, m_{norm}, $s_{\text{geo}}(v)$, π_c, and the final reweighting are computed online. Additional details are provided in Sec. A.3 of the supplementary material.

4 Experiments

4.1 Setups

Models. We evaluate on three targets: **LlamaGen-XL-T2I** [28] (top-k=1000, CFG=3.5) and **Lumina-mGPT-7B** [22] (top-k=2000, CFG=3.0) for the text-to-image (T2I) task; and **LlamaGen-XL-C2I** (top-k=1000, CFG=4.0) for the class-to-image (C2I) task.

Table 1: Main Results on Text-to-Image (T2I) Task. We evaluate the efficiency and quality of various speculative decoding methods on LlamaGen-XL-T2I Stage II and Lumina-mGPT-7B. Efficiency is measured by *Speedup* and *Acc. Len.* (average accepted tokens per step), while image quality is evaluated using *FID* and *CLIP* score.

Models	Methods	Speedup↑	Acc. Len.↑	FID↓	CLIP↑
LlamaGen-XL-T2I	Vanilla	1.00×	–	45.5	28.6
	EAGLE-2	0.97×	1.20	43.1	28.9
	LANTERN	1.70×	2.40	47.2	28.6
	SJD	1.61×	1.69	48.1	28.7
	LISA (Ours)	**2.11×**	**3.05**	46.7	28.5
Lumina-mGPT-7B	Vanilla	1.00×	–	30.1	33.0
	EAGLE-3	1.86×	2.77	30.2	32.1
	LANTERN	2.56×	3.63	33.9	32.7
	SJD	2.05×	2.23	31.1	31.3
	GSD (G=25)	3.77×	3.53	33.1	31.2
	ZipAR-16	2.31×	3.12	32.6	31.2
	AR-Sample	2.35×	2.61	30.9	26.1
	LISA (Ours)	**4.12×**	**4.79**	31.0	31.5

Table 2: Comparison of performance on LlamaGen-XL-C2I. We report efficiency metrics including *Speedup* and *Acc. Len.*, together with four standard generation quality metrics: *FID*, *IS*, *Precision*, and *Recall*.

Method	Speedup	Acc. Len.	FID↓	IS↑	Prec.↑	Rec.↑
Vanilla	1.0×	1.00	9.78	152.3	0.89	0.38
EAGLE-2	0.8×	1.07	9.82	150.4	0.90	0.46
EAGLE-3	0.7×	1.00	9.88	144.2	0.83	0.47
LANTERN	1.04×	1.50	10.72	149.5	0.87	0.41
Ours (κ=1.0, λ=0.6)	1.39×	2.02	10.50	148.6	0.86	0.39
Ours (κ=1.1, λ=0.4)	**1.53×**	**2.69**	12.60	142.1	0.76	0.42

Datasets and Metrics. For T2I, we evaluate 5,000 images from the MS-COCO 2017 validation set [21] using FID and CLIP score. For C2I, we generate 20,000 samples from ImageNet-1K [5] and report FID, IS, Precision, and Recall.

Efficiency is measured with end-to-end PyTorch inference on a single RTX 4090. We report (1) **speedup**, defined as the ratio of vanilla target-model latency to method latency per image, and (2) **accept length**, defined as the average number of draft tokens accepted per decoding step under each method's verification rule. Latency includes drafter and target forward passes, stochastic sampling, GASV computation, and output generation.

Baselines. We compare two speculative decoding paradigms: (1) Drafter-target methods, including EAGLE-2 [18], EAGLE-3 [19], and Lantern [11]; (2) Jacobi-style and parallel sampling, including SJD [31], GSD [27], AR-Sample [24], and ZipAR [9]. For the C2I task, we specifically tune Lantern's hyperparameters

Fig. 5: Visual comparison on LlamaGen-XL-T2I (Stage II). Top row: Vanilla generations. Bottom row: LISA (κ=1.1, λ=0.4). Our framework achieves 2.11× speedup while preserving the original image composition and structural details.

(LanternK, δ) to maintain generation quality, as detailed in the supplementary material.

Drafter Training. All drafters adopt the **EAGLE-3** architecture and are trained for 6 epochs on A100 GPUs (LR=3×10^{-4}, BF16). Training data includes a 100k LAION-COCO subset for **LlamaGen-XL-T2I** and a 30k target-generated set for **Lumina-mGPT-7B**. More details are provided in the supplementary material.

4.2 Main Results

Text-to-Image (T2I). We evaluate the performance of LISA on the MS-COCO benchmark using two representative models: **LlamaGen-XL-T2I Stage II** and **Lumina-mGPT-7B**. As reported in Table 1, our method outperforms existing speculative decoding baselines in terms of efficiency while maintaining competitive image quality. Specifically, LISA achieves a **2.11× speedup** on LlamaGen-XL-T2I and reaches a substantial **4.12× acceleration** on Lumina-mGPT-7B. The results indicate that LISA provides a favorable speed–quality trade-off, with FID and CLIP remaining comparable to competitive methods across different model architectures.

Class-to-image (C2I). As in Table 2, LISA achieves competitive performance on ImageNet using **LlamaGen-XL-C2I**. By adjusting the verification parameters

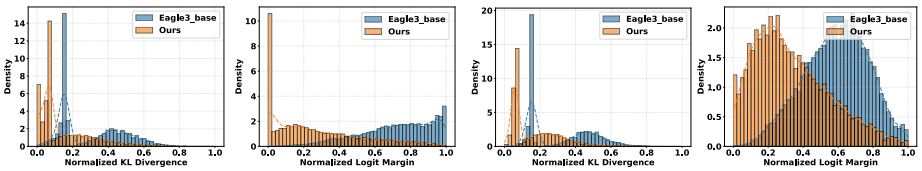
Fig. 6: Visual comparison on Lumina-mGPT-7B. We showcase samples from the **PartiPrompts** benchmark. Top row: Vanilla generations. Bottom row: LISA (κ=1.1, λ=0.4). Visual consistency confirms that our framework preserves complex structural details across diverse prompts while providing inference speedup.

Fig. 7: Visualization results on LlamaGen-XL-C2I. We showcase samples across several ImageNet classes. From top to bottom, rows correspond to different (κ, λ) configurations. Our approach maintains visual quality across these accelerated regimes.

(κ, λ), our method enables a controllable trade-off between speed and quality. Specifically, LISA reaches a peak speedup of **1.53**× with an average accept length of **2.69**, outperforming baseline speculative decoding methods such as EAGLE-3 and LANTERN in efficiency. The results show that LISA maintains competitive generation quality under different acceleration settings.

4.3 Qualitative Results

Visual comparisons in Fig. 5–7 validate LISA's effectiveness across T2I and C2I tasks. Results on LlamaGen and Lumina-mGPT demonstrate that LISA maintains high fidelity, precise prompt alignment, and complex structural details, confirming that our acceleration preserves the target models' generative capabilities. LISA produces realistic, coherent images across various (κ, λ) configurations in accelerated regimes. Overall, the results suggest that LISA maintains competitive visual quality and semantic consistency with the target models under different acceleration settings. Further visualizations are in the supplementary material.



(a) KL divergence (T=0.5) (b) Logit margin (T=0.5) (c) KL divergence (T=1.0) (d) Logit margin (T=1.0)

Fig. 8: Ablation on drafter alignment across sampling temperatures. We compare EAGLE-3 and LISA drafters on LlamaGen-XL-C2I using Top-K KL divergence (K=512) and logit margin at T=0.5 and T=1.0. LISA shifts both distributions toward zero with lower variance, indicating improved alignment.

Table 3: Ablation on GASV parameters across models. We evaluate the impact of drift threshold κ and cluster-posterior threshold λ at T=1.0. *w/o GASV* denotes standard exact-match verification. This table focuses on efficiency variations under different GASV configurations. Note that **Ours** (*w/o* GASV) already outperforms EAGLE-3 due to locality-informed distillation during training.

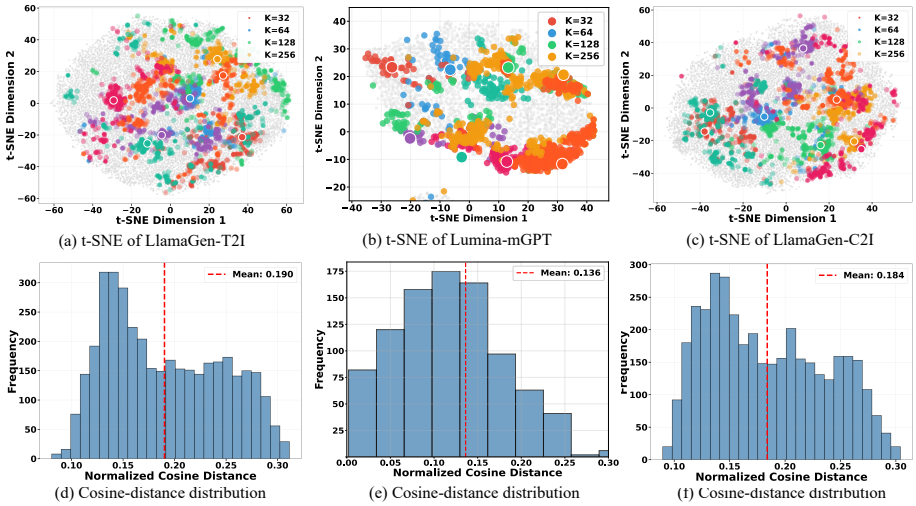
Models	Drafter	w/o GASV	Varying κ (λ=0.4)			Varying λ (κ=1.0)	
			κ=0.6	κ=0.8	κ=1.0	λ=0.6	λ=0.8
LlamaGen-XL-T2I	EAGLE-3	1.00	1.18	1.25	1.32	1.15	1.08
	Ours	1.24	2.16	2.55	2.92	2.80	2.49
LlamaGen-XL-C2I	EAGLE-3	1.00	1.06	1.08	1.25	1.18	1.12
	Ours	1.07	1.76	2.16	2.45	2.02	1.81

4.4 Ablation Studies

Locality-Informed Distillation (Training-Side). We compare LISA and EAGLE-3 drafters on LlamaGen-XL-C2I under standard verification to isolate training-side effects. Figure 8 reports Top-K KL divergence (K=512) and logit margin at T=0.5 and T=1.0. LISA shifts both distributions toward zero with lower variance, indicating improved drafter–target alignment before geometric verification.

Geometry-Aware Soft Verification (Inference-Side). We evaluate GASV with both EAGLE-3 and our locality-distilled drafter on LlamaGen-XL models. Table 3 reports accept length under different κ and λ settings. GASV improves acceptance for both drafters, with the distilled drafter reaching 2.92 on LlamaGen-XL-T2I. Larger κ yields more permissive verification, while larger λ makes acceptance more conservative. These results support GASV as an effective inference-side complement to locality-informed distillation. Qualitative comparisons are provided in Appendix B.4.

Generalization to a Medusa-style Drafter. On LlamaGen-XL-T2I, applying the locality-informed distillation to a Medusa-style drafter improves the accept length from 1.19 to 1.53 and the speedup from 1.04× to 1.37×. Adding GASV further improves the accept length, with comparable generation quality (Table 4).



(a) t-SNE of LlamaGen-T2I (b) t-SNE of Lumina-mGPT (c) t-SNE of LlamaGen-C2I

(d) Cosine-distance distribution (e) Cosine-distance distribution (f) Cosine-distance distribution

Fig. 9: Geometric probe of codebooks. We project the token embeddings into 2D space for visualization (a–c) and compute high-dimensional K-NN distances in the original embedding spaces (d–f). Across the evaluated models, the projected tokens form multiple locally compact regions, while the K-NN distance distributions reveal heterogeneous neighborhood density rather than a uniform arrangement. Although the 2D projections are qualitative, the consistent local structure observed in both views suggests that visual codebooks retain nontrivial geometric organization after discretization. This provides empirical support for using locality as a design prior in locality-informed distillation and geometry-aware verification.

Locality Analysis of Discrete Visual Tokens. To investigate LISA's structural motivation, we conduct a geometric probe on several codebooks. As visualized in Fig. 9, the evaluated tokenizers exhibit a recurring locality pattern, with tokens forming compact clusters and mean distances ranging from 0.136 to 0.190. This observation suggests that vector quantization can retain part of the spatial redundancy and semantic continuity of visual signals. Such locality serves as an empirical structural prior for LISA: the framework uses this geometric coherence to improve drafter alignment and to relax verification within semantically coherent neighborhoods (Sec. 3.3). These results support the use of codebook locality as a practical design prior for both training and verification.

Sensitivity of Parameters. LISA exhibits stable performance across the evaluated settings, as its calibration assets, including the geometry budget β_c, are derived from the target model's local prediction-space structure. Detailed parameter configurations and calibration procedures are provided in the the supplementary material. We further provide a full component-wise decomposition of LISA, isolating training-side and inference-side contributions, in Appendix B.1.

Table 4: Generalization to a Medusa-style drafter on LlamaGen-XL-T2I. We report accept length, end-to-end speedup, FID, and CLIP score. *w/o GASV* uses standard speculative verification.

Method	Acc. Len.↑	Speedup↑	FID↓	CLIP↑
Medusa-style	1.19	1.04×	46.9	28.5
LISA-Medusa w/o GASV	1.53	1.37×	46.0	28.7
LISA-Medusa w/ GASV	**2.25**	**1.82**×	46.2	28.4

5 Conclusion

While speculative decoding has been effective for LLMs, methods like EAGLE-3 provide more limited gains for visual autoregressive models due to their flatter token distributions. We address this challenge with **LISA**, which jointly improves training-side alignment and inference-side verification. By exploiting the geometry of VQ codebooks, LISA combines (i) *locality-informed distillation* to improve drafter alignment in high-entropy regions, and (ii) *geometry-aware soft verification* to increase acceptance under stochastic sampling. Experiments on MS-COCO and ImageNet demonstrate up to **4.12**× acceleration while maintaining competitive generation quality. Overall, LISA achieves a favorable speed–quality trade-off for visual autoregressive generation. Future work may extend this geometry-guided framework to broader generative settings, including video generation and multimodal large language models.

Acknowledgements

This paper is supported by Young Scientists Fund of the National Natural Science Foundation of China (NSFC) (No. 62506305), and in part by Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (No. 2024R01007), and in part by Key Research and Development Program of Zhejiang Province (No. 2025C01026), and in part by Scientific Research Project of Westlake University (No. WU2025WF003).

References

1. Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J.D., Chen, D., Dao, T.: Medusa: Simple llm inference acceleration framework with multiple decoding heads. In: ICML (2024)
2. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: CVPR (2022)
3. Chen, C., Borgeaud, S., Irving, G., Lespiau, J.B., Sifre, L., Jumper, J.: Accelerating large language model decoding with speculative sampling. arXiv preprint arXiv:2302.01318 (2023)
4. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)

5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
6. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
7. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: CVPR (2022)
8. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: CVPR (2025)
9. He, Y., Chen, F., He, Y., He, S., Zhou, H., Zhang, K., Zhuang, B.: Zipar: Accelerating autoregressive image generation through spatial locality. arXiv preprint arXiv:2412.04062 (2024)
10. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
11. Jang, D., Park, S., Yang, J.Y., Jung, Y., Yun, J., Kundu, S., Kim, S.Y., Yang, E.: Lantern: Accelerating visual autoregressive models with relaxed speculative decoding. In: ICLR (2025)
12. Jiang, K., Huang, J.: A survey on vision autoregressive model. arXiv preprint arXiv:2411.08666 (2024)
13. Jin, Y., Li, J., Gu, T., Liu, Y., Zhao, B., Lai, J., Gan, Z., Wang, Y., Wang, C., Tan, X., Ma, L.: Efficient multimodal large language models: A survey. Visual Intelligence **3**(1), 27 (2025)
14. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: SOSP (2023)
15. Leviathan, Y., Kalman, M., Matias, Y.: Fast inference from transformers via speculative decoding. In: ICML (2023)
16. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. In: NeurIPS (2024)
17. Li, Y., Lv, C., Wang, H.: Freqexit: Enabling early-exit inference for visual autoregressive models via frequency-aware guidance. In: NeurIPS (2025)
18. Li, Y., Wei, F., Zhang, C., Zhang, H.: Eagle-2: Faster inference of language models with dynamic draft trees. In: NeurIPS. pp. 7421–7432 (2024)
19. Li, Y., Wei, F., Zhang, C., Zhang, H.: EAGLE-3: Scaling up inference acceleration of large language models via training-time test. In: NeurIPS (2025)
20. Liao, B., Li, Y., Jian, S., Wang, H.: Parallel jacobi decoding for fast autoregressive image generation. In: CVPR (2026)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
22. Liu, D., Zhao, S., Zhuo, L., Lin, W., Xin, Y., Li, X., Qin, Q., Qiao, Y., Li, H., Gao, P.: Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv preprint arXiv:2408.02657 (2024)
23. Liu, X., Dong, P., Hu, X., Chu, X.: Longgenbench: Long-context generation benchmark. arXiv preprint arXiv:2410.04199 (2024)
24. Ma, X., Zhao, F., Ling, P., Qiu, H., Wei, Z., Yu, H., Huang, J., Zeng, Z., Ma, L.: Towards better & faster autoregressive image generation: From the perspective of entropy. In: NeurIPS (2025)

25. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831. PMLR (2021)
26. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: NeurIPS (2019)
27. So, J., Shin, J., Kook, H., Park, E.: Grouped speculative decoding for autoregressive image generation. In: ICCV (2025)
28. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
29. Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: Dycoke: Dynamic compression of tokens for fast video large language models. In: CVPR (2025)
30. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
31. Teng, Y., Shi, H., Liu, X., Ning, X., Dai, G., Wang, Y., Li, Z., Liu, X.: Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. In: ICLR (2025)
32. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: NeurIPS (2024)
33. Wang, H., Jin, X., Lu, L., Li, C., Chen, J., Liu, Q., Wang, H.: Earlytom: Early token compression completes fast video understanding. In: CVPR (2026)
34. Wang, Y., Ren, S., Lin, Z., Han, Y., Guo, H., Yang, Z., Zou, D., Feng, J., Liu, X.: Parallelized autoregressive visual generation. In: CVPR (2025)
35. Xia, H., Yang, Z., Dong, Q., Wang, P., Li, Y., Ge, T., Liu, T., Li, W., Sui, Z.: Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. arXiv preprint arXiv:2401.07851 (2024)
36. Xiong, J., Liu, G., Huang, L., Wu, C., Wu, T., Mu, Y., Yao, Y., Shen, H., Wan, Z., Huang, J., et al.: Autoregressive models in vision: A survey. arXiv preprint arXiv:2411.05902 (2024)
37. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
38. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
39. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 (2022)
40. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Randomized autoregressive visual generation. In: ICCV (2025)
41. Zhang, L., Wang, X., Huang, Y., Xu, R.: Learning harmonized representations for speculative sampling. In: ICLR (2025)
42. Zheng, L., Yin, L., Xie, Z., Sun, C.L., Huang, J., Yu, C.H., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J.E., et al.: Sglang: Efficient execution of structured language model programs. In: NeurIPS (2024)
43. Zhu, J., Wang, H., Su, M., Wang, Z., Wang, H.: Obs-diff: Accurate pruning for diffusion models in one-shot. arXiv preprint arXiv:2510.06751 (2025)