IMPROVING SAMPLE DIVERSITY IN AUTOREGRESSIVE TEXT-TO-IMAGE GENERATION VIA CLUSTER TRUNCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While diffusion models achieve state-of-the-art image quality for text-to-image (T2I) generation, recent work has demonstrated that they suffer from sample diversity collapse. In this work, we investigate whether autoregressive (AR) image generation models can deliver greater sample diversity at an acceptable level of image quality. With recent advances in quality and efficiency, AR models have emerged as a viable alternative to diffusion-based image generation. Beyond enabling new use cases such as interleaved image-text generation, their sequential generation process makes them compatible with a wide range of token-based decoding strategies originally developed to improve diversity in text generation. Motivated by the potential of a better diversity-quality tradeoff in the AR paradigm, we present, to our knowledge, the first systematic study of sample diversity in next-token AR image generation models. We show that two key properties of AR image generation, persistently high token-level entropy and substantial redundancy in visual token spaces, limit the effectiveness of existing token-level decoding methods for diversity enhancement. We therefore propose p-*less cluster*, a new decoding strategy that performs entropy-based truncation sampling at cluster level rather than at token level. We evaluate our approach and baseline decoding methods across four autoregressive T2I models and two datasets using a comprehensive suite of metrics spanning image quality, prompt alignment, and diversity. Our results show that p-less cluster unlocks the greatest diversity across most evaluated autoregressive T2I models and datasets while keeping image quality and prompt alignment at an acceptable level.

1 INTRODUCTION

Text-to-image (T2I) generation has witnessed remarkable progress in recent years, with diffusion-based models such as Flux (Black Forest Labs, 2024), Stable Diffusion (Stability AI, 2024), and Imagen (Saharia et al., 2022) producing photorealistic outputs that closely follow complex text prompts. Despite this progress, diversity collapse has remained a persistent limitation with T2I generation. Diversity collapse manifests at two levels: *distributional diversity* (the model ignores rare or underrepresented modes of the data distribution) and *sample diversity* (repeated sampling for the same prompt yields little variation in composition, style, or content) (Gandikota & Bau, 2026). Critically, these two axes of diversity are distinct and do not always improve together: for example, SDXL-DMD2 (Yin et al., 2024), a distilled diffusion model, achieves superior distributional diversity as measured by FID (Heusel et al., 2017) compared to the base model, yet exhibits markedly lower per-prompt sample diversity (Gandikota & Bau, 2026). In this work, we focus on the latter: *sample diversity*, i.e., generating a diverse set of images for the same prompt. This capability is particularly important in creative applications, where users are often presented with multiple generated candidates to choose from (Parmar et al., 2026; Angelova, 2025).

Recently, autoregressive (AR) image generation (Sun et al., 2024; Wang et al., 2024; Chern et al., 2024; Chen et al., 2025) has emerged as a compelling alternative paradigm to diffusion-style T2I generation. Such models have enabled new use cases such as interleaved image and text generation by unifying language and image generation in a single transformer-based architecture. This has important implications for how sampling is conducted in T2I generation: by formulating image

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

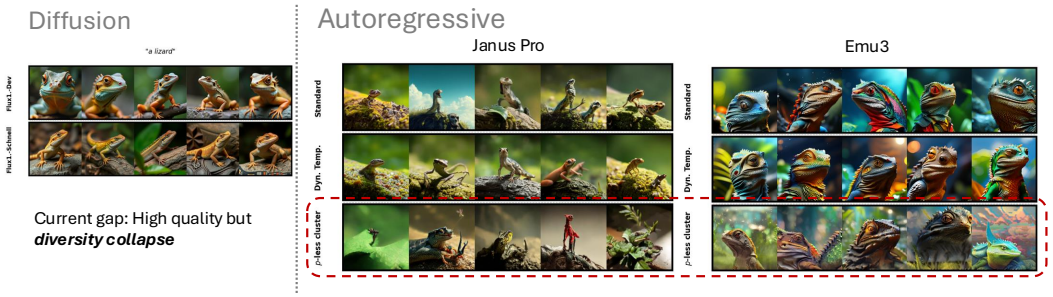


Figure 1: p-less cluster enables **diverse i.i.d. sampling** across autoregressive (AR) image generative models while maintaining acceptable image quality, avoiding the issue of diversity collapse.

synthesis as sequential token prediction, AR models can leverage decoding strategies developed previously for language generation to promote diversity.

In the realm of text generation, a wide range of sampling approaches have been proposed which are directly applicable to AR image generation models. Methods such as top-p (Holtzman et al., 2020), top-K (Fan et al., 2018), min-p (Nguyen et al., 2024), and p-less (Tan et al., 2026) enable greater diversity in language generation without significantly sacrificing output quality by truncating the token probability distribution to a subset of higher-probability tokens, thereby allowing higher temperature values to be applied during sampling. Despite the utility of these sampling strategies for extracting high-quality diverse outputs from LLMs, prior studies on autoregressive T2I have not studied how choice of truncation-based sampling methods impact the sample diversity of generated images.

Our work aims to close this gap by addressing the question: *Can decoding strategies developed for text generation unlock greater sample diversity in autoregressive image models while keeping image quality and prompt alignment at an acceptable level?* Motivated by the high token-level entropy yet substantial redundancy of visual token spaces in AR image models, we propose *p-less cluster*, a cluster-level truncation-based sampling method designed to improve sample diversity while keeping image quality and prompt alignment acceptable. Instead of truncating over individual tokens, p-less cluster truncates over clusters of similar tokens, encouraging exploration across semantically distinct modes while filtering out low-probability regions that often lead to artifacts.

We conduct comprehensive experiments using four leading AR T2I generation models and benchmark against two diffusion-style T2I model over two prompt datasets. Through a suite of evaluation metrics spanning image quality, prompt alignment, and diversity, we demonstrate how p-less cluster extracts greater image diversity across most models and datasets than baseline approaches for sampling. Additionally, we characterize the trade-off between image diversity and quality (Parmar et al., 2026) across temperatures and show that our method remains acceptable over a wider temperature range than baseline decoding methods. We further substantiate our results with qualitative analyses and ablation studies on our approach.

In summary, our contributions are as follows:

- To our knowledge, we conduct the first study of per-prompt sample diversity in token-wise autoregressive text-to-image models, spanning four models and three different vector-quantized codebook paradigms.
- We propose *p-less cluster*, a cluster-aware decoding strategy that performs two-stage sampling with truncation over clusters of visual tokens.
- We conduct extensive experiments across four AR image generation models, showing that our method reaches the highest diversity at an acceptable level of quality and alignment in most settings, and remains acceptable over a wider range of temperatures than standard sampling.
- Through quantitative and qualitative analyses, we validate the robustness of our findings and identify regimes where cluster-level sampling provides the largest diversity gains.

2 RELATED WORK

2.1 DIVERSITY COLLAPSE IN TEXT-TO-IMAGE GENERATIVE MODELS

A large body of work has focused on improving sample diversity in T2I generation, primarily within diffusion-based models. For example, Parmar et al. (2026) propose scaling group inference to generate diverse and high-quality samples, while Kim et al. (2026) introduce contrastive noise optimization to encourage diverse outputs during diffusion sampling. In contrast, improving diversity in autoregressive (AR) image generation remains underexplored. Existing efforts such as DiverseAR (Yang et al., 2025) enhance diversity in bitwise or next-scale AR frameworks (e.g., Infinity (Han et al., 2025)), but do not address next-token AR generation. To address this gap, our work focuses on this setting, aiming to improve diversity under the standard token-by-token autoregressive formulation.

2.2 AUTOREGRESSIVE IMAGE GENERATION

Pioneering AR image generation operated in pixel space with convolutional (Van den Oord et al., 2016) or recurrent neural networks (Van Den Oord et al., 2016). Lately, motivated by the success of language models, current AR image generation models have adopted transformers and tokenization techniques, which can be broadly categorized into next-token, next-scale, and mask-based paradigms. Next-token prediction generates an image sequentially via discrete tokens, such as via vector quantized Variational Autoencoder (Razavi et al., 2019) or Generative Adversarial Networks (Yu et al., 2022a). Sun et al. (2024) shows that this framework could achieve state-of-the-art performance. Next-scale prediction treats images as a series of multi-resolution maps, performing a coarse-to-fine generation (Tian et al., 2024; Han et al., 2025; Liu et al., 2025), with the potential to better model global structural nuances during early stages. Mask-based AR models dispense with the strict adherence to a predefined sequential generation order (Li et al., 2024; Yao et al., 2025). Our work focuses on the next-token paradigm since it is the most mature, benefiting from a vast ecosystem of established techniques in NLP, and serves as a diverse testbed for new methodological improvements (Mu et al., 2025).

2.3 DECODING STRATEGIES FOR AUTOREGRESSIVE IMAGE GENERATION

Prior research on decoding strategies for AR image generation has largely focused on improving inference efficiency rather than diversity. For instance, ZipAR (He et al., 2024) exploits spatial locality to enable parallel decoding and accelerate generation. Similarly, speculative decoding methods such as Speculative Jacobi-Denoising Decoding (Teng et al., 2025) leverage iterative refinement to speed up autoregressive sampling. Beyond speed, Ma et al. (2025) propose an entropy-based decoding strategy that adapts token generation order and temperature based on spatial entropy, motivated by the observation that images exhibit non-uniform information density across regions. While this approach provides a principled way to improve both efficiency and quality, it does not explicitly target sample diversity. One common approach to increase i.i.d. sample diversity at inference in autoregressive models is to increase temperature applied to discrete token probability distribution. In this work, we focus on decoding strategies designed specifically to enhance diversity at higher temperature in AR image generation, and compare against Ma et al. (2025) as a strong baseline.

3 METHODOLOGY

3.1 PRELIMINARIES

Let $p_\theta(x \mid c)$ denote an AR image generation model that generates an image x conditioned on a text prompt c. Our goal is to design a decoding procedure $\mathcal{D}$ such that a set of M images, sampled independently as

$$x_1, \ldots, x_M \overset{\text{i.i.d.}}{\sim} \mathcal{D}(p_\theta, c),$$

achieves higher set-level diversity while keeping image quality and prompt alignment at an acceptable level. Importantly, we focus on i.i.d. sampling without introducing cross-sample coupling or sequential dependence, ensuring that all samples can be generated in parallel and that diversity improvements arise purely from the decoding strategy.

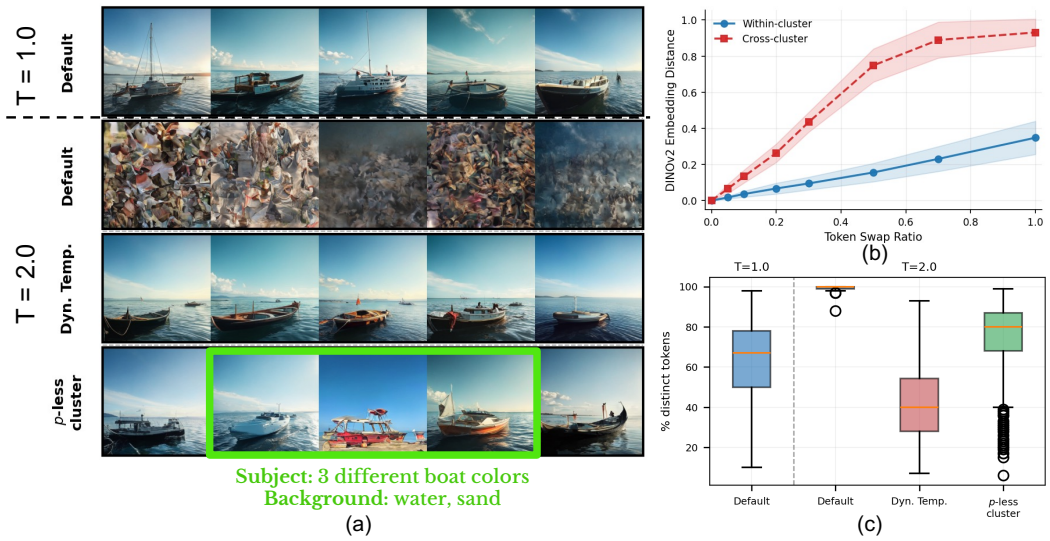


Figure 2: **Comparison of decoding strategies on image- and token-level diversity (Janus Pro).** (a) Visual comparison; (b) impact of within- vs. cross-cluster token swaps on image diversity; (c) percentage of distinct tokens across generation steps (100 samples). At high temperature, default sampling admits too many tokens, leading to degeneration. Dynamic temperature mitigates this by reducing the effective temperature, but limits diversity due to fewer distinct tokens. p-less cluster achieves a better tradeoff by combining cluster-level sampling with truncation.

3.2 p-LESS CLUSTER SAMPLING

We propose a cluster-aware decoding framework for AR image generation that consists of (1) constructing a partition of the token vocabulary, (2) performing cluster-level probability filtering using the p-less entropy-based truncation (Tan et al., 2026), and (3) sampling via a two-stage procedure. Step (1) is executed offline to ensure no substantial cost is incurred at inference time.

Token Vocabulary Partitioning via K-Means. Let V denote the size of the discrete image-token vocabulary. We partition the vocabulary into n clusters $\{\mathcal{C}_j\}_{j=1}^{n}$ using unsupervised K-Means on codebook vectors (codebook statistics are given in Table 2). The resulting partition is computed once offline.

Cluster-Level p-less Filtering. At each decoding step, the model produces a token distribution $p \in \Delta^{V-1}$, where Δ^{V-1} denotes the probability simplex over the V tokens. We aggregate token probabilities into cluster probabilities:

$$q_j = \sum_{v \in \mathcal{C}_j} p_v, \quad j = 1, \ldots, n.$$

We then apply p-less (Tan et al., 2026) truncation to the cluster distribution, masking clusters with probability below

$$\hat{p}_k = \frac{1}{\exp\left(H_k(q)\right)},$$

where H_k denotes the Rényi entropy of order k, followed by renormalization. This adaptively selects a subset of semantically coherent regions in the vocabulary.

Two-Stage Sampling. First, a cluster index $\hat{j}$ is drawn from the filtered cluster distribution. Second, a token is sampled from the original token-level probabilities restricted to the selected cluster:

$$\hat{v} \sim \frac{p_v}{\sum_{u \in \mathcal{C}_{\hat{j}}} p_u}, \quad v \in \mathcal{C}_{\hat{j}}.$$

This design allows diversity control at the cluster level while preserving the model's fine-grained preferences within each cluster.

3.3 Intuition for cluster-level truncation sampling

Our approach is motivated by two key properties that distinguish visual generation from traditional AR text generation, which are summarized in the following two observations.

Observation 1: High token-level entropy during generation. Compared to language models, AR image models maintain relatively high entropy across many decoding steps (as shown in Figure 6).

Observation 2: High redundancy in the token space. Despite this high entropy, Figure 2(b) demonstrates that randomly swapping all tokens with other tokens from the same cluster results in relatively small image-level changes, with DINOv2 embedding distance remaining below 0.4. This indicates that diversity at the discrete token level does not necessarily translate into meaningful diversity at the image level. In contrast, swapping tokens across clusters produces substantially greater image diversity, suggesting that tokens *across* different clusters are more semantically divergent.

Together, these observations explain why default sampling exhibits diversity collapse despite the high entropy of AR image models (Figure 2(a)). Although many tokens are admitted during sampling, much of this variation occurs within semantically redundant regions of the token space, leading to visually similar outputs. Ma et al. (2025) investigate non-default temperature schedules for AR image generation, proposing a decoding strategy that dynamically decays temperature based on token-level entropy. While entropy-based temperature adjustment has improved the quality–diversity balance in text generation (Zhang et al., 2024), it is less effective in the visual domain due to the persistently high token-level entropy of AR image models (Observation 1). In practice, high entropy causes the method to reduce the effective temperature during generation, which mitigates artifacts but simultaneously collapses diversity by restricting the effective candidate set (as shown in Figure 2(c)).

Cluster-level truncation effectively mitigates this limitation. Leveraging insights from Observation 2, our method operates over clusters of semantically similar tokens to reduce redundancy in the candidate space. By first sampling at the cluster level, p-less cluster enables the model to preserve multiple meaningful modes of the distribution rather than collapsing to a set of similar-looking tokens. However, sampling at high temperature might admit too many clusters which can lead to degeneration, similar to default sampling (Figure 2(a)(c)). To address this, we combine cluster-level sampling with the p-less entropy-based truncation (Tan et al., 2026), which filters out undesirable clusters based on cluster-level entropy while retaining a diverse set of plausible modes (formalized in Appendix B).

This design yields a favorable balance between diversity and quality, as illustrated in Figure 2(a). By explicitly encouraging sampling across clusters while controlling tail mass, our method increases the number of distinct tokens (Figure 2(c)) and promotes cross-cluster exploration while keeping images coherent. Consequently, it achieves higher image-level diversity compared to token-level methods such as dynamic temperature, which reduces diversity by overly suppressing non-dominant tokens under high token-level entropy.

3.4 Key design choices

Offline vs. online clustering. We adopt offline clustering rather than performing clustering at inference time. While online clustering could in principle adapt to the local token distribution at each decoding step, it introduces substantial computational overhead. As inference is already latency-sensitive in AR generation, performing clustering per step would be time cost prohibitive.

p-less for truncation. Although our cluster-based framework is compatible with a range of truncation strategies, we adopt p-less due to its strong empirical performance in high-temperature text generation (Tan et al., 2026). Unlike fixed-threshold methods such as top-K or top-p, p-less adapts the truncation threshold based on entropy, making it well-suited to both token-level and cluster-level distributions. This adaptive behavior is particularly important in high-entropy regimes typical of visual generation.

Hyperparameters. Besides temperature, which is common to all decoding strategies, our method introduces two additional hyperparameters: (i) k, the order of the Rényi entropy used for cluster-level

probability truncation, and (ii) n, the number of clusters. To ensure practical usability and fair comparison with baselines, we restrict the search space to a small, fixed set: $k \in \{1.5, 2.0\}$ and $n \in \{500, 1000\}$. We find this range sufficient to achieve competitive performance across models. In an extended sweep on Anole (Appendix E.3), a larger n tends to increase diversity, while a smaller n is more robust to quality degradation at high temperatures; we find $n = 500$ optimal for most models.

4 Experiment setup

4.1 Base models

We apply our decoding strategy to a diverse set of recent AR text-to-image models: **Janus Pro-7B** (Chen et al., 2025), **LlamaGen (Stage 1)** (Sun et al., 2024), **Emu3** (Wang et al., 2024), and **Anole** (Chern et al., 2024). These models vary in architecture (Janus Pro-7B, Emu3, and Anole are unified understanding-generation architectures, while LlamaGen is a generation-only backbone), tokenizer design, and decoding defaults, providing a robust testbed for evaluating decoding strategies across heterogeneous AR systems. Table 2 summarizes the visual codebook statistics for each model.

4.2 Datasets

We use two complementary text-to-image benchmarks for evaluation: **PartiPrompts** (Yu et al., 2022b), a suite of prompts spanning a wide range of semantic categories and difficulty levels, and **GenEval** (Ghosh et al., 2023), which focuses on compositional reasoning with object-centric prompts. We select these datasets to cover both abstract and concrete prompt regimes. PartiPrompts emphasizes open-ended and imaginative descriptions (e.g. *a smiling banana wearing a bandana*), while GenEval targets precise compositional alignment (e.g., *a photo of a traffic light next to a person*).

For GenEval, we use all 553 prompts covering six categories (single-object, two-object, counting, colors, color-attribute binding, and compositional relations). For PartiPrompts, we use all 535 prompts from the *Basic*, *Complex*, and *Challenge* categories. For each prompt, we generate a set of 5 images using the model-specific default classifier-free guidance (CFG) scale: 3.0 for Anole, 5.0 for Emu3, 5.0 for Janus Pro-7B, and 7.5 for LlamaGen (unless stated otherwise).

4.3 Evaluation metrics

We adopt established automatic metrics following Parmar et al. (2026) while also supplementing them with a VLM-as-judge protocol we developed. This was done to address known limitations of feature-based scores (Hu et al., 2023; Cho et al., 2023) such as CLIPScore (Hessel et al., 2021), which is widely used to measure both image fidelity and prompt alignment but can behave like a bag of words on compositional prompts (Lin et al., 2024). Off-the-shelf vision-language models agree with human judgments of prompt alignment more closely than CLIPScore (Lin et al., 2024), and strong multimodal judges rate image quality and alignment with correlations approaching the agreement between human raters (Ku et al., 2024). We therefore use a vision-language model (InternVL3-14B (Zhu et al., 2025)) to assign scores on a 1–5 scale to each image, following an explicit rubric for each dimension. We use an open-weight judge so that the evaluation is reproducible and not subject to changes in proprietary APIs; it is not derived from any of the generators it scores. All scores are averaged over three independent runs to reduce evaluation variance. See Appendix F for complete details.

Overall, we evaluate performance along three axes: *image quality*, *prompt alignment*, and *sample diversity*, where improving *sample diversity* is the main focus of our work. Image quality and prompt alignment are primarily used as quality constraints to ensure that diversity gains do not come at the expense of unacceptably poor image quality. Following Parmar et al. (2026), we compute standard errors using bootstrapping with 1000 resamples.

Image quality. We measure image quality using VLM-as-judge scores (1–5 scale), where the model evaluates perceptual realism and absence of artifacts.

Prompt alignment. We use (i) VLM-as-judge scores (1–5 scale), and (ii) CLIPScore (Hessel et al., 2021), reported as a secondary metric for comparability with prior work.

Sample diversity. We use two metrics computed over a set of 5 generated images per prompt: (i) DINOv2 embedding distance (Oquab et al., 2023), and (ii) Vendi Score (Friedman & Dieng, 2022).

Other metrics. We do not report FID, which measures distribution-level fidelity rather than per-prompt sample diversity (Section 1) and needs reference images that PartiPrompts lacks. We score both prompt sets with the same VLM rubric instead of GenEval's detection-based metric, so "GenEval" denotes the prompt set throughout; learned preference models such as HPS (Wu et al., 2023) are not evaluated.

4.4 Baselines

We compare our method against the following baselines (see complete hyperparameter settings in Appendix D): **(1) Default decoding.** We adopt each model's standard decoding strategy: multinomial sampling for Janus Pro-7B and LlamaGen, top-p sampling ($p = 0.9$) for Anole, and top-K ($K = 2048$) for Emu3. **(2) Dynamic temperature decoding.** We include entropy-based dynamic temperature adjustment (Ma et al., 2025) as a strong baseline. This method modulates the sampling temperature based on token-level entropy, prioritizing regions with higher spatial information density. **(3) Diffusion model baselines.** We additionally evaluate the state-of-the-art Flux.1-Dev and Flux.1-Schnell models (Black Forest Labs, 2024) using default settings: 50 steps with $\mathrm{CFG} = 3.5$ for Flux.1-Dev, and 4 steps for Flux.1-Schnell. As our work focuses on AR model diversity, we do not compare against specific diffusion-based diversity enhancement techniques.

5 Results

Every decoding strategy faces a trade-off between fidelity and diversity, as raising the temperature increases diversity until images diverge from the prompt (Figure 4). Since any method can report higher diversity by accepting lower quality, we compare methods by the diversity they reach *while remaining acceptable*, i.e., with VLM quality and alignment above 3.0 (see Appendix F for rubric details). Each method runs at its own temperature which was chosen under this quality constraint on a tuning subset (Appendix D).

Table 1: **Diversity at an acceptable level of quality and alignment (default CFG).** Each AR decoding strategy runs at the temperature T that maximizes diversity on the tuning subset subject to VLM quality and alignment above 3.0 (Appendix D). Quality and alignment on the full benchmark are shown next to diversity to make the trade-off explicit; † marks values below 3.0. Diffusion baselines use their default settings. Bold marks the highest mean DINOv2 per model; $\pm$ is the bootstrap standard error (1000 resamples). Vendi and CLIPScore are reported in Tables 4 and 5.

Model	Decoding Strategy	GenEval				PartiPrompts			
		T	Quality	Alignment	DINOv2	T	Quality	Alignment	DINOv2
Flux.1-Dev	Default	–	–	–	0.311$\pm$0.016	–	–	–	0.354$\pm$0.013
Flux.1-Schnell	Default	–	–	–	0.270$\pm$0.006	–	–	–	0.277$\pm$0.008
Anole	Default	1.5	3.82	3.21	0.548$\pm$0.006	1.5	4.15	3.43	0.535$\pm$0.005
	Dynamic Temperature	2.0	4.37	3.57	0.505$\pm$0.007	1.5	4.32	3.56	0.531$\pm$0.008
	p-less cluster (Ours)	2.5	3.48	3.04	**0.563**$\pm$0.006	2.5	3.82	3.29	**0.542**$\pm$0.007
Emu3	Default	2.0	4.02	2.96†	**0.441**$\pm$0.007	3.0	4.68	3.80	0.364$\pm$0.007
	Dynamic Temperature	4.0	4.38	3.38	0.405$\pm$0.007	3.0	4.70	3.86	0.360$\pm$0.007
	p-less cluster (Ours)	3.0	4.08	3.08	**0.441**$\pm$0.007	3.5	4.23	3.30	**0.400**$\pm$0.007
Janus Pro	Default	2.5	3.76	3.48	**0.374**$\pm$0.006	2.0	3.96	3.40	0.426$\pm$0.007
	Dynamic Temperature	3.5	4.76	4.76	0.222$\pm$0.005	4.0	4.50	4.01	0.375$\pm$0.007
	p-less cluster (Ours)	4.0	4.10	4.06	0.328$\pm$0.006	4.0	3.72	3.18	**0.454**$\pm$0.007
LlamaGen	Default	1.5	3.58	3.10	0.526$\pm$0.005	1.5	3.85	3.42	0.474$\pm$0.006
	Dynamic Temperature	4.0	3.61	3.08	0.541$\pm$0.006	2.0	3.91	3.48	0.471$\pm$0.006
	p-less cluster (Ours)	3.0	3.49	2.94†	**0.543**$\pm$0.005	3.0	3.83	3.42	**0.479**$\pm$0.006

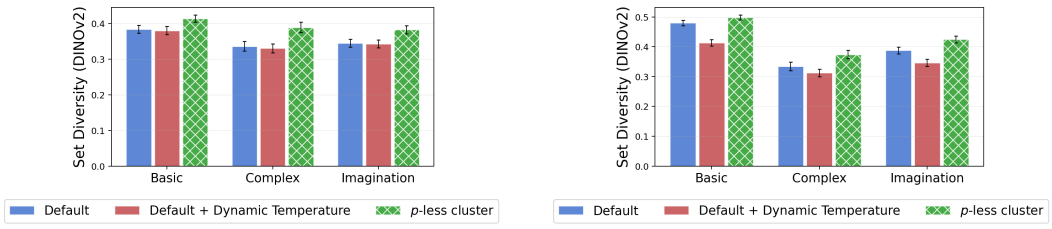
5.1 Main results

AR models achieve higher diversity than diffusion baselines at default CFG. As shown in Table 1 (and Tables 4 and 5 for Vendi), all four AR models outperform diffusion baselines on both diversity metrics across GenEval and PartiPrompts, with the exception of Emu3 on PartiPrompts, which is comparable to Flux.1-Dev. Relative improvements in DINOv2 of Default decoding over

Flux.1-Dev range from 20% to 76% in the remaining settings. This suggests that AR models possess a stronger inherent capacity for diversity, likely due to their token-wise generation process. Within the diffusion baselines, we observe that Flux.1-Schnell consistently underperforms Flux.1-Dev on both diversity metrics. This is consistent with prior findings that distilled diffusion models trade off diversity for efficiency (Gandikota & Bau, 2026).

p-less cluster sampling reaches the highest acceptable diversity in most settings. Table 1 (and Tables 4 and 5) shows that p-less cluster sampling achieves the highest mean diversity on both GenEval and PartiPrompts for Anole and LlamaGen. The clearest gains over the best baseline are for Emu3 and Janus Pro on PartiPrompts ($+0.036$ and $+0.028$ DINOv2) and for Janus Pro at low CFG (Tables 6 and 7). Against Dynamic Temperature, p-less cluster is more diverse in all eight settings, well beyond the standard errors in five. On GenEval, p-less cluster matches or underperforms Default sampling for Emu3 and Janus Pro (analyzed below). Full results including Vendi and CLIPScore are provided in Appendix E.1. As the trade-off predicts, its quality and alignment are lower than those of Dynamic Temperature and usually lower than those of Default: p-less cluster remains acceptable at higher temperatures and therefore operates closer to the threshold. Scores stay above 3.0 except for two alignment values (Table 1): 2.94 for p-less cluster on LlamaGen and 2.96 for Default on Emu3, both on GenEval.

p-less cluster increases the diversity gap at low CFG for Janus Pro. We observed that the maximum diversity across different decoding methods is below 0.5 for DINOv2 at default CFG for Janus Pro and Emu3. Therefore, we further investigated how these two models perform in the low-CFG setting. For Emu3, none of the decoding strategies achieves an average VLM alignment score above 3.0 across all temperatures in the low-CFG regime, indicating severe quality degradation. Tables 6 and 7 provide results for Janus Pro, which did meet the quality requirement in this low-CFG setting (full temperature sweep in Figure 12). We observe that p-less cluster outperforms the baseline methods by an even larger margin in this setting. Lowering CFG and changing the decoding strategy are thus complementary: on GenEval, lowering CFG from 5 to 1 raises Default's diversity from 0.374 to 0.442, and p-less cluster raises it to 0.552 with quality and alignment above 3.0 (Table 6).



(a) Emu3 (b) Janus Pro-7B

Figure 3: **DINOv2 diversity by PartiPrompts challenge category.** p-less cluster achieves larger gains in the *Complex* and *Imagination* categories with more open-ended and abstract prompts.

p-less cluster sampling is most effective for complex, open-ended prompts. To better understand these gains, we analyze diversity by prompt category for Emu3 and Janus Pro, the two models where p-less cluster shows more mixed results. For both models, diversity improvements are largest in the *Complex* and *Imagination* categories, where prompts admit multiple plausible visual realizations (Figure 3). In contrast, for concrete object-centric prompts in the *Basic* category, which resemble GenEval prompts (e.g., "a photo of a car"), p-less cluster yields its smallest gains for both models. This trend is less pronounced for Anole and LlamaGen (Figure 11), suggesting that the effectiveness of p-less cluster depends on both prompt complexity and the model's inherent diversity capacity. Indeed, at default CFG no method exceeds 0.45 DINOv2 on GenEval for Emu3 or Janus Pro, consistent with peaked output distributions that leave few clusters above the truncation threshold. For Janus Pro, p-less cluster peaks at the largest temperature in our grid (T=4.0) while still scoring about 4.1 on VLM quality and alignment (Figure 8), so the searched range, not image quality, limits its diversity.

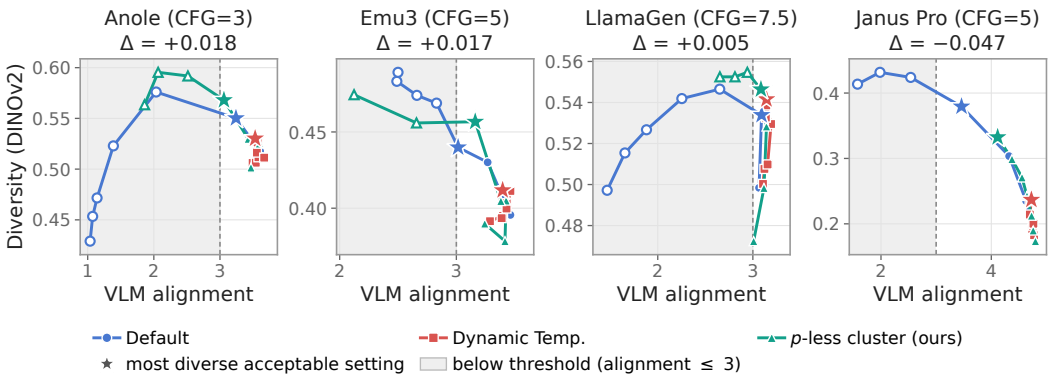


Figure 4: **Diversity–alignment trade-off on GenEval (tuning subset, default CFG).** Curves trace T=1.0 to 4.0 in steps of 0.5; hollow markers fall below the acceptability threshold (shaded). Stars mark each method's most diverse acceptable setting; Δ is the DINOv2 difference between the p-less cluster star and the best baseline star. Janus Pro at CFG=1 is shown in Figure 12.

5.2 Diversity-alignment-quality tradeoff

To characterize the diversity-alignment-quality tradeoff, we swept the temperature of every decoding strategy on smaller subsets of GenEval (120 prompts) and PartiPrompts (90 prompts) to limit computational cost (Appendix D). Figure 4 illustrates this diversity-alignment tradeoff for GenEval dataset. Every method traces a curve along which higher temperature buys diversity at the cost of alignment; the methods differ in how far their curves reach before crossing the acceptability threshold. On GenEval, p-less cluster stays above the threshold over a wider temperature range than Default for all four models ($T \leq 2.5$ vs. 1.5 for Anole and LlamaGen, 3.0 vs. 2.0 for Emu3, 4.0 vs. 2.5 for Janus Pro) and reaches the most diverse acceptable setting for Anole, Emu3, and LlamaGen. Default leaves the acceptable region at lower temperatures, and for Anole and LlamaGen it then loses diversity as well as alignment as its images degenerate. For Janus Pro, Default's most diverse acceptable setting is more diverse than that of p-less cluster (0.38 vs. 0.33), although p-less cluster is still well above the threshold at the largest temperature we tested. Across both datasets (Figures 8 and 9), the most diverse acceptable setting of p-less cluster exceeds that of Dynamic Temperature in all eight model–dataset pairs and that of Default in six, with the exceptions being Janus Pro on GenEval and Emu3 on PartiPrompts. Dynamic temperature adjustment yields inconsistent gains and can reduce diversity in some cases (Table 1). Furthermore, it does not offer meaningful tradeoffs between diversity and alignment; instead, different temperature values are clustered closely for all models except LlamaGen (Figure 4). Qualitative comparisons in Figure 2 and Figure 13 support these findings. Default and dynamic temperature decoding tend to produce visually similar samples for a given prompt, reflecting diversity collapse. In contrast, p-less cluster sampling generates a broader range of outputs, with variation in object position, color, and scene composition, while preserving prompt fidelity.

6 Conclusion

In this work, we conducted the first comprehensive study of how decoding strategies impact sample diversity in next-token AR T2I models. Inspired by our observation that token-level diversity does not necessarily translate into diversity at the pixel level, we proposed p-less cluster: a new truncation-based decoding method for AR T2I which first samples a cluster of visually similar tokens before selecting a token within the cluster. Through extensive experiments and analyses, we demonstrated that p-less cluster increases sample diversity while keeping image quality and prompt alignment at an acceptable level across a range of AR T2I models and prompt settings, particularly those which involve more complex or abstract descriptions. Our study highlights the potential for AR models to improve image sample diversity through principled decoding methods which are designed to address the unique characteristics of sampling in AR T2I generation.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, and implementing methods. We have not used generative AI tools for designing or providing feedback on research methodology or experiments. Assistance with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analyses are not applicable to this work. Additionally, we used generative AI tools for assistance in drafting parts of a research paper as well as creating & modifying scientific figures or images. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available datasets and models, whose licenses are listed in Appendix D.3. Limitations and broader impacts of our method are discussed in Appendix A.

REPRODUCIBILITY STATEMENT

The models, datasets, evaluation metrics, and baselines used in our study are described in Section 4. Implementation details, the hyperparameter tuning protocol, and the selected hyperparameters for every model and decoding strategy are provided in Appendix D (Table 3). The complete VLM-as-judge prompt and protocol are given in Appendix F. Both datasets are publicly available, and code will be provided upon acceptance.

REFERENCES

Nadezhda Angelova. Integration of artificial intelligence into traditional graphic designer software products–a relief and a challenge. In *AIP Conference Proceedings*, volume 3274, pp. 080008. AIP Publishing LLC, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. URL https://arxiv.org/abs/2501.17811.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. URL https://arxiv.org/abs/2407.06135.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1304–1313, March 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023. URL https://arxiv.org/abs/2310.11513.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Han_Infinity_Scaling_Bitwise_AutoRegressive_Modeling_for_High-Resolution_Image_Synthesis_CVPR_2025_paper.html.

Yefei He, Feng Chen, Yuanyu He, Shaoxuan He, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Zipar: Parallel auto-regressive image generation through spatial locality. *arXiv preprint arXiv:2412.04062*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. URL https://arxiv.org/abs/2104.08718.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EVRMnAREc3.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024. URL https://aclanthology.org/2024.acl-long.663/.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2404.01291.

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Infinitystar: Unified spacetime autoregressive modeling for visual generation. *Advances in neural information processing systems*, 38, 2025.

Xiaoxiao Ma, Feng Zhao, Pengyang Ling, Haibo Qiu, Zhixiang Wei, Hu Yu, Jie Huang, Zhixiong Zeng, and Lin Ma. Towards better & faster autoregressive image generation: From the perspective of entropy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=LiQH1MOCMs.

Jiteng Mu, Nuno Vasconcelos, and Xiaolong Wang. Editar: Unified conditional generation with autoregressive models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7899–7909, 2025.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*, 2024.

Maxime Oquab, Timothée Darcet, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Jackson Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IyTNxjTuWT.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Stability AI. Introducing Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5, 2024. Accessed: 2026-05-06.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. URL https://arxiv.org/abs/2406.06525.

Runyan Tan, Shuang Wu, and Phillip Howard. p-less sampling: A robust hyperparameter-free approach for llm decoding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yao Teng, Fu-Yun Wang, Xian Liu, Zhekai Chen, Han Shi, Yu Wang, Zhenguo Li, Weiyang Liu, Difan Zou, and Xihui Liu. Speculative jacobi-denoising decoding for accelerating autoregressive text-to-image generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=y2eWc6jrlu.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. URL https://arxiv.org/abs/2409.18869.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Ying Yang, Zhengyao Lv, Tianlin Pan, Haofan Wang, Binxin Yang, Hubery Yin, Chen Li, and Chenyang Si. Diversear: Boosting diversity in bitwise autoregressive image generation. *arXiv preprint arXiv:2512.02931*, 2025.

Ting Yao, Yehao Li, Yingwei Pan, Zhaofan Qiu, and Tao Mei. Denoising token prediction in masked autoregressive models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18024–18033, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487, 2024.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *International Conference on Learning Representations*, 2022a.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022b. URL https://arxiv.org/abs/2206.10789.

Shimao Zhang, Yu Bao, and Shujian Huang. Edt: Improving large language models' generation by entropy-based dynamic temperature sampling. *arXiv preprint arXiv:2403.14541*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Limitations & Broader Impacts

Limitations. Our approach operates purely at inference time, and its effectiveness is therefore bounded by the base model's underlying token distribution. In particular, gains in diversity depend on the model assigning sufficient probability mass to multiple desirable modes; when the distribution is overly peaked (e.g., as observed for Janus Pro under default CFG), the potential for improvement is inherently limited.

Additionally, the method introduces two hyperparameters (Section 3.4). Although we restrict tuning to a small 2x2 combinations, selecting optimal values still requires model-specific calibration.

Broader impacts. Though our work does not have direct societal impacts, given that AI-generated images are being used for foundation model training, improving sample diversity in generative models mitigates the risk of mode collapse, ensuring that underrepresented demographic groups and rare cultural nuances are accurately reflected in synthetic data.

B Theoretical justification

Token-level sampling treats each token independently, even though autoregressive image models often spread probability mass across many visually similar tokens within a cluster. As a result, token-level filtering can remove tokens that collectively represent a meaningful generation mode, reducing diversity. Cluster-level p-less instead operates on aggregate cluster probabilities, preserving all tokens from clusters that survive truncation. The tokens most affected by this difference are those filtered at the token level but potentially retained at the cluster level, which we term *High-Support Low-Probability* (HSLP) tokens (defined by (HSLP-1) and (HSLP-2) below). We show empirically that HSLP tokens are common in practice and prove that cluster-level p-less assigns them strictly higher probability than token-level p-less.

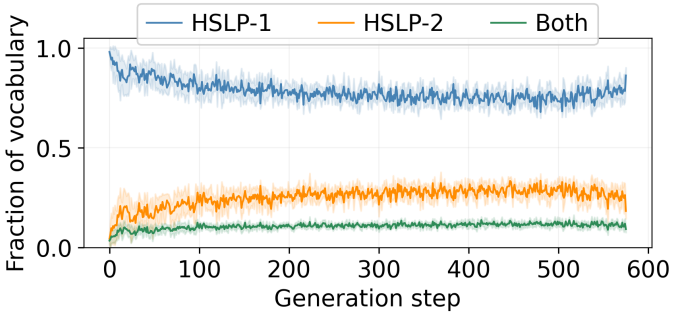


Figure 5: **Fraction of vocabulary tokens satisfying the HSLP conditions** (HSLP-1) **and** (HSLP-2) **for Janus Pro-7B across decoding steps.** A reasonable portion of tokens (approximately 10%) satisfy both conditions, indicating that they are filtered by token-level p-less but would survive cluster-level p-less, thus receiving a non-zero probability under our method.

B.1 Empirical prevalence of HSLP tokens

Figure 5 shows the fraction of tokens satisfying the HSLP conditions for Janus Pro-7B across decoding steps. Approximately 10% of tokens at each step are filtered out by token-level p-less yet belong to a cluster that survives cluster-level filtering. This motivates the formal analysis in Proposition B.1 below: such tokens receive strictly higher probability under cluster-level p-less, and recovering them can contribute to improved diversity.

B.2 Theoretical analysis of token admission

Proposition B.1 (HSLP token admission)**.** ***Assumptions and notation.*** *As in Section 3.2, let* $p \in \Delta^{V-1}$ *denote the token distribution at a decoding step, let* $\{\mathcal{C}_j\}_{j=1}^{n}$ *be the partition of the*

vocabulary into n clusters, and let $q_j = \sum_{v \in \mathcal{C}_j} p_v$ be the induced cluster distribution. For a distribution r, write $\hat{p}_k(r) = 1/\exp(H_k(r))$ for the p-less threshold of order k, so that token-level p-less keeps the tokens with $p_v \geq \hat{p}_k(p)$ and p-less cluster keeps the clusters with $q_j \geq \hat{p}_k(q)$. For $k = 2$, $\hat{p}_2(r) = \sum_i r_i^2$ is the collision probability. Let $v^* \in \mathcal{C}_{j^*}$ be a token such that $p_{v^*} > 0$ and satisfying the following High-Support Low-Probability (HSLP) conditions:

$$p_{v^*} < \hat{p}_k(p), \tag{HSLP-1}$$

$$q_{j^*} \geq \hat{p}_k(q), \tag{HSLP-2}$$

and assume that q is not uniform over the clusters with non-zero probability. Let $\tilde{p}^{tok}$ and $\tilde{p}^{clu}$ denote the distributions from which token-level p-less and p-less cluster sample the next token, respectively, i.e., after truncation and renormalization.

Claim. *Under the two HSLP conditions,*

$$\tilde{p}^{tok}_{v^*} < p_{v^*} < \tilde{p}^{clu}_{v^*}. \tag{1}$$

Proof. **Part 1: $\tilde{p}^{\text{tok}}_{v^*} = 0$.** By (HSLP-1), v^* falls strictly below the token-level threshold and is filtered out by token-level p-less. After normalization,

$$\tilde{p}^{\text{tok}}_{v^*} = 0 < p_{v^*}, \tag{2}$$

where the strict inequality follows from the assumption $p_{v^*} > 0$.

Part 2: $\tilde{p}^{\text{clu}}_{v^*} > p_{v^*}$. Let $\mathcal{S} = \{j : q_j \geq \hat{p}_k(q)\}$ denote the set of clusters that survive cluster-level filtering; by (HSLP-2), $j^* \in \mathcal{S}$. Under two-stage sampling,

$$\tilde{p}^{\text{clu}}_{v^*} = \tilde{q}_{j^*} \cdot p(v^* \mid \mathcal{C}_{j^*}), \tag{3}$$

where $\tilde{q}_{j^*}$ denotes the normalized cluster probability after filtering, and $p(v^* \mid \mathcal{C}_{j^*})$ is the within-cluster conditional probability of v^*:

$$\tilde{q}_{j^*} = \frac{q_{j^*}}{\sum_{j \in \mathcal{S}} q_j}, \qquad p(v^* \mid \mathcal{C}_{j^*}) = \frac{p_{v^*}}{q_{j^*}}. \tag{4}$$

Substituting (4) into (3) gives

$$\tilde{p}^{\text{clu}}_{v^*} = \frac{p_{v^*}}{Z}, \qquad Z = \sum_{j \in \mathcal{S}} q_j. \tag{5}$$

We have $0 < Z < 1$. For $k \neq 1$, $\hat{p}_k(q) = \left(\sum_j q_j \, q_j^{k-1}\right)^{1/(k-1)}$ is the power mean of order $k-1$ of the non-zero cluster probabilities, weighted by q (for $k = 1$, it is their weighted geometric mean). It therefore lies between the smallest and the largest non-zero cluster probability, strictly so because q is not uniform over its support. Hence the most probable cluster survives ($Z > 0$), and the least probable cluster with non-zero probability is filtered out ($Z < 1$). Therefore,

$$\tilde{p}^{\text{clu}}_{v^*} = \frac{p_{v^*}}{Z} > p_{v^*}. \tag{6}$$

Conclusion. Combining (2) and (6),

$$\tilde{p}^{\text{tok}}_{v^*} = 0 < p_{v^*} < \frac{p_{v^*}}{Z} = \tilde{p}^{\text{clu}}_{v^*}. \qquad \square$$

Remark. In the degenerate case where all clusters survive, i.e., when q is uniform over its support, $Z = 1$ and cluster-level p-less reduces to multinomial sampling.

C Additional analysis of AR image generation

C.1 Comparison of entropy distribution between text and image generation

To characterize how decoding uncertainty differs between language and image generation within the same model architecture, we compare per-token Shannon entropy across text and image generation modes for Janus Pro and Emu3. For each benchmark prompt, the model is tasked with generating both an image and a descriptive image caption conditioned on the same topic. Specifically, the text branch receives the instruction: *"Topic: {prompt}. Generate a descriptive image caption using N tokens."* Both branches are evaluated under identical temperature and CFG settings, ensuring that any observed entropy differences are attributable to the modality rather than generation hyperparameters.

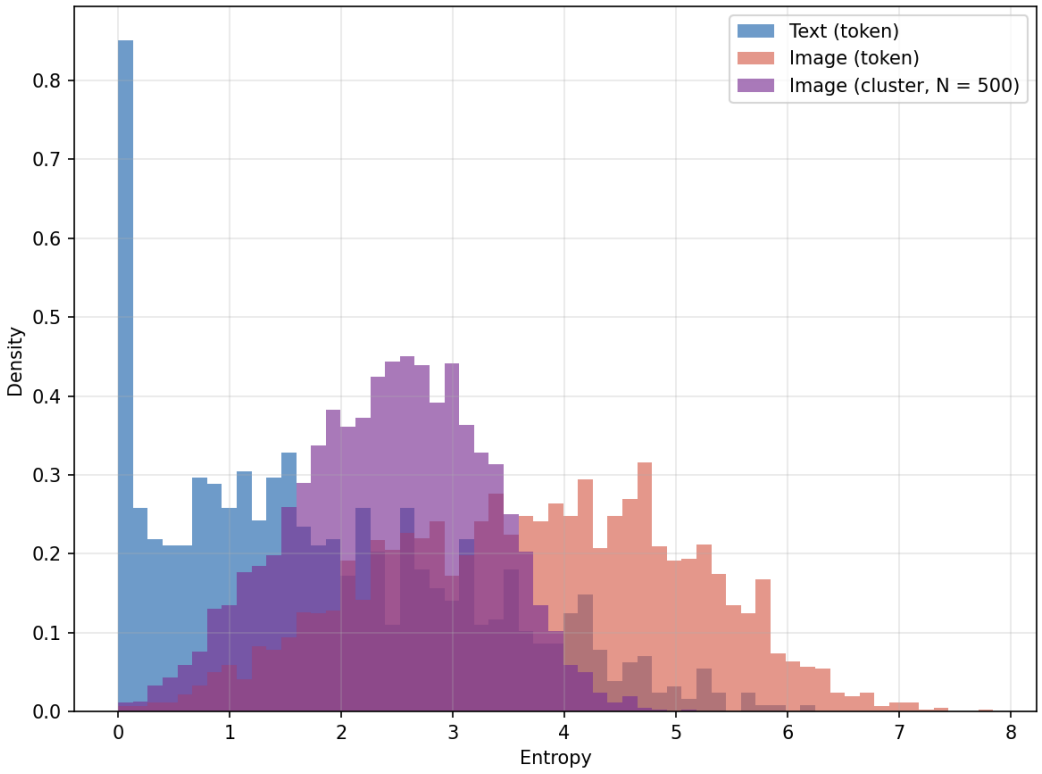


Figure 6: **Shannon entropy distribution over decoding steps for text vs. image generation in Janus Pro, aggregated across 10 prompts.** The image generation mode sustains substantially higher per-token entropy throughout decoding compared to the text generation mode under identical temperature and CFG settings, confirming that AR image models operate in a fundamentally higher-entropy regime.

C.2 Additional illustration of different decoding strategies at high temperature

Figure 7 mirrors Figure 2 but adds 2 more methods for comparison: p-less without clustering and Default cluster (clustering without p-less).

C.3 Robustness of decoding strategies w.r.t. temperature

Figures 8 and 9 illustrate the relationship between temperature, set alignment, and set diversity across different models and decoding strategies, using the same prompt subsets as Section 5.2. The results reveal a fundamental diversity–alignment tradeoff that no decoding strategy fully escapes: higher temperatures generally boost Set Diversity (DINOv2) but degrade Set Alignment (VLM) across all four models. Beyond a temperature of 3.0, most methods fall below the minimum alignment gate (> 3.0). p-less cluster sampling offers the best balance (Section 5.2). Nevertheless, even p-less cluster eventually succumbs to alignment degradation at high temperatures. Dynamic Temperature occupies the opposite end of the spectrum: it maintains the most stable alignment and quality scores and rarely collapses catastrophically, but consistently yields lower peak diversity because it lowers effective temperature when entropy is high. Across all models, its highest-diversity-in-gate points are lower than those of p-less cluster.

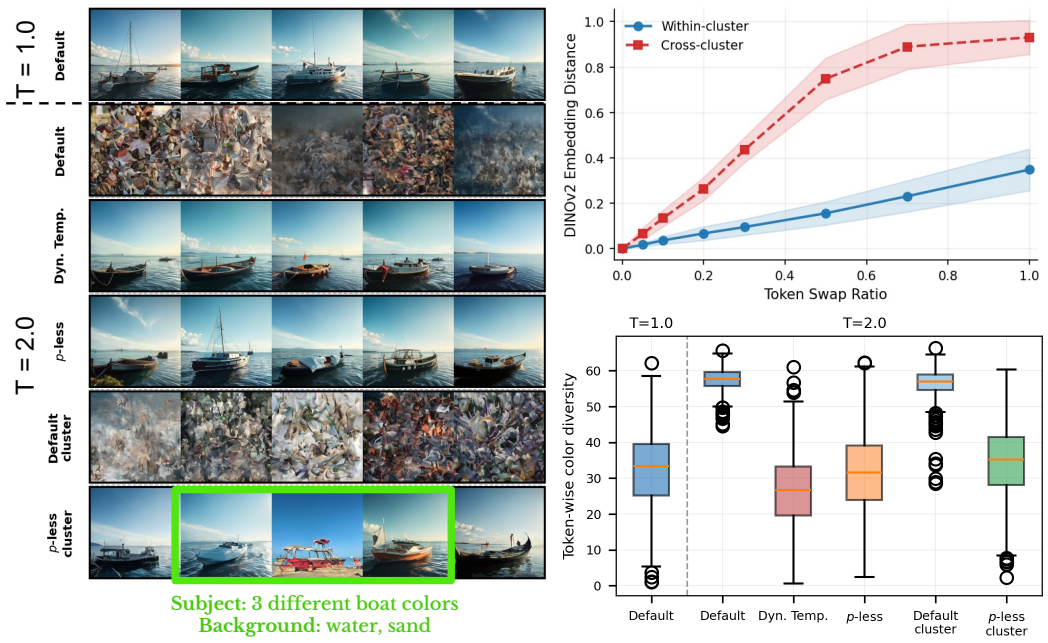


Figure 7: **Qualitative and quantitative comparison of how different decoding strategies affect image-level and token-level diversity (Janus Pro).** At high temperature, both Default and Default cluster sampling admits too many tokens which leads to degeneration. Both Dynamic Temperature and p-less mitigates quality issues at the cost of limited diversity due to low number of distinct tokens. p-less cluster strikes a better balance by combining cluster-level sampling (increasing diversity) and truncation (maintaining quality).

D Implementation Details

All experiments are conducted on NVIDIA A100 GPUs. We provide implementation details for each model and decoding strategy below.

D.1 Diffusion Models

We use Flux.1-Dev and Flux.1-Schnell (Black Forest Labs, 2024) with their default inference settings.[1] For Flux.1-Dev, we apply classifier-free guidance (CFG=3.5) with 50 denoising steps. For Flux.1-Schnell, we use 4 inference steps without CFG.

D.2 Autoregressive Models

Additional details of the AR models employed in our study are provided in Table 2, including visual codebook statistics and underlying codebook architectures.

Default decoding. We use each model's standard decoding strategy: multinomial sampling for Janus Pro-7B and LlamaGen, top-p sampling ($p = 0.9$) for Anole, and top-K sampling ($K = 2048$) for Emu3.

Dynamic temperature (Ma et al., 2025). We follow the recommended hyperparameter settings from the original paper for the two hyperparameters: α, the decay rate of the temperature, and θ, the lower bound of the temperature. Specifically, we use α=3.0 and θ=0.6 for LlamaGen, and α=2.5 and θ=0.6 for Anole, Emu3, and Janus Pro-7B. The latter settings are adopted from those recommended for Lumina-mGPT in the original paper, which the authors note are appropriate for well-trained models that benefit from smoother temperature schedules.

[1] Default inference settings are from `https://github.com/black-forest-labs/flux`.

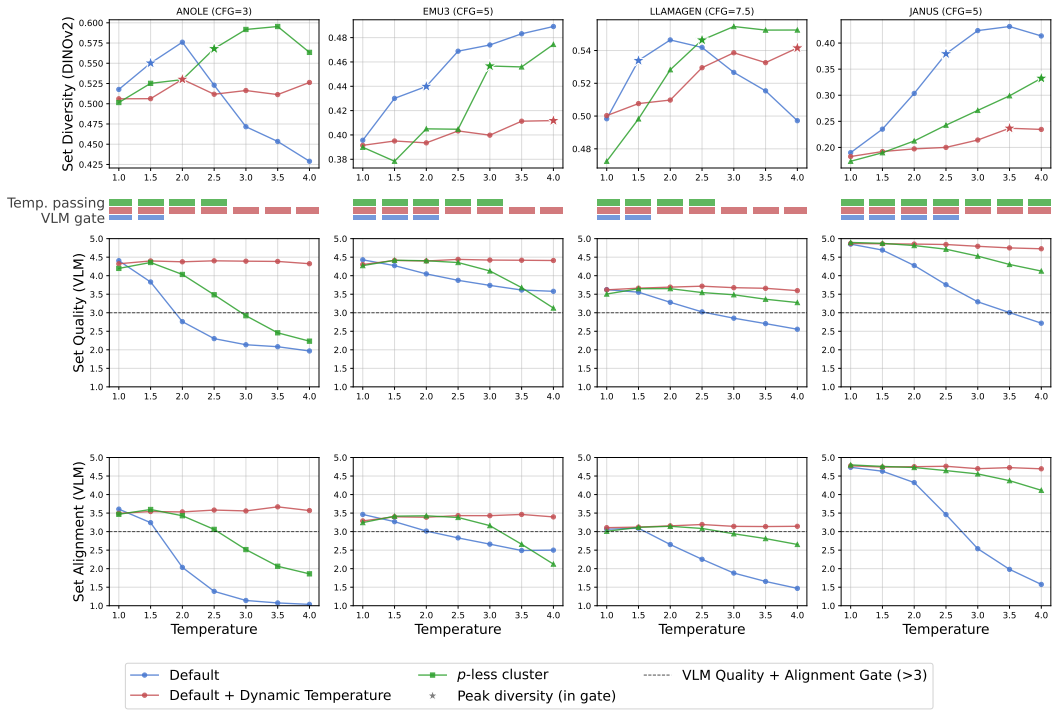


Figure 8: **Robustness of decoding strategies across sampling temperatures with respect to Diversity (top) and Alignment (bottom) on GenEval.** Starred markers indicate the highest-diversity point that remains above the alignment gate (dashed line at VLM Alignment > 3.0). p-less cluster (green) remains above the quality gate across a wider range of temperatures than Default sampling, though no wider than Dynamic Temperature. At the starred points, p-less cluster achieves higher diversity than Dynamic Temperature for all models.

Table 2: Visual codebook statistics across autoregressive models. Models are grouped by shared codebook architectures.

Group	Model	Vocab Size	Emb. Dim.	L2 Norm		Pairwise Dist.	
				Mean	Std.	Mean	Median
VQ-GAN	**Anole**	8,192	256	7.8043	11.6681	13.9432	13.3072
MoVQ-GAN	**Emu3**	32,768	4	0.4961	0.7123	0.8541	0.8110
VQ-VAE	**Janus Pro 7B LlamaGen**	16,384	8	1.0358	0.0325	1.4354	1.4603

p-less cluster sampling. We use a small 2x2 grid for the two hyperparameters: k, the order of the Rényi entropy used for cluster-level truncation, and n, the number of clusters. Specifically, we evaluate $k \in \{1.5, 2.0\}$ and $n \in \{500, 1000\}$.

Hyperparameter tuning protocol. We conduct hyperparameter tuning for each model, decoding strategy, and benchmark using a smaller subset of prompts: 120 prompts from GenEval (20 prompts randomly sampled from 6 categories) and 90 prompts from PartiPrompts (30 prompts randomly sampled from 3 categories). We select the configuration that maximizes diversity metrics while maintaining acceptable image quality and alignment (VLM Quality and Alignment Score > 3.0). The full sweeps are shown in Figures 8 and 9; three reported settings differ from the most diverse acceptable DINOv2 setting in those sweeps. For p-less cluster on LlamaGen/GenEval, the reported T=3.0 has a tuning-subset alignment of 2.94, marginally below the threshold; the most diverse setting that satisfies it is T=2.5 (0.546 vs. 0.555 DINOv2 on the subset). For Default on Emu3/PartiPrompts, T=3.5 also satisfies the threshold and has higher subset DINOv2 than the reported T=3.0 (0.410 vs.

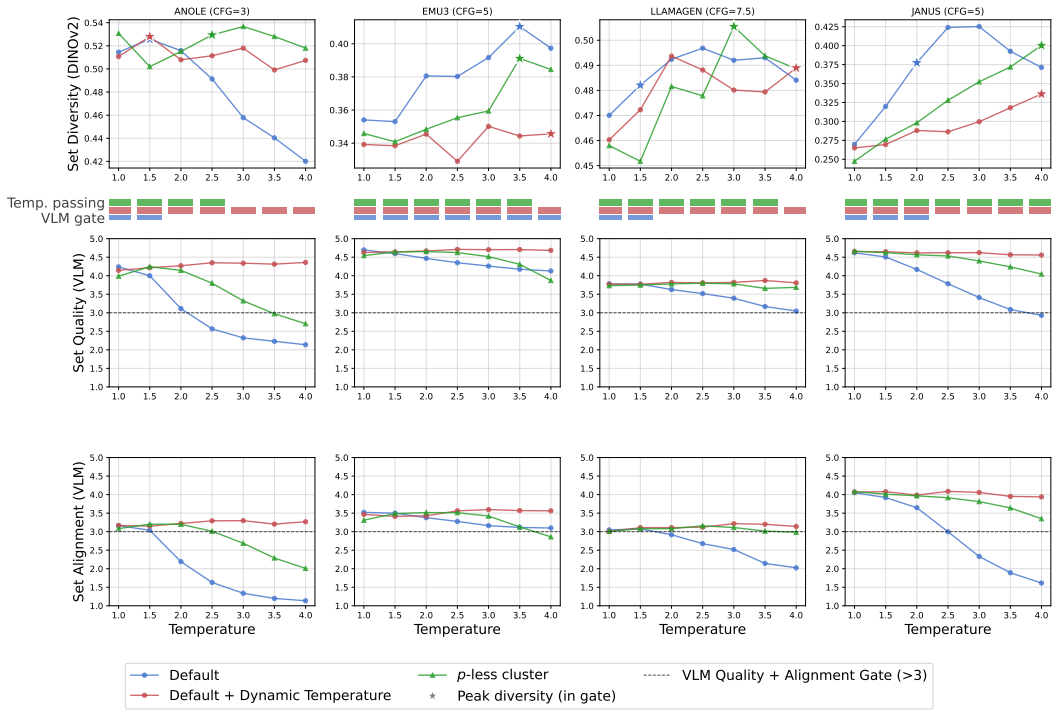


Figure 9: **Robustness of decoding strategies across sampling temperatures with respect to Diversity (top) and Alignment (bottom) on PartiPrompts.** Starred markers indicate the highest-diversity point that remains above the alignment gate (dashed line at VLM Alignment > 3.0). p-less cluster (green) remains above the quality gate across a wider range of temperatures than Default sampling for all models except Emu3, though no wider than Dynamic Temperature. At the starred points, p-less cluster achieves higher diversity than Dynamic Temperature for Emu3, LlamaGen, and Janus Pro, and comparable diversity for Anole.

0.392). For p-less cluster on Anole/PartiPrompts, T=1.0 has marginally higher subset DINOv2 than the reported T=2.5 (0.531 vs. 0.529).

Optimal hyperparameter setting for all methods. Table 3 summarizes the optimal hyperparameters for each model, decoding strategy, and benchmark. For p-less cluster sampling, we find that a smaller number of clusters ($n = 500$) generally provides better robustness to quality degradation across temperatures, while a larger number of clusters ($n = 1000$) can yield higher diversity gains at the cost of stability.

D.3 Licensing and Usage

Dataset	License
GenEval (Ghosh et al., 2023)	MIT
PartiPrompts (Yu et al., 2022b)	Apache-2.0

Table 3: Optimal hyperparameters per model, decoding strategy, and benchmark. T: temperature; α, θ: decay rate and lower bound of dynamic temperature; k: Rényi entropy order; n: number of token clusters; p: nucleus threshold; K: top-K vocabulary truncation.

Model	Method	GenEval	PartiPrompts
Anole	Default	T=1.5, p=0.9	T=1.5, p=0.9
	Dyn. Temp.	T=2.0, α=2.5, θ=0.6	T=1.5, α=2.5, θ=0.6
	p-less Cluster	T=2.5, k=1.5, n=1000	T=2.5, k=1.5, n=500
Emu3	Default	T=2.0, K=2048	T=3.0, K=2048
	Dyn. Temp.	T=4.0, α=2.5, θ=0.6	T=3.0, α=2.5, θ=0.6
	p-less Cluster	T=3.0, k=2.0, n=500	T=3.5, k=2.0, n=500
Janus Pro-7B	Default	T=2.5	T=2.0
	Dyn. Temp.	T=3.5, α=2.5, θ=0.6	T=4.0, α=2.5, θ=0.6
	p-less Cluster	T=4.0, k=1.5, n=500	T=4.0, k=1.5, n=500
LlamaGen	Default	T=1.5	T=1.5
	Dyn. Temp.	T=4.0, α=3.0, θ=0.6	T=2.0, α=3.0, θ=0.6
	p-less Cluster	T=3.0, k=1.5, n=500	T=3.0, k=2.0, n=500

Model	License
Janus Pro (Chen et al., 2025)	MIT (code), DeepSeek Model License (weights)
LlamaGen (Sun et al., 2024)	MIT
Anole (Chern et al., 2024)	Chameleon Research License
Emu3 (Wang et al., 2024)	Apache-2.0
Flux.1-Dev (Black Forest Labs, 2024)	FLUX.1-dev Non-Commercial
Flux.1-Schnell (Black Forest Labs, 2024)	Apache-2.0

E Additional experimental results

E.1 Full quality-alignment-diversity results

Tables 4 and 5 report the complete results on GenEval and PartiPrompts, including quality, alignment, and diversity metrics for all model–strategy pairs.

E.2 Ablation: p-less vs. p-less cluster

Table 8 compares p-less and p-less cluster at the default CFG value across all four models and both benchmarks, complementing the qualitative comparison in Figure 7. p-less cluster improves Vendi diversity in six of the eight settings, with the most substantial gains on Janus Pro-7B (+16.54% on GenEval, +10.12% on PartiPrompts), while differences on the other three models are modest (between -0.81% and $+3.05\%$).

E.3 Effect of cluster size

To investigate the effect of cluster size, we fix $k = 2.0$ and conduct experiments on an extended range of cluster size $\{500, 1000, 1500, 2000\}$. As illustrated in Figure 10, while diversity gains tend to increase with a larger number of clusters, using fewer clusters provides greater robustness to quality degradation across temperatures. This is because a smaller number of clusters induces more aggressive truncation, reducing the variance of the effective candidate set and preventing sampling from low-probability regions that often lead to artifacts. In contrast, a larger number of clusters enables finer-grained control over diversity by preserving more modes of the distribution, but it is also more sensitive to temperature. Retaining low-probability clusters increases diversity, yet it

Table 4: Full results on the **GenEval** benchmark comparing Default, Dynamic Temperature, and p-less cluster decoding strategies. Quality is measured by a VLM judge score, alignment by VLM and CLIPScore, and diversity by DINOv2 and Vendi Score. Diffusion baselines use their default settings and are reported for diversity only. Bold indicates the best value per model per metric.

Model	Decoding Strategy	Quality	Alignment		Diversity	
		VLM	VLM	CLIPScore	DINOv2	Vendi
Flux.1-Dev (default settings)		–	–	–	0.311±0.016	2.336±0.069
Flux.1-Schnell (default settings)		–	–	–	0.270±0.006	2.186±0.024
Anole	Default	3.820±0.022	3.207±0.042	0.305±0.001	0.548±0.006	3.494±0.027
	Dynamic Temperature	**4.369**±0.020	**3.572**±0.047	**0.316**±0.001	0.505±0.007	3.247±0.032
	p-less cluster (Ours)	3.483±0.021	3.038±0.042	0.302±0.001	**0.563**±0.006	**3.549**±0.025
Emu3	Default	4.019±0.023	2.961±0.046	0.287±0.001	**0.441**±0.007	**2.988**±0.030
	Dynamic Temperature	**4.375**±0.020	**3.381**±0.047	**0.301**±0.001	0.405±0.007	2.809±0.029
	p-less cluster (Ours)	4.084±0.022	3.076±0.045	0.294±0.001	**0.441**±0.007	2.972±0.031
Janus Pro-7B	Default	3.760±0.026	3.475±0.038	0.303±0.001	**0.374**±0.006	**2.684**±0.027
	Dynamic Temperature	**4.759**±0.028	**4.762**±0.037	**0.337**±0.001	0.222±0.005	1.994±0.023
	p-less cluster (Ours)	4.097±0.022	4.059±0.033	0.320±0.001	0.328±0.006	2.480±0.026
LlamaGen	Default	3.575±0.023	**3.101**±0.044	**0.296**±0.001	0.526±0.005	3.424±0.026
	Dynamic Temperature	**3.610**±0.023	3.082±0.042	0.295±0.001	0.541±0.006	3.491±0.026
	p-less cluster (Ours)	3.485±0.041	2.942±0.089	0.290±0.001	**0.543**±0.005	**3.511**±0.024

Table 5: Full results on the **PartiPrompts** benchmark comparing Default, Dynamic Temperature, and p-less cluster decoding strategies. Metrics follow the same protocol as Table 4. Bold indicates the best value per model per metric.

Model	Decoding Strategy	Quality	Alignment		Diversity	
		VLM	VLM	CLIPScore	DINOv2	Vendi
Flux.1-Dev (default settings)		–	–	–	0.354±0.013	2.514±0.056
Flux.1-Schnell (default settings)		–	–	–	0.277±0.008	2.227±0.033
Anole	Default	4.154±0.018	3.428±0.035	0.296±0.001	0.535±0.005	3.401±0.024
	Dynamic Temperature	**4.320**±0.025	**3.563**±0.051	**0.297**±0.002	0.531±0.008	3.353±0.037
	p-less cluster (Ours)	3.816±0.024	3.289±0.049	0.290±0.002	**0.542**±0.007	**3.466**±0.030
Emu3	Default	4.684±0.020	3.798±0.049	0.295±0.002	0.364±0.007	2.618±0.032
	Dynamic Temperature	**4.704**±0.017	**3.862**±0.047	**0.295**±0.002	0.360±0.007	2.611±0.031
	p-less cluster (Ours)	4.233±0.027	3.303±0.047	0.287±0.002	**0.400**±0.007	**2.794**±0.031
Janus Pro-7B	Default	3.959±0.041	3.403±0.056	0.288±0.002	0.426±0.007	2.947±0.031
	Dynamic Temperature	**4.503**±0.024	**4.007**±0.045	**0.302**±0.002	0.375±0.007	2.707±0.033
	p-less cluster (Ours)	3.724±0.044	3.179±0.059	0.281±0.002	**0.454**±0.007	**3.080**±0.031
LlamaGen	Default	3.849±0.024	3.421±0.048	0.291±0.002	0.474±0.006	3.192±0.027
	Dynamic Temperature	**3.913**±0.025	**3.479**±0.052	**0.293**±0.002	0.471±0.006	3.164±0.030
	p-less cluster (Ours)	3.831±0.024	3.418±0.048	0.289±0.001	**0.479**±0.006	**3.213**±0.026

simultaneously raises the risk of sampling from unreliable regions of the distribution, resulting in degraded image quality. For most models, we find $n = 500$ is the optimal number of clusters. More details on model-specific hyperparameters are provided in Appendix D.

E.4 Diversity breakdown by PartiPrompts challenge category (all models)

Figure 11 extends Figure 3 in the main text by showing DINOv2 diversity broken down by PartiPrompts challenge category for all four AR models.

Table 6: Full results on the **GenEval** benchmark for Janus Pro-7B at low CFG (CFG = 1.0)

Model	Decoding Strategy	Quality	Alignment		Diversity	
		VLM	VLM	CLIPScore	DINOv2	Vendi
Janus Pro-7B	Default	4.417±0.020	3.987±0.035	**0.319**±0.001	0.442±0.008	2.977±0.034
	Dynamic Temperature	**4.503**±0.018	**4.115**±0.034	0.319±0.003	0.426±0.016	2.912±0.069
	p-less cluster (Ours)	3.804±0.023	3.174±0.042	0.295±0.001	**0.552**±0.006	**3.506**±0.028

Table 7: Full results on the **PartiPrompts** benchmark for Janus Pro-7B at low CFG (CFG = 1.0)

Model	Decoding Strategy	Quality	Alignment		Diversity	
		VLM	VLM	CLIPScore	DINOv2	Vendi
Janus Pro-7B	Default	4.271±0.027	3.358±0.056	0.276±0.002	0.553±0.010	3.489±0.045
	Dynamic Temperature	**4.379**±0.035	**3.376**±0.074	**0.295**±0.003	0.476±0.013	3.142±0.059
	p-less cluster (Ours)	4.053±0.029	3.130±0.056	0.273±0.002	**0.589**±0.010	**3.646**±0.042

E.5 Diversity–alignment trade-off for Janus Pro at low CFG

Table 8: Vendi diversity score comparison of p-less vs. p-less cluster across models and benchmarks at default CFG. p-less cluster values are those in Tables 4 and 5. p-less cluster yields a diversity gain over p-less in most model-dataset settings, with the largest gains on Janus Pro-7B.

Model	Dataset	p-less	p-less cluster	% Gain / Loss
Anole	GenEval	3.578	3.549	−0.81%
	PartiPrompts	3.443	3.466	+0.67%
Emu3	GenEval	2.884	2.972	+3.05%
	PartiPrompts	2.779	2.794	+0.54%
Janus Pro-7B	GenEval	2.128	2.480	+16.54%
	PartiPrompts	2.797	3.080	+10.12%
LlamaGen	GenEval	3.441	3.511	+2.03%
	PartiPrompts	3.230	3.213	−0.53%

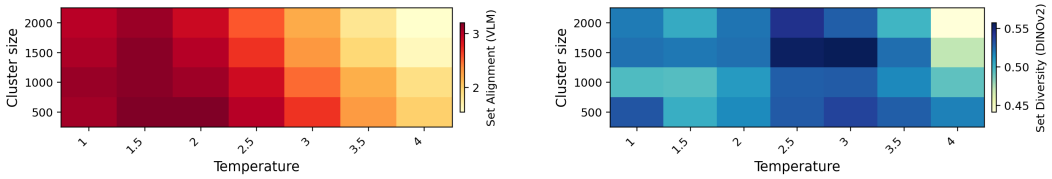
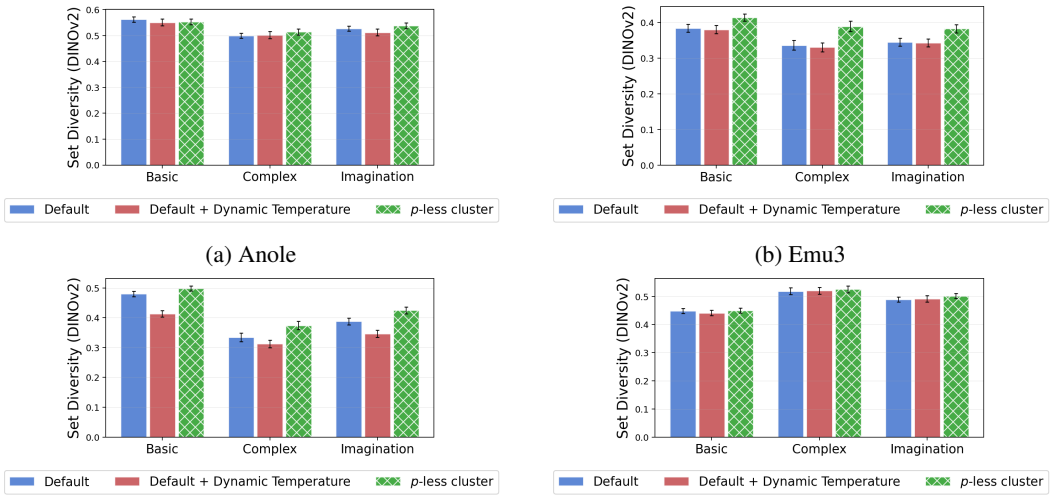


Figure 10: **Effect of number of clusters on prompt alignment (left) and diversity (right) for Anole on PartiPrompts.** Smaller number of clusters increases robustness with respect to image quality at higher temperatures and exhibits more stability in sample diversity.



(a) Anole

(b) Emu3

(c) Janus Pro

(d) LlamaGen

Figure 11: **DINOv2 diversity by PartiPrompts challenge category across all four AR models.** p-less cluster consistently achieves the highest diversity in the *Complex* and *Imagination* categories across all models. For Anole and LlamaGen, the gains extend uniformly across all categories, whereas for Emu3 and Janus Pro the improvements are more concentrated in open-ended prompts.

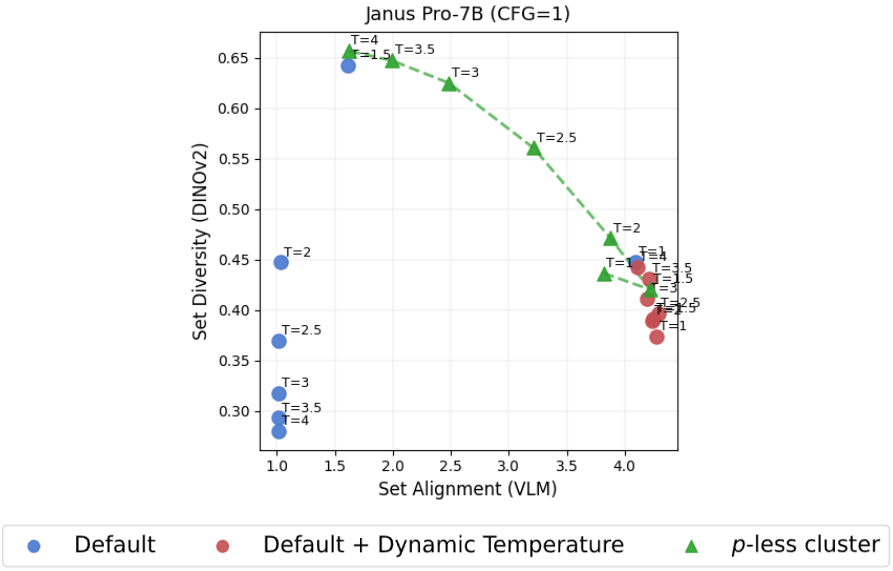


Figure 12: **Diversity–alignment trade-off for Janus Pro at CFG=1.0 on the GenEval tuning subset.** Labels give the sampling temperature. Default's alignment drops to about 1.6 at T=1.5 and to about 1 for $T \geq 2$, whereas p-less cluster reaches a DINOv2 of about 0.56 at T=2.5 with alignment above 3.0 (full-benchmark results in Table 6).

F VLM-AS-JUDGE PROMPT AND PROTOCOL

The following prompt was used for all VLM-as-judge evaluations. The model is instructed to act as an expert image quality evaluator, scoring each generated image on both visual quality and prompt alignment. The protocol ensures consistency and transparency in the evaluation process.

Listing A.1: VLM-as-Judge Evaluation Prompt

You are an expert image quality evaluator. You will assess a generated image against its source prompt with precision and consistency.

Source Prompt

"{prompt_text}"

Evaluation criteria

1. Visual quality (1–5) Assess the intrinsic quality of the image, independent of the prompt:

- 5 – Excellent: Sharp, well-lit, natural colors, zero artifacts, professional-grade composition
- 4 – Good: Minor imperfections (slight blur, minor noise) that don't detract from the image
- 3 – Acceptable: Noticeable issues (soft focus, unnatural lighting, minor distortions) but still coherent
- 2 – Poor: Significant artifacts, distortions, or incoherence that impairs the image
- 1 – Very Poor: Severely corrupted, unrecognizable, or fundamentally broken

2. Prompt alignment (1–5) Assess how faithfully the image represents the prompt. Consider: subject, attributes (color, size, texture), spatial relationships, style, and mood:

- 5 – Perfect: All prompt elements present and accurately depicted
- 4 – Good: Core elements present; minor attributes missing or slightly off
- 3 – Partial: Main subject recognizable but several attributes or details are wrong/missing
- 2 – Poor: Only loosely related to the prompt; major elements absent or incorrect
- 1 – No match: Image bears no meaningful relation to the prompt

Scoring rules

- Score each dimension **independently** -- a technically beautiful image can still score 1 on alignment
- Be **consistent**: the same image shown twice should receive the same scores
- Avoid **leniency bias**: reserve 5s for images that genuinely meet all criteria
- If the prompt is ambiguous, judge alignment against the most reasonable interpretation

Output format Return ONLY a valid JSON object. No preamble, no trailing text.

```
{
  "quality_score": <int 1-5>,
  "alignment_score": <int 1-5>,
  "quality_reasoning": "<one sentence citing specific visual evidence>",
  "alignment_reasoning": "<one sentence citing prompt elements present or missing>"
}
```

G QUALITATIVE EXAMPLES

Figures 13–16 show qualitative comparisons for all 4 models on GenEval and PartiPrompts.

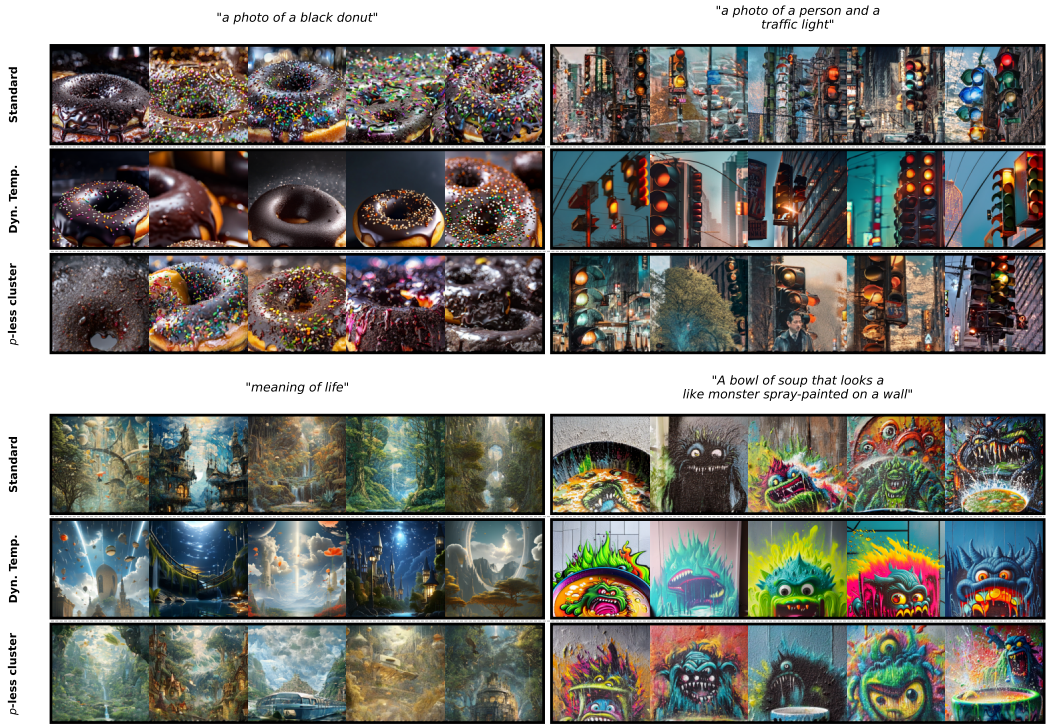


Figure 13: **Qualitative comparison for Emu3.** Hyperparameters match Table 1. Top: GenEval prompts; Bottom: PartiPrompts. Default and Dynamic Temperature produce visually similar samples for the same prompt, reflecting diversity collapse. p-less cluster method generates samples with greater variation in object position, color, and scene composition while preserving prompt fidelity. Additional qualitative examples for the other models are provided in Figures 14–16.

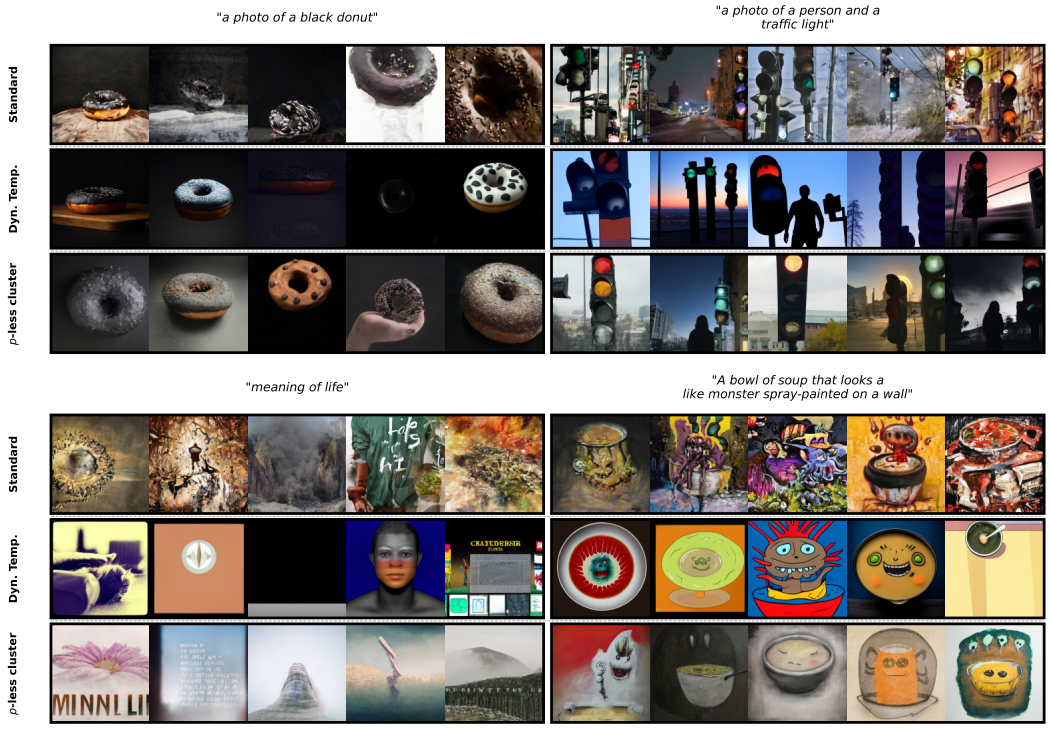


Figure 14: **Qualitative comparison for Anole.** Top: GenEval prompts; Bottom: PartiPrompts. Hyperparameters match Table 4.

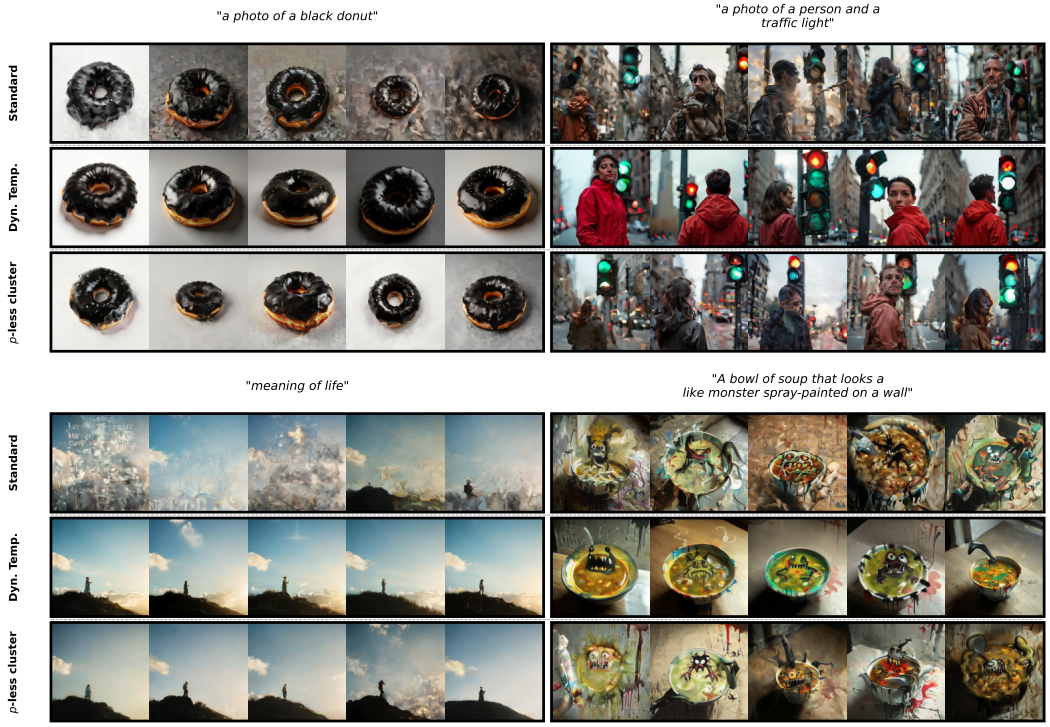


Figure 15: **Qualitative comparison for Janus Pro** at default CFG. Top: GenEval prompts; Bottom: PartiPrompts. Hyperparameters match Table 4.

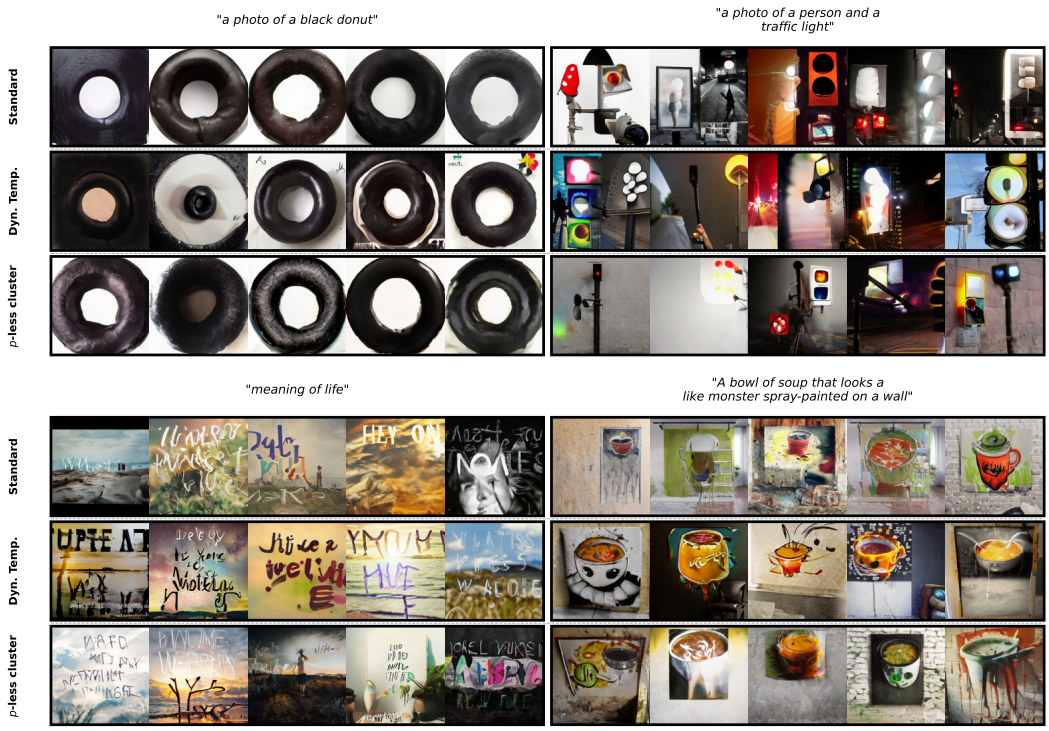


Figure 16: **Qualitative comparison for LlamaGen.** Top: GenEval prompts; Bottom: PartiPrompts. Hyperparameters match Table 4.