Revisiting Greedy Decoding for Visual Question Answering: A Calibration Perspective

Anonymous ACL submission

Abstract

Stochastic sampling strategies are widely adopted in large language models (LLMs) to balance output coherence and diversity. These heuristics are often inherited in Multimodal LLMs (MLLMs) without task-specific justification. However, we contend that stochastic decoding can be suboptimal for Visual Question Answering (VQA). VQA is a closed-ended task with head-heavy answer distributions where uncertainty is usually epistemic, arising from missing or ambiguous visual evidence rather than plausible continuations. In this work, we provide a theoretical formalization of the relationship between model calibration and predictive accuracy, and derive the sufficient conditions for greedy decoding optimality. Extensive experiments provide empirical evidence for the superiority of greedy decoding over stochastic sampling across multiple benchmarks. Furthermore, we propose Greedy Decoding for Reasoning Models, which outperforms both stochastic sampling and standard greedy decoding in multimodal reasoning scenarios. Overall, our results caution against naively inheriting LLMs decoding heuristics in MLLMs and demonstrate that greedy decoding can be an efficient yet strong default for VQA.

1 Introduction

Large Language Models (LLMs) have emerged as the standard for various generative tasks due to their ability to produce diverse and high-fidelity outputs. To navigate the inherent trade-off between coherence and diversity, stochastic sampling are widely employed during inference. Common sampling strategies achieve this balance by truncating the probability tail to prevent degeneration while maintaining output variation (Fan et al., 2018; Holtzman et al., 2020; Nguyen et al., 2024). Alternatively, entropy-dependent strategies attempt to desmooth the model distribution, treating the output as a mixture of the true distribution and uniform noise to improve generation quality (Hewitt et al., 2022).

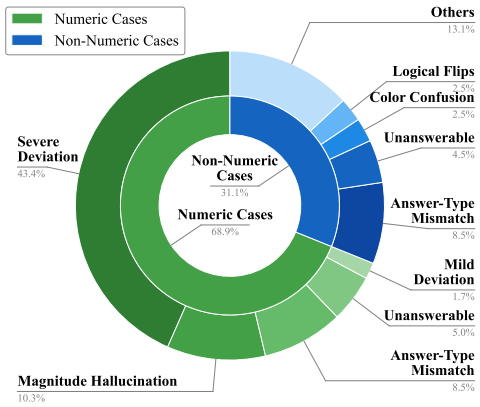


Figure 1: **Failure cases of stochastic sampling on ChartQA.** We sample 177 instances in which greedy decoding produces the correct answer, whereas stochastic sampling yields an incorrect one.

These stochastic decoding heuristics are often adopted for Multimodal LLMs (MLLMs) without task-specific justification. However, we contend that Visual Question Answering (VQA) differs from open-ended text generation, rendering sampling during inference suboptimal. First, most VQA tasks are closed-ended decision problems that require precise outputs. Therefore, the answer distribution is *head-heavy*, dominated by high-frequency tokens such as numbers, colors, or boolean values (Whitehead et al., 2022). For instance, approximately 89% of answers VQAv2 (Antol et al., 2015) with real MS-COCO images are single-token, with boolean responses alone constituting 38% of the distribution. This head-heaviness is strong enough that a trivial "always answer yes" baseline reaches 29.7% accuracy. Second, VQA relies on visual grounding where uncertainty is primarily *epistemic* rather than aleatoric. In open-ended generation, uncertainty is often aleatoric, representing a choice between multiple valid creative trajectories. In contrast, uncertainty in VQA typically stems from missing or ambiguous vi-

sual evidence (Li et al., 2023; Leng et al., 2024; Favero et al., 2024), under which stochastic sampling risks expanding the candidate set toward low-probability and low-agreement distractors instead of grounded, valid answers. For example, the primary error types introduced by stochastic sampling on ChartQA (Masry et al., 2022) are severe deviations (43.4%) for numeric questions and answer-type mismatches (8.5%) for non-numeric ones, whereas semantically equivalent near-misses accounts for only 2.5% of total failures; see Figure 1 and Section 4.4 for a detailed breakdown.

In this paper, we revisit decoding strategies from a calibration perspective. We first review existing calibration metrics (Wang, 2023) and formalize their theoretical relationship to predictive accuracy. We then derive sufficient conditions under which greedy decoding is optimal. Extensive experiments show that these conditions are met by MLLMs on VQA tasks, where greedy decoding consistently outperforms stochastic sampling. To summarize, our contributions are:

1. We provide a theoretical formalization of the relationship between model calibration metrics and predictive accuracy, and derive sufficient conditions under which greedy decoding is optimal, enabling principled decoding strategy selection via a development set;
2. We conduct extensive evaluations across diverse benchmarks and models and provide empirical evidence for the superiority of greedy decoding over stochastic sampling on VQA tasks;
3. We propose *Greedy Decoding for Reasoning Models* (GDRM) to improve the reasoning model performance which anchors answer tokens to a greedy prediction without compromising the reasoning capability.

2 Related Work

Decoding in LLMs. LLMs employ sampling to balance output diversity and coherence. Temperature scaling modulates the distribution sharpness by rescaling logits but does not restrict the vocabulary, allowing low-probability tokens to emerge at higher temperatures (Holtzman et al., 2020). To mitigate degeneration from the unreliable tail, heuristic truncation methods are widely adopted. Top-k sampling (Fan et al., 2018) constrains the candidate set to the k most probable tokens, while top-p (nucleus) sampling (Holtzman et al., 2020) retains the smallest set whose cumulative probability mass exceeds a predefined threshold p. However, these static cutoffs involve a difficult trade-off where restrictive thresholds may omit plausible continuations and lenient settings risk incoherence. This limitation has motivated more principled truncation strategies that adapt to contextual uncertainty. Hewitt et al. (2022) propose ϵ-sampling and entropy-dependent η-sampling to dynamically adjust the candidate set. Most recently, Min-p sampling (Nguyen et al., 2024) scales the truncation threshold relative to the probability of the top token, $p_{\max}$. This approach enforces stricter pruning when the model is confident and permits greater diversity when the distribution is flat, providing a more robust mechanism for balancing coherence and variation. Recent studies have shown advantages of deterministic decoding (Song et al., 2025; Li et al., 2025a). However, these findings remain heuristic and lack rigorous theoretical analysis.

Decoding in MLLMs. MLLMs typically inherit LLM decoding heuristics. However, multimodal generation susceptible to object hallucination, where entities in model outputs lack visual grounding (Li et al., 2023). Recent progress, therefore, focuses on guided or constrained decoding to improve visual grounding. Visual Contrastive Decoding (Leng et al., 2024) contrasts distributions under real versus perturbed images to downweight tokens driven by language priors. Variants along these lines re-score or filter candidates to emphasize visual faithfulness (Ghosh et al., 2024; An et al., 2025; Su et al., 2025). In contrast, our analysis targets the complementary question of deterministic versus stochastic decoding, and thus is applicable on top of these approaches.

3 Theoretical Framework

In this section, we first establish the notation and problem setup (Section 3.1). We then extend existing calibration metrics to strategy-specific metrics (*i.e.*, ECE^{α} and BS^{α}) for any sampling strategy parameterized by α (Section 3.2). Finally, we derive sufficient conditions for greedy optimality using these metrics (Section 3.3).

3.1 Problem Setup.

For an image–question pair (I, x), let $\mathcal{A}$ be a normalized finite answer set (*e.g.,* a VQA vocabulary after string canonicalization). The ground

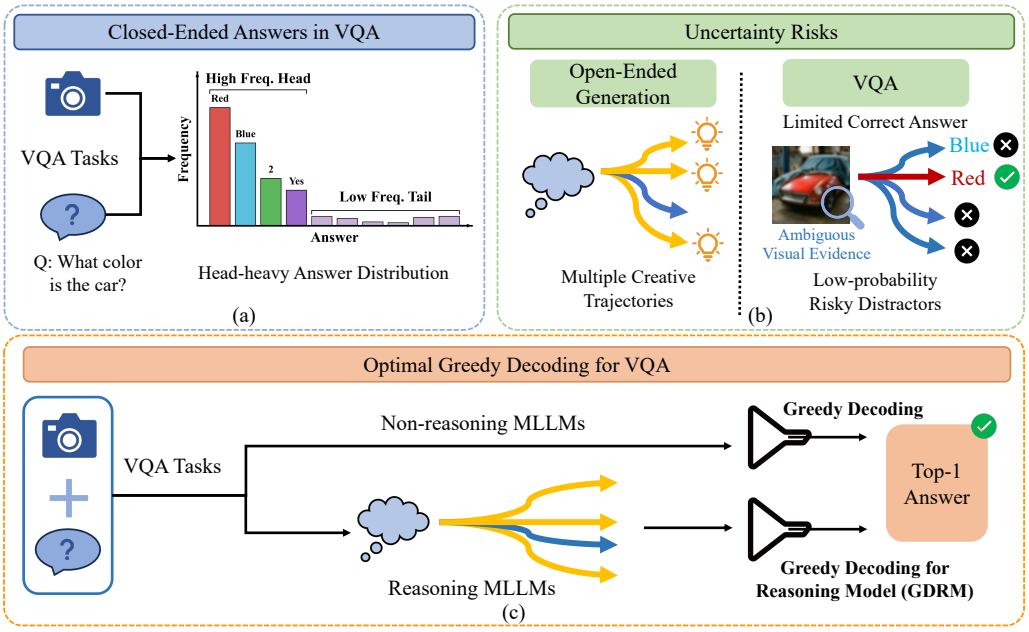


Figure 2: **(a)** VQA Answer distributions are head-heavy (*e.g.*, numbers, boolean values, and colors); **(b)** VQA uncertainty is primarily epistemic; opening the tail risk introducing low-probability distractors; **(c)** GDRM anchors final answer tokens to the greedy prediction while preserving internal reasoning capabilities.

truth *answer-level posterior* is defined as $p(\cdot|I,x) : \mathcal{A} \to [0,1]$ with $\sum_{a\in\mathcal{A}} p(a|I,x) = 1$. Let $q(\cdot|I,x) : \mathcal{A} \to [0,1]$ denote the posterior estimated by a MLLM.

In practice, an answer a is decoded from a sequence of tokens $y_{1:T}$ via a decoding function D, such that $a = D(y_{1:T})$, where y_t is the token generated at step t and T is the sequence length. The answer probability is computed as the joint product of the token probabilities.

Consider an arbitrary candidate selection strategy parameterized by α (*e.g.*, *top-k*, *top-p*, or *min-p*), where α represents the corresponding hyperparameter value (k for *top-k*, p for *top-p* and p_{base} for *min-p*). For a given input pair (I,x), let $S_t^\alpha(I,x,y_{<t})$ be the set of candidate tokens selected by the strategy at step t. Under this strategy, tokens are sampled from S_t^α proportionally to the model's original likelihood

$$y_t \sim \underbrace{\frac{p_\theta(\cdot \mid y_{<t}, I, x)\ \mathbf{1}\{\cdot \in S_t^\alpha\}}{\sum_{y\in S_t^\alpha} p_\theta(\cdot \mid y_{<t}, I, x)}}_{:=p_\theta^\alpha(y|I,x,y_{<t})}.$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and p_θ^α denotes the resulting sampling distribution.

Ideally, α is chosen to maximize the expectation of producing correct predictions. This objective is defined as the following optimization problem:

$$\max_\alpha J(\alpha) := \mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}[p(a|I,x)] \qquad (1)$$

where $q^\alpha(a|I,x) = \sum_{D(y_{1:T})=a} p_\theta^\alpha(y_{1:T}|I,x)$ is the probability of selecting answer a under the strategy parameter α. Note that optimization of $J(\alpha)$ is generally intractable in practice due to the unknown ground truth distribution p. In the following sections, we identify the sufficient conditions under which the greedy strategy (taking the answer with the highest confidence) is optimal, and connect them to the calibration metrics. We denote the corresponding answer generated by the greedy strategy as a^1.

3.2 Calibration Measurement

Several metrics exist for measuring the calibration of deep learning models (Wang, 2023). In this work, we adopt two widely applied measures: Expected Calibration Error (ECE) (Guo et al., 2017) and the Brier Score (BS) (Brier, 1950) (definition in Appendix A.1). We define strategy-specific calibration metrics ECE^α and BS^α as follows.

ECE^α. We define ECE^α directly in terms of the distributions p and q as the expected absolute error

over the samples generated by strategy α:

$$\mathrm{ECE}^{\alpha} := \mathbb{E}_{I,x,a\sim q^{\alpha}(\cdot|I,x)} \left[|q(a|I,x) - p(a|I,x)| \right], \tag{2}$$

$$\mathrm{ECE}^{1} := \mathbb{E}_{I,x} \left[|q(a^1|I,x) - p(a^1|I,x)| \right], \tag{3}$$

where ECE^1 is the calibration error of the greedy strategy (taking the highest-confidence token).

BS^{α}. Similarly, we define the strategy-specific Brier Score as the expected normalized squared difference:

$$\mathrm{BS}^{\alpha} := \mathbb{E}_{I,x,a\sim q^{\alpha}(\cdot|I,x)} \left[\frac{|q(a|I,x) - p(a|I,x)|^2}{q(a|I,x)} \right], \tag{4}$$

$$\mathrm{BS}^{1} := \mathbb{E}_{I,x} \left[\frac{|q(a^1|I,x) - p(a^1|I,x)|^2}{q(a^1|I,x)} \right], \tag{5}$$

where BS^1 is the calibration error of the greedy strategy (always taking the token with the highest confidence). Note that in Eq. 10 we substitute the denominator $p(a|I,x)$ in the original definition of BS (Appendix A.1) with $q(a|I,x)$ (see Appendix A.2 for details).

3.3 Greedy Optimality

Our primary theoretical contribution is summarized in the following theorem.

Theorem 3.1. *Let ECE^{α}, ECE^1, BS^{α}, and BS^1 be defined as in (2), (3), (10) and (11). Greedy decoding strategy is optimal if, for any α that does not correspond to the greedy strategy, at least one of the following conditions holds:*

① $G_1^{\alpha} := \mathbb{E}_{I,x,a\sim q^{\alpha}}[q(a^1|I,x) - q(a|I,x)] - ECE^1 - ECE^{\alpha} \geq 0$

② $G_2^{\alpha} := \mathbb{E}_{I,x,a\sim q^{\alpha}}\left[q(a^1|I,x) - \frac{1+q^2(a|I,x)}{2q(a|I,x)}\right] - ECE^1 + \frac{BS^{\alpha}}{2} \geq 0$

The proof of Theorem 3.1 is in Appendix A.3.

Theorem 3.1 suggests that the greedy decoding is optimal under different conditions. Condition ① can be satisfied when model assigns high confidence to the top-1 answer while maintaining low confidence for others (*i.e.*, large $\mathbb{E}_{I,x,a\sim q^{\alpha}}[q(a^1|I,x) - q(a|I,x)]$), and is well calibrated across all answers (*i.e.*, low ECE^1 and ECE^{α}). These requirements align with observed VQA behavior: well-calibrated models exhibit bimodal confidence, with correct predictions concentrated at high confidence and incorrect ones at low confidence, enabling accuracy gains via abstention on uncertain cases (Whitehead et al., 2022; Khan and Fu, 2024). Condition ② is easier to satisfy when the top-1 prediction is both confident and relatively well-calibrated (*i.e.*, high $q(a^1|I,x)$ and low ECE^1), and lower-ranked ones are substantially less calibrated (*i.e.*, large BS^{α} for sampling with parameter α).

4 Experiments

4.1 Benchmark

Datasets. We use 6 English benchmarks: (1) MMMU (Yue et al., 2024), a multidisciplinary multimodal benchmark spanning STEM, humanities, and professional domains; (2) ChartQA (Masry et al., 2022), which tests reasoning over charts and plots; (3) BLINK (Fu et al., 2024), a VQA dataset containing visual commonsense problems; (4) MM-HallBench (Sun et al., 2024), a multimodal hallucination benchmark; (5) MMLU (Hendrycks et al., 2021), a text-only multiple-choice QA benchmark; and (6) CapArena (Cheng et al., 2025), a detailed image captioning benchmark. All results are reported on the official validation splits, which contain 900 (MMMU), 1,920 (ChartQA), 1,901 (BLINK), 96 (MM-HallBench), and 1,531 (MMLU) examples, and 600 images (CapArena).

Models. We evaluate three open-source MLLMs: Qwen2.5-VL (3B and 7B) (Bai et al., 2025b), LLaVA-v1.5 (7B) (Liu et al., 2023), and Qwen3-VL-Thinking (4B) (Bai et al., 2025a). All models are evaluated in the zero-shot setting.

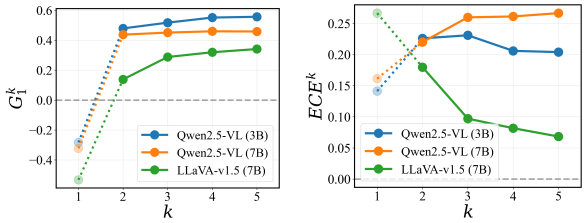
Evaluation metrics. Overall accuracy is reported on all datasets except MM-HallBench and CapArena. For multiple-choice questions, accuracy is computed via exact match; for free-form questions, we use normalized soft matching. For stochastic sampling, results are averaged using four different random seeds. Prompt formatting and inference details are in Appendix B. We report Answer Score and Hallucination Rate for MM-HallBench, and Average Score and Length for CapArena (detailed definitions are in Appendix B.5).

4.2 Baseline

Decoding strategies. We compare eight decoding strategies: greedy decoding, temperature, top-k (Holtzman et al., 2020), top-p (Holtzman et al., 2020), min-p (Nguyen et al., 2024), ε-Sampling, η-Sampling (Hewitt et al., 2022), and Beam Search.

Table 1: Accuracy (in %) of Beam Search on ChartQA benchmark using Qwen2.5-VL (3B).

Method	b=3	b=5	b=10
Beam Search	80.99	81.41	81.30
Greedy		**83.12**	



(a) G_1^k versus k (b) ECE^k versus k

Figure 3: Empirical evidence Theorem 3.1 on ChartQA using Qwen2.5-VL (3B and 7B), and LLaVA-v1.5 (7B). (a) G_1^k ($\alpha := k$ for *top-k*) is non-negative for all $k > 1$. (b) Expected calibration errors ECE^k ($\alpha := k$ for *top-k*) across different token rank k.

Hyperparameters. For each strategy, we sweep a compact grid of commonly used hyperparameters and report the best setting on the validation split. For stochastic sampling methods, we vary temperature T jointly with the strategy-specific hyperparameters; greedy decoding requires no sweep. All sampling baselines in Table 7 use the officially recommended decoding parameters for Qwen3-VL Thinking models, as specified in the Qwen3-VL Technical Report (Bai et al., 2025a). Hyperparameter details are provided in Appendix B.3.

4.3 Main Results

Table 2, Table 3, and Table 4 compare greedy decoding with stochastic sampling on MMMU, ChartQA, and BLINK using Qwen2.5-VL (3B and 7B) and LLaVA-v1.5 (7B), respectively. Greedy decoding consistently outperforms stochastic strategies, with the margin over the best stochastic method exceeding the standard deviation in 24 of 27 settings. Notebly, for Qwen2.5-VL, greedy dominates all settings with gap-to-deviation ratios of up to 18.8×. The only exception is a marginal gain from Top-p sampling for LLaVA-v1.5 (7B) on MMMU (35.67% vs. 34.78%). As predicted by Theorem 3.1, this stems from the model's high top-1 calibration error (ECE^1; Figure 3b), validating our theorem's diagnostic value (see Section 4.4 for detailed analysis). Table 1 compares beam search with greedy decoding using Qwen2.5-VL (3B) on ChartQA, where answers are multi-token rather than multiple-choices (*i.e.*, `A/B/C/D`). Greedy decoding achieves higher accuracy than beam search despite a lower computational budget. Finally, human evaluation of ChartQA outputs by two researchers yields high agreement with string-match labels (Cohen's $\kappa = 0.95$), suggesting our results are not driven by string-matching artifacts.

Figure 4 shows that stochastic sampling is highly sensitive to temperature across models and datasets. Specifically, performance degrades rapidly as temperature increases, reflecting an inverse relationship between randomness and accuracy that aligns with the head-heavy answer distributions in VQA. Conversely, as temperature decreases, sampling approximates greedy decoding and performance improves. This convergence toward greedy decoding as the optimal limit is precisely the phenomenon our theoretical framework explains. Overall, these results indicate that greedy decoding is an effective and robust decoding strategy for VQA.

4.4 Analytical Results

Answer Distribution. We conduct a statistical analysis of the answer distributions across VQA and text-only QA datasets. MMMU, BLINK, and MMLU utilize a multiple-choice answer format, inducing answer distributions concentrated over specific tokens `A/B/C/D`. Although ChartQA permits more diverse responses, its answer distribution exhibits a similar highly concentrated pattern. Specifically, 75.16% of answers are numerical. Among the non-numerical ones, the vast majority fall into limited categories: 66.04% are single words, 23.48% represent country names, and 11.74% are binary responses. These results provide empirical evidence that VQA answer distributions are head-heavy, dominated by high-frequency tokens such as numbers, colors, and binary responses.

Empirical Evidence of Greedy Optimality. We demonstrate empirical evidence of Theorem 3.1 using the top-k sampling. Figure 3a visualizes G_1^k (*i.e.*, G_1^α where $\alpha := k$) across three models on ChartQA benchmark. We can observe that $G_1^k > 0$ holds for all $k > 1$. As a result, greedy decoding is provably the optimal decoding strategy under Condition ① in Theorem 3.1, which is consistent with the experiment results demonstrated in Table 2, 3 and 4. We observe similar behavior for other sampling methods. For instance, using top-p sampling ($p = 0.9$) with Qwen2.5-VL (3B) on ChartQA

Table 2: Accuracy (mean ± std, in %) of different decoding and sampling strategies on MMMU, ChartQA, and BLINK using Qwen2.5-VL (3B), averaged over 3 random seeds. Best and second-best are **bolded** and underlined.

Method	MMMU			ChartQA			BLINK		
	τ=0.7	τ=1.0	τ=2.0	τ=0.7	τ=1.0	τ=2.0	τ=0.7	τ=1.0	τ=2.0
Temp' Only	43.00 ± 0.67	41.11 ± 0.81	33.67 ± 0.91	79.58 ± 0.40	76.14 ± 0.32	48.33 ± 1.60	32.67 ± 1.09	28.35 ± 0.97	23.72 ± 1.06
Top-k	42.53 ± 1.29	41.08 ± 0.87	33.42 ± 1.39	79.61 ± 0.37	76.11 ± 0.26	51.89 ± 0.41	32.67 ± 1.20	30.82 ± 0.96	26.35 ± 0.59
Top-p (nucleus)	42.69 ± 0.43	41.30 ± 1.14	28.20 ± 1.83	81.13 ± 0.43	78.93 ± 0.47	47.71 ± 1.52	31.61 ± 1.28	31.67 ± 0.74	27.14 ± 0.56
Min-p	42.64 ± 1.58	40.20 ± 0.96	28.64 ± 0.99	79.62 ± 0.36	76.09 ± 0.49	45.66 ± 2.59	33.30 ± 0.38	30.98 ± 0.75	29.93 ± 1.32
ε-Sampling	42.45 ± 1.90	40.50 ± 1.34	35.03 ± 1.45	79.91 ± 0.32	76.53 ± 0.51	52.46 ± 1.63	31.09 ± 1.04	31.19 ± 0.65	26.25 ± 0.42
η-Sampling	43.00 ± 1.76	40.44 ± 0.62	28.09 ± 1.72	79.66 ± 0.34	76.46 ± 0.43	48.16 ± 0.56	31.09 ± 0.91	31.04 ± 0.27	25.51 ± 0.95
Greedy		**45.44**			**83.12**			**35.32**	

Table 3: Accuracy (mean ± std, in %) of different decoding and sampling strategies on MMMU, ChartQA, and BLINK using Qwen2.5-VL (7B), averaged over 3 random seeds. Best and second-best are **bolded** and underlined.

Method	MMMU			ChartQA			BLINK		
	τ=0.7	τ=1.0	τ=2.0	τ=0.7	τ=1.0	τ=2.0	τ=0.7	τ=1.0	τ=2.0
Temp' Only	47.78 ± 0.74	45.67 ± 0.88	40.22 ± 1.02	78.96 ± 0.36	77.97 ± 0.41	58.38 ± 1.28	37.56 ± 0.83	35.61 ± 0.79	32.25 ± 0.94
Top-k	47.78 ± 0.91	46.00 ± 0.76	38.22 ± 1.31	78.96 ± 0.33	77.92 ± 0.35	61.30 ± 0.96	37.56 ± 0.88	35.61 ± 0.82	35.35 ± 0.67
Top-p (nucleus)	46.56 ± 0.69	46.33 ± 0.92	40.33 ± 1.17	80.00 ± 0.39	79.69 ± 0.44	67.60 ± 1.12	38.77 ± 0.71	37.45 ± 0.68	34.03 ± 0.73
Min-p	47.33 ± 1.12	44.11 ± 0.95	44.33 ± 0.86	71.41 ± 1.84	71.56 ± 1.76	71.51 ± 1.69	29.40 ± 1.21	30.30 ± 1.08	30.25 ± 1.02
ε-Sampling	49.67 ± 1.36	45.44 ± 1.01	40.33 ± 1.09	80.16 ± 0.31	78.18 ± 0.48	69.79 ± 1.24	37.30 ± 0.76	35.77 ± 0.71	31.78 ± 0.81
η-Sampling	49.56 ± 1.28	44.67 ± 0.83	39.11 ± 1.26	80.16 ± 0.34	78.73 ± 0.46	66.20 ± 1.07	37.30 ± 0.79	35.82 ± 0.64	33.82 ± 0.77
Greedy		**52.56**			**82.13**			**41.56**	

yields $G_1^p = 0.173 > 0$ (*i.e.*, G_1^α where $\alpha := p$).

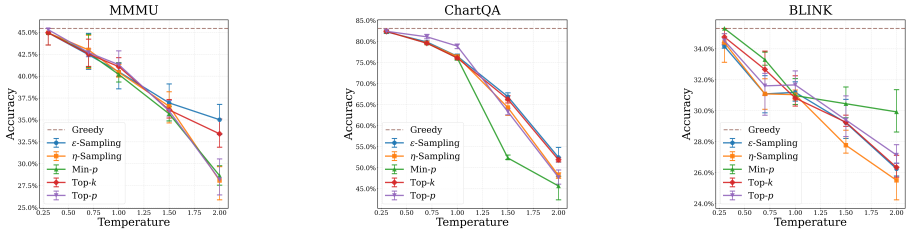
Failure Analysis. We investigate the failure modes of stochastic sampling strategies by analyzing 177 cases where greedy decoding yields correct answers while stochastic decoding fails. Figure 1 shows the detailed failure modes for both numeric and non-numeric cases. We can observe that numeric questions account for the majority of failure cases, with more than half of them exhibiting severe deviations (>20% error) rather than minor inaccuracies. Further qualitative inspection reveals other specific failure modes, such as order-of-magnitude hallucinations (*e.g.*, 2.6 → 26000), answer-type mismatch (*e.g.*, 79 → `"Yes"`), and erroneous abstentions (*e.g.*, 1.0 →`"Unanswerable"`). Non-numeric failures exhibit similar tail-drift artifacts, including answer-type mismatch (*e.g.*, `"Disapprove"` → 53), color confusion (*e.g.*, `"Light Blue"` → `"Gray"`), and logical flips (*e.g.*, `"Yes"`→`"No"`). These failure modes indicate that VQA distributions are head-heavy, concentrating valid answers within the high-probability region. Consequently, stochastic errors represent low-probability tail drift rather than semantically correct alternatives.

Hallucination. We evaluate different decoding strategies on the MM-HallBench using Qwen2.5-VL (3B). Table 5 compares answer score and hallucination rate under different sampling and deterministic decoding methods including beam search and greedy decoding. We can observe that, consistent with other VQA benchmarks, greedy decoding achieves the highest average answer score and lowest hallucination rate. This performance validates our motivation that stochastic sampling expands the candidate set into the low-probability tail, frequently associated with ambiguous visual evidence and therefore at a higher risk of hallucination. Recent work (Li et al., 2025b) further corroborates this, empirically showing that greedy decoding consistently outperforms stochastic sampling across models on the POPE benchmark (Rohrbach et al., 2018), even when coarse binary metrics obscure the full quality difference (Table 10; Appendix C.2).

Calibration Error versus Model Performance. We calculate the calibration errors of MLLMs and analyze their relation with model performance. Figure 3b illustrates the rank-conditioned calibration errors, denoted as ECE^k (G_1^α where $\alpha := k$). We observe distinct calibration errors across models: Qwen2.5-VL (3B and 7B) exhibits strong calibration at the top-1 rank, and calibration degrades at larger k. Conversely, LLaVA-1.5 (7B) displays an inverse trend, where calibration error is high at $k = 1$, and decreases as k increases. The high calibration error at the top-1 token of LLaVA-1.5 (7B) explains that the greedy decoding outperforms the optimal stochastic sampling method by an absolute

Table 4: Accuracy (mean ± std, in %) of different decoding and sampling strategies on MMMU, ChartQA, and BLINK using LLaVA-1.5 (7B), averaged over 3 random seeds. Best and second-best are **bolded** and underlined.

Method	MMMU			ChartQA			BLINK		
	τ=0.7	τ=1.0	τ=2.0	τ=0.7	τ=1.0	τ=2.0	τ=0.7	τ=1.0	τ=2.0
Temp' Only	32.44 ± 0.84	32.89 ± 0.91	31.44 ± 1.02	21.93 ± 0.73	19.69 ± 0.66	8.80 ± 1.18	38.35 ± 0.72	38.77 ± 0.69	35.61 ± 0.88
Top-k	31.33 ± 0.97	29.22 ± 1.08	29.11 ± 1.14	22.03 ± 0.61	18.33 ± 0.79	9.69 ± 1.05	38.35 ± 0.67	39.77 ± 0.64	35.66 ± 0.81
Top-p (nucleus)	32.33 ± 0.76	**35.67 ± 0.88**	31.11 ± 0.95	22.03 ± 0.58	19.69 ± 0.63	9.74 ± 0.96	38.82 ± 0.61	39.45 ± 0.62	36.14 ± 0.74
Min-p	32.44 ± 1.03	32.00 ± 0.92	29.67 ± 1.21	21.51 ± 0.81	18.18 ± 0.72	2.97 ± 1.43	38.35 ± 0.75	38.77 ± 0.71	33.77 ± 0.97
ε-Sampling	32.56 ± 0.89	31.22 ± 0.96	28.22 ± 1.17	21.51 ± 0.76	18.18 ± 0.68	11.77 ± 1.09	38.24 ± 0.69	38.56 ± 0.66	34.24 ± 0.86
η-Sampling	34.78 ± 0.93	32.11 ± 0.85	28.44 ± 1.10	21.09 ± 0.72	19.01 ± 0.65	2.76 ± 1.51	37.08 ± 0.83	37.08 ± 0.79	37.08 ± 0.71
Greedy	34.78			**23.54**			**39.93**		



(a) MMMU (b) ChartQA (c) BLINK

Figure 4: Accuracy of different decoding/sampling strategies on MMMU, ChartQA and BLINK benchmarks using Qwen2.5-VL (3B) with different temperatures.

Table 5: Average Score and Hallucination Rate of different decoding/sampling strategies on MM-HallBench using Qwen2.5-VL (3B).

Strategy	Score ↑	Hallucination Rate (%) ↓
Temp' Only	3.48	36
Top-k	3.38	42
Top-p (nucleus)	3.58	36
Min-p	3.32	43
ε-Sampling	3.36	35
η-Sampling	3.57	38
Beam Search	3.54	**34**
Greedy	**3.64**	**34**

Table 6: Performance of different decoding strategies ($\tau = 1.0$) on MMLU and CapArena benchmarks using Qwen2.5-VL (3B) and (7B).

Method	MMLU	CapArena	
	Acc (%)	Avg. Score	Avg. Length
Temp' Only	71.23	-7.17	156.34
Top-k	71.23	5.17	158.33
Top-p (nucleus)	71.23	2.17	155.06
Min-p	72.50	5.83	153.77
ε-Sampling	71.23	0.67	156.53
η-Sampling	71.79	0.83	160.55
Greedy	**73.21**	**11.83**	148.75

gain of 1.51% for LLaVA-1.5 (7B), compared to larger gains of 1.99% and 1.97% for Qwen2.5-VL 3B and 7B, respectively. Empirically, the calibration error results suggest that models with superior top-1 calibration benefit more significantly from greedy decoding, as their high-confidence predictions at top-1 token are more reliable.

4.5 Text-only QA and Open-ended Generation

Although motivated by the head-heavy nature of close-ended VQA answer distributions, our findings extend beyond this specific setting in two key directions. First, the derivation of greedy optimality in Theorem 3.1 depends solely on the probabilistic structure of the answer-level posterior and remains mathematically valid when visual input is omitted. As a result, greedy decoding is optimal for text-only QA provided either Condition ① or Condition ② holds. Evaluations on the MMLU benchmark (Table 6) confirm this generalizability, with greedy decoding outperforming stochastic strategies. Specifically, G_1^k remains positive for all $k > 1$ across the MMLU dataset (see Appendix C.1 for details). Second, we demonstrate generalizability to open-ended generation by evaluating detailed image captioning on CapArena (Cheng et al., 2025). As shown in Table 6, greedy decoding achieves the highest average score, despite slightly shorter outputs. These findings suggest that for factual image description where accuracy is prioritized, greedy decoding remains a strong default.

5 Greedy Decoding for Reasoning Model

Recent advancements in MLLMs with reasoning capabilities, such as Qwen3-VL-Thinking (Bai et al., 2025a), have demonstrated superior performance compared to non-reasoning models on

Table 7: Accuracy (in %) of different decoding/sampling strategies on MMMU, ChartQA and BLINK benchmarks using Qwen3-VL-Thinking (4B). Baseline: Stochastic Sampling and Beam Search. GDRM: Greedy Decoding for Reasoning Models. Greedy is the Greedy Decoding Baseline.

	MMMU		**ChartQA**		**BLINK**	
Method	Baseline	GDRM	Baseline	GDRM	Baseline	GDRM
Temp' Only	53.22	54.44 (+1.22)	80.94	81.09 (+0.15)	46.08	48.29 (+2.21)
Top-k	54.11	54.78 (+0.67)	80.78	81.46 (+0.68)	44.77	47.71 (+2.94)
Top-p (nucleus)	54.56	54.67 (+0.11)	81.56	81.82 (+0.26)	45.13	45.98 (+0.85)
Min-p	53.78	55.33 (+1.55)	77.71	78.02 (+0.31)	45.29	45.40 (+0.11)
ε-Sampling	54.22	54.78 (+0.56)	80.62	80.78 (+0.16)	46.61	46.45 (-0.16)
η-Sampling	51.89	54.22 (+2.33)	81.25	81.41 (+0.16)	47.13	47.66 (+0.53)
Beam Search	59.78	60.92 (+1.14)	81.51	81.56 (+0.05)	47.40	48.39 (+0.99)
Greedy	54.11		81.46		45.71	

various multimodal tasks. Conventionally, it is believed that introducing uncertainty is essential for the reasoning process, and thus greedy decoding is often considered suboptimal for these models (Naik et al., 2024). However, motivated by Theorem 3.1, we demonstrate that multimodal reasoning can still benefit from greedy decoding. Recent study shows that effective reasoning correlates with increasing token probabilities over the generation trajectory (Liu et al., 2025). Thus, while the internal reasoning process can exhibit significant stochasticity, the final outputs are often produced with high confidence (large $q(a^1|I, x)$), following sufficient deliberation. For VQA tasks with unique ground-truth answers, a high-confidence prediction that aligns with the correct answer results in a low calibration error ECE[1], which facilitates the satisfaction of both conditions outlined in Theorem 3.1.

Following the aforementioned theoretical insight, we propose GDRM. Specifically, by incorporating the generated reasoning trace r into the input context, we transform the original reasoning VQA task (I, x) into a standard prediction task $(I, x+r)$. We then apply greedy decoding to generate the final answer tokens given this augmented context. The formal definition of our proposed GDRM is given as follows:

$$\begin{cases} r \sim \textsc{Sampling}(p_\theta(\cdot \mid I, x)), & \text{reasoning tokens} \\ \hat{y} = \textsc{Greedy}(p_\theta(\cdot \mid I, x + r)), & \text{answer tokens} \end{cases}$$

where r denotes the reasoning tokens, $\hat{y}$ denotes the answer tokens, SAMPLING denotes any given sampling strategy, and GREEDY denotes greedy decoding strategy.

We evaluate GDRM on VQA benchmarks using Qwen3-VL-Thinking (4B) model. As shown in Table 7, GDRM consistently outperforms standard stochastic sampling across all benchmarks, notably boosting Temperature Sampling on BLINK by over 2% by decoupling reasoning exploration from answer generation. Furthermore, it improves upon the naive greedy decoding baseline (*e.g.*, reaching 60.92% on MMMU vs. 54.11% baseline), demonstrating that combining probabilistic/beam search reasoning with deterministic answering is superior to a fully deterministic approach. Additional results on Qwen3-VL-30B-A3B-Thinking (Table 11; Appendix C.3) show that GDRM consistently outperforms all stochastic baselines, demonstrating its scalability to larger models. In summary, GDRM offers a greater stability, effectively boosting the reasoning model performance by anchoring the answer tokens to the top-1 prediction.

6 Conclusion

We revisit greedy decoding for MLLMs on VQA from a calibration perspective. We demonstrate that the head-heavy answer distribution and epistemic uncertainty inherent to VQA make stochastic sampling suboptimal, as it often expands the candidate set toward low-probability distractors rather than semantically correct alternatives. Our theoretical framework formalizes the relationship between model calibration and predictive accuracy, identifying sufficient conditions under which greedy decoding is optimal. Our extensive experiments across diverse VQA benchmarks demonstrate the superior performance of greedy decoding to stochastic sampling. Furthermore, we propose GDRM, a decoding strategy that anchors final answers to top-1 predictions without compromising the model's underlying reasoning capabilities.

Together, our findings caution against naively inheriting LLMs decoding heuristics in MLLMs without task-specific justification and demonstrate that greedy decoding can be an efficient yet strong default for VQA.

Limitations

Benchmark Scope. Although our evaluation covers diverse benchmarks, the effectiveness of these decoding strategies in highly specialized domains (*e.g.*, medical VQA) remains unexplored.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. 2025. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14077–14094.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 889–898. Association for Computational Linguistics.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. 2024. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. *arXiv preprint arXiv:2405.15683*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

John Hewitt, Christopher D. Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3388–3403. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*. ArXiv:1904.09751.

Zaid Khan and Yun Fu. 2024. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10854–10863.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Xueyan Li, Guinan Su, Mrinmaya Sachan, and Jonas Geiping. 2025a. Sample smart, not hard: Correctness-first decoding for better reasoning in llms. *arXiv preprint arXiv:2510.05987*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. 2025b. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Peijie Liu, Fengli Xu, and Yong Li. 2025. Token signature: Predicting chain-of-thought gains with token decoding feature in large language models. In *International Conference on Machine Learning (ICML)*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. 2024. Diversity of thought improves reasoning abilities of llms. *Preprint*, arXiv:2310.07088.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Cheng Wang. 2023. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint*, arXiv:2308.01222. Revised version v3, submitted 2 August 2023, last revised 10 May 2024.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision (ECCV)*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Theory Details

A.1 Calibration metrics

Expected Calibration Error (ECE) ECE partitions the predicted probability space into N bins, $b_n := \{i \mid q(\hat{a}_i|I_i, x_i) \in U_n\}$, where U_n denotes a predefined probability interval for the n-th bin. It computes the weighted average of the discrepancy between empirical accuracy and average confidence within each bin:

$$\mathrm{ECE} := \sum_{n=1}^{N} \frac{|b_n|}{M} \left|\mathrm{acc}(b_n) - \mathrm{conf}(b_n)\right|, \qquad (6)$$

where M is the total number of samples, and

$$\mathrm{acc}(b_n) := \frac{1}{|b_n|} \sum_{i \in b_n} \mathbf{1}(\hat{a}_i = a_i^*), \qquad (7)$$

$$\mathrm{conf}(b_n) := \frac{1}{|b_n|} \sum_{i \in b_n} q(\hat{a}_i|I_i, x_i). \qquad (8)$$

Here, $\hat{a}_i$ is the model's prediction and a_i^* is the ground truth label. Conceptually, $\mathrm{acc}(b_n)$ approximates the expectation $\mathbb{E}[p(a|I, x) \mid q(a|I, x) \in U_n]$, while $\mathrm{conf}(b_n)$ approximates $\mathbb{E}[q(a|I, x) \mid q(a|I, x) \in U_n]$. A more thorough mathematical investigation of ECE is provided in Appendix A.1.

We provide more detailed mathmatical explanation for ECE and BS in the following propositions.

Proposition A.1. *Assuming that $q(a|I,x)$ takes a unique value for each a, I, x, and that we have access to infinite data for every (I,x). Consider constructing one bin for each unique $q(a|I,x)$ value. ECE in our problem setup approximates $\mathbb{E}_{I,x,a\sim p(\cdot|I,x)}[|q(a|I,x) - p(a|I,x)|]$*

Proof. Since $q(a|I,x)$ is unique for each (I,x,a) and we assign one bin per $q(a|I,x)$, each bin corresponds to a specific (I,x,a) combination. Then we have:

$$\begin{aligned}\text{ECE} &= \sum_{n=1}^{N} \frac{|b_n|}{N} |\text{acc}(b_n) - \text{conf}(b_n)| \\ &\approx \sum_{I,x,a} p(I,x,a) |p(a|I,x) - q(a|I,x)| \\ &= \mathbb{E}_{I,x,a\sim p(\cdot|I,x)}[|q(a|I,x) - p(a|I,x)|].\end{aligned}$$

□

Brier Score (BS) The Brier Score measures the mean squared error between the predicted probability and the actual label. While the classical Brier Score assumes deterministic labels, we can derive an adapted version that accounts for the discrepancy between distributions using a binning approach:

$$\text{BS} := \sum_{n=1}^{N} \frac{|b_n|}{M} \frac{|\text{acc}(b_n) - \text{conf}(b_n)|^2}{\text{acc}(b_n)}. \quad (9)$$

This metric approximates the expected squared difference between the predicted and true probabilities across the answer space, averaged over the dataset. A more thorough mathematical investigation of BS is also provided in Appendix A.1.

To analyze the calibration of a specific parameterized strategy α, we extend the above metrics into a continuous expectation framework.

Proposition A.2. *Assuming that $q(a|I,x)$ takes a unique value for each a, I, x, and that we have access to infinite data for every (I,x). Consider constructing one bin for each unique $q(a|I,x)$ value. Then BS in our problem setup approximates $\mathbb{E}_{I,x}[\sum_a |q(a|I,x) - p(a|I,x)|^2]$.*

Proof. The proof follows similar steps to the proof of Proposition A.1:

$$\begin{aligned}&\text{BS} \frac{1}{N} \sum_{n=1}^{N} |b_n| \frac{|\text{acc}(b_n) - \text{conf}(b_n)|^2}{\text{acc}(b_n)} \\ &\approx \sum_{I,x,a} p(I,x,a) \frac{|q(a|I,x_i) - p(a|I,x)|^2}{p(a|I,x)} \\ &\approx \mathbb{E}_{I,x,a\sim p(\cdot|I,x)} \left[\frac{|q(a|I,x_i) - p(a|I,x)|^2}{p(a|I,x)} \right] \\ &= \mathbb{E}_{I,x}[\sum_a |q(a|I,x) - p(a|I,x)|^2]\end{aligned}$$

□

These results motivate our definition of ECE^α and BS^α with p and q.

A.2 Strategy-specific Brier Score

The strategy-specific Brier Score is defined as the expected normalized squared difference:

$$\text{BS}^\alpha := \mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)} \left[\frac{|q(a|I,x) - p(a|I,x)|^2}{q(a|I,x)} \right], \quad (10)$$

$$\text{BS}^1 := \mathbb{E}_{I,x} \left[\frac{|q(a^1|I,x) - p(a^1|I,x)|^2}{q(a^1|I,x)} \right], \quad (11)$$

where BS^1 is the calibration error of the greedy strategy (always taking the token with the highest confidence).

In the definition, we substitute the denominator $p(a|I,x)$ in the original definition of BS (Appendix A.1) with $q(a|I,x)$. Since the denominator serves only as a positive weighting over answers, both of them yield a valid calibration error, differing only in how deviations are weighted across answers. We choose $q(a|I,x)$ for two reasons. First, it is known and strictly positive for top-k answers, making the metric stable and directly estimable, whereas $p(a|I,x)$ is unknown and may be arbitrarily small or zero. Second, it enables algebraic decomposition in the proof of Theorem 3.1 (see Appendix A.3 for details): the bound leading to Eq. 3.1 naturally yields a $\mathbb{E}[|q-p|^2/q]$ term (*i.e.*, our BS^α), whereas utilizing $p(a|I,x)$ would break this tractable bound and obscure the resulting sufficient condition.

A.3 Proof of Theorem 3.1

Proof. We first introduce the correctness metric (C^α) for the strategy with the parameter α:

$$C^\alpha := \mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}[p(a|I,x)] \quad (12)$$

We start by bounding C^α using ECE^α and q:

$$\begin{aligned}
C^\alpha &= \mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}[p(a|I,x)] \\
=&\mathbb{E}_{I,x,,a\sim q^\alpha(\cdot|I,x)}[q(a|I,x) \\
&+ (p(a|I,x) - q(a|I,x))] \\
\leq&\mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}[q(a|I,x) \\
&+ |p(a|I,x) - q(a|I,x)|] \\
\leq&\mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}[q(a|I,x)] + \text{ECE}^\alpha
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
C^\alpha =&\mathbb{E}_{I,x,,a\sim q^\alpha(\cdot|I,x)}[q(a|I,x) \\
&+ (p(a|I,x) - q(a|I,x))] \\
\geq&\mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}[q(a|I,x) \\
&- |p(a|I,x) - q(a|I,x)|] \\
=&\mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}[q(a|I,x)] - \text{ECE}^\alpha
\end{aligned}$$

Similarly, we can have $C^1 \leq \mathbb{E}_{I,x}[q(a^1|I,x)] + \text{ECE}^1$ and $C^1 \geq \mathbb{E}_{I,x}[q(a^1|I,x)] - \text{ECE}^1$. Applying this to inequality (3.1), we obtain:

$$\begin{aligned}
C^\alpha &\leq \mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}[q(a|I,x)] + \text{ECE}^\alpha \\
\leq&\mathbb{E}_{I,x}[q(a^1|I,x)] - \text{ECE}^1 \leq C^1
\end{aligned}$$

On the other hand, we can also bound C^α as follows:

$$\begin{aligned}
C^\alpha &= \mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}\left[\frac{p(a|I,x)q(a|I,x)}{q(a|I,x)}\right] \\
=&\mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}\left[\frac{p(a|I,x)^2 + q(a|I,x)^2}{2q(a|I,x)}\right. \\
&\left.- \frac{|p(a|I,x) - q(a|I,x)|^2}{2q(a|I,x)}\right] \\
\leq&\mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}\left[\frac{1 + q(a|I,x)^2}{2q(a|I,x)}\right] - \frac{\text{BS}^\alpha}{2}
\end{aligned}$$

Combining this bound with the previous bound on C^1 using ECE^1, and assuming inequality (3.1) holds, we get:

$$\begin{aligned}
C^\alpha &\leq \mathbb{E}_{I,x,a\sim q^\alpha(\cdot|I,x)}\left[\frac{1 + q(a^\alpha|I,x)^2}{2q(a^\alpha|I,x)}\right] - \frac{\text{BS}^\alpha}{2} \\
&\leq \mathbb{E}_{I,x}[q(a^1|I,x)] - \text{ECE}^1 \leq C^1
\end{aligned}$$

The inequality $C^1 \geq C^\alpha$, for all α that gives non-greedy strategy suggests that greedy strategy is optimal. □

B Implementation Details

B.1 Prompt and Formatting

For each dataset, we provide a global system prompt. All models use a standardized instruction template with explicit role and output style cues. For multiple-choice questions, we append the choices and instruct the model to answer with the option letter only (*e.g.,* `A/B/C/D`). For free-form questions, we instruct the model to answer with a short phrase. We disable chain-of-thought style prompting and request only to output the answers without reasoning.

B.2 Inference setting

Inference is run on a single A6000 with 48GB memory for all models. We cap decoding at 64 tokens for free-form answers and 8 tokens for multiple-choice questions.

B.3 Hyperparameters

We search: $T \in \{0.3, 0.7, 1.0, 1.5, 2.0\}$, $k \in \{20, 50\}$, $p \in \{0.9, 0.95\}$, $p_{base} \in \{0.05, 0.1\}$, $\varepsilon \in \{0.0009, 0.0006\}$, and $\eta \in \{0.0006, 0.0002\}$. For the Qwen3-VL Thinking series models, we use the officially recommended decoding hyperparameters detailed in Table 8.

Table 8: Hyperparameters used for Qwen3-VL Thinking models.

Parameter	Value
temperature	1.0
top-p	0.95
top-k	20

B.4 System Prompts

This section provides the exact system prompts utilized for each VQA benchmark during our evaluation. These prompts were designed to constrain the model's output to the specific format required by each dataset's evaluation script.

BLINK For the BLINK benchmark, the following system prompt was used to ensure the model responded with only the correct option letter(s).

```
You are being evaluated on BLINK.
Answer with ONLY the option letter(s)
for the correct choice (A/B/C/...).
Do not output any words, punctuation, or
explanation.
```

ChartQA For the ChartQA benchmark, we employed a more detailed system prompt to guide the model in generating answers in various formats (*e.g.*, numeric, "Yes/No", multiple-choice).

```
You are a vision-language model specialized in chart question answering. Answer with ONLY the final result (no explanation).
Rules:
• For numeric answers: use digits and include the unit exactly as shown in the chart (e.g., %, °C, $, K).
• For yes/no: reply "Yes" or "No" only.
• For multiple-choice: output the option letter(s) only (e.g., "B" or "A,C") with no spaces.
• If the question cannot be answered from the chart, reply exactly: "unanswerable".
```

MMMU No system prompt was used for the MMMU benchmark. The evaluations were without any system prompt being passed to the models.

B.5 Metrics

Hallucination Benchmark. The Answer Score is a GPT-4o-assigned quality rating on a 1–5 Likert scale evaluating correctness, completeness, and relevance of the generated answer. The Hallucination Rate is the fraction of responses flagged by GPT-4o as containing content not supported by the image. Both metrics follow the evaluation protocol established in prior work (Sun et al., 2024).

The detailed evaluation prompt is as follows:

```
Please act as an impartial and objective judge and evaluate the quality of the response provided by a Large Multimodal Model (LMM) to the user question. Your evaluation should be mainly based on whether the response is informative, and whether the response contains any hallucination. Hallucination, in this context, refers to a situation where the LMM generates a response that includes information not present or implied in the image or previous conversation. A hallucination could be a false claim about an object, action, emotion, or any other detail that is not grounded in the image.
For clarity, consider these examples:
...
With these examples in mind, please help me evaluate whether the response by the LMM is informative, and whether hallucination exists in it, based on the comparison between the LMM's response and the factual information provided in the image contents, question, and the standard human-generated answer below.
Please note that the standard human-generated answer may only contain factual information but may not give a detailed analysis. Also, the standard human-generated answer may not be completely comprehensive in describing all the objects and their attributes, so please be a bit more cautious during evalutation. LMM's detailed analysis or reasoning should be encouraged.
To evaluate the LMM responses, first, begin your evaluation by providing a short explanation. Second, after providing your explanation, you must rate the response by choosing from the following options:
- Rating: 6, very informative with good analysis or reasoning, no hallucination
- Rating: 5, very informative, no hallucination - Rating: 4, somewhat informative, no hallucination - Rating: 3, not informative, no hallucination
- Rating: 2, very informative, with hallucination - Rating: 1, somewhat informative, with hallucination - Rating: 0, not informative, with hallucination
```

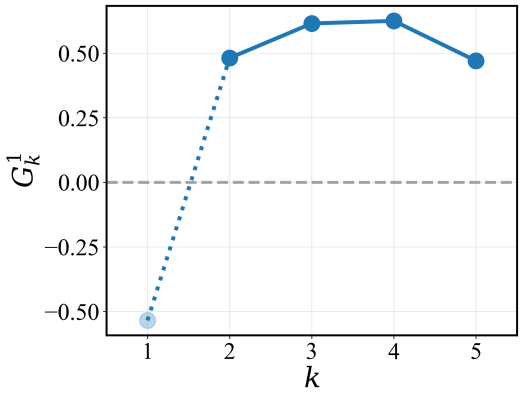


Figure 5: Empirical evidence (G_1^k vs. k) for Theorem 3.1 on MMLU using Qwen2.5-VL (3B)

Table 9: Accuracy (in %) of different decoding/sampling strategies on MMLU benchmark using Qwen2.5-VL (3B)..

Method	MMLU	
	τ=0.7	τ=2.0
Temp' Only	72.86	71.93
Top-k	71.23	71.23
Top-p (nucleus)	71.23	69.47
Min-p	71.58	70.18
ε-Sampling	71.23	68.77
η-Sampling	70.36	71.79
Greedy	**73.21**	

Open-ended Image Captioning. The Average Score is a GPT-4o-as-a-judge measurement obtained via pairwise battles between the evaluated model and three reference captioners (GPT-4o, CogVLM2, and MiniCPM-V) over 600 images, computed as the mean of the three per-baseline net win rates (in percent, range $[-100, 100]$). The Average Length is the mean length of the generated captions, reported alongside the Average Score to contextualize verbosity. Both metrics follow the CapArena-Auto evaluation protocol established in prior work (Cheng et al., 2025).

C Additional Results

C.1 Additional results on the MMLU dataset

Empirical evidence on the MMLU dataset. Fig 5 shows that the G_1^k remains positive for all $k > 1$ on the MMLU dataset using Qwen2.5-VL (3B). Theorem 3.1 proves that in this case, greedy decoding is optimal.

Results with different temperatures. Table 9 provides additional results using Qwen2.5-VL (3B) across different temperature settings. The results demonstrate that although lowering the temperature slightly mitigates the negative effects of stochasticity, greedy decoding consistently outperforms stochastic sampling methods.

C.2 Detailed results on the POPE Benchmark

Table 10 shows average Accuracy and F1 scores on the POPE benchmark (Rohrbach et al., 2018) using greedy decoding and stochastic sampling (nucleus) strategies. Greedy (vanilla) decoding consistently outperforms nucleus sampling (vanilla) across all four evaluated models on both metrics (Li et al., 2025b). The performance gap is particularly pronounced in models like MiniGPT-4 (Zhu et al., 2023), where nucleus sampling suffers a severe degradation in both Accuracy (-16.2%) and F1 score (-14.78%). These results are consistent with our findings on the MM-HallBench dataset in the main text, providing additional empirical evidence that stochastic sampling expands the candidate set into the low-probability tail, frequently associated with ambiguous visual evidence and therefore at a higher risk of hallucination.

Table 10: Comparison of average Accuracy and F1 scores, averaged across random, popular, and adversarial object splits, between Greedy and Nucleus sampling across different models on POPE benchmark. Results are adopted from Li et al. (2025b). Best results are in **bold**.

Model	Avg. Accuracy		Avg. F1	
	Greedy	Nucleus	Greedy	Nucleus
LLaVA-1.5 (Liu et al., 2023)	**84.79**	81.26	**85.61**	82.40
MiniGPT-4 (Zhu et al., 2023)	**76.76**	60.56	**76.82**	62.04
Shikra (Chen et al., 2023)	**81.32**	78.94	**82.01**	80.18
InstructBLIP (Dai et al., 2023)	**84.36**	78.83	**84.64**	79.74

C.3 Greedy Decoding for Reasoning Model

To evaluate the scalability of Greedy Decoding for Reasoning Model (GDRM), we conduct additional experiments on ChartQA using Qwen3-VL-30B-A3B-Thinking, a larger Mixture-of-Experts reasoning model comprising 30B total parameters (3B active). As shown in Table 11, GDRM consistently outperforms all stochastic sampling baselines, suggesting that its effectiveness scales to larger-scale models.

Table 11: Comparison of Baseline and GDRM accuracy across different sampling strategies for Qwen3-VL-30B-A3B-Thinking.

	ChartQA	
Method	Baseline	GDRM
Top-k	84.68	85.40 (+0.72)
Top-p (nucleus)	84.15	84.70 (+0.55)
ε-Sampling	84.22	84.86 (+0.64)

C.4 Detailed results with different random seeds

We list the evaluation results with different random seeds using Qwen2.5-VL (3B) from Table 12 to Table 23.

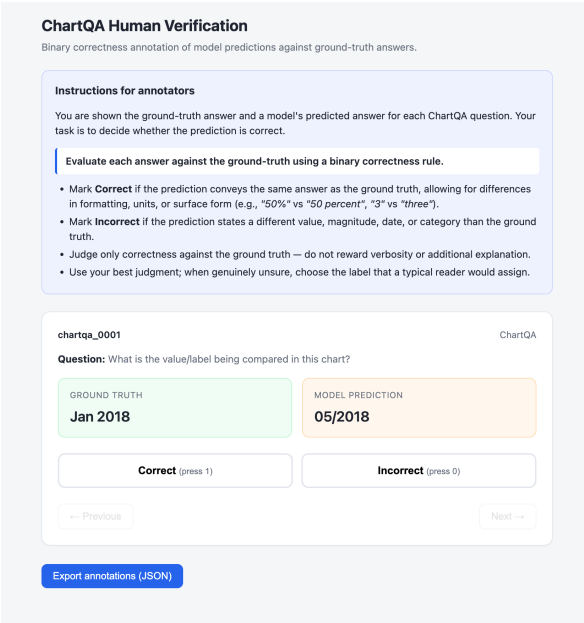


Figure 6: The annotation interface used for the human verification study on ChartQA. Annotators are shown the ground-truth answer and the model prediction, follow the binary correctness rule stated in the instructions, and record a Correct/Incorrect judgment.

D Human Verification Interface

To assess whether our automatic string-matching evaluation reflects true answer correctness, we conducted a human verification study on ChartQA. Two researchers independently judged each model prediction against the ground-truth answer using a binary correctness rule (correct vs. incorrect), through the lightweight annotation interface shown in Figure 6. For each instance, the interface presents the ground-truth answer alongside the model prediction together with the full annotation instructions, and records a binary correctness judgment. The verbatim instruction shown to annotators is: *"Evaluate each answer against the ground-truth using a binary correctness rule."* Annotators were instructed to mark a prediction *Correct* if it conveys the same answer as the ground truth up to differences in formatting, units, or surface form, and *Incorrect* if it states a different value, magnitude, date, or category. Agreement between the human annotations and the string-match labels is high (Cohen's $\kappa = 0.95$), indicating that our results are not driven by string-matching artifacts.

D.1 Annotator Recruitment and Compensation

The human verification was carried out by two researchers from the project team, who participated voluntarily. Because the task was a brief internal validation conducted by members of the research team rather than recruited external workers, no monetary compensation was provided or required.

D.2 Annotator Consent

Prior to annotation, both researchers were shown the following instruction, which also served as the consent statement: *"You are asked to voluntarily judge model-generated answers against the public ChartQA ground-truth labels using a binary correctness rule. The task involves only model outputs and publicly available benchmark data, with no personal or sensitive information. Your anonymized binary judgments will be aggregated solely to measure the inter-annotator agreement reported in this paper. By proceeding, you consent to this use."* Both annotators reviewed this statement and consented before participating.

E Use of AI Assistants in Research

In our study, generative AI assistants are used sparingly and in accordance with the guidelines on ACL's Policy on AI Writing Assistance. We utilize ChatGPT for basic paraphrasing and grammar checks. These tools are applied minimally to ensure the authenticity of our work and to adhere strictly to the regulatory standards set by ACL. Our use of these AI tools is focused, responsible, and aimed at supplementing rather than replacing human input and expertise in our research.

Table 12: Accuracy (in %) of different decoding/sampling strategies on MMMU benchmark (seed=42) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	45.33	44.22	42.11	37.33	33.89
Top-p (nucleus)	45.44	42.44	40.33	35.44	28.67
Min-p	45.44	42.56	39.89	35.22	27.56
ε-Sampling	45.44	43.00	41.44	38.56	34.56
η-Sampling	45.44	44.22	40.33	38.22	29.22
Greedy			45.44		

Table 13: Accuracy (in %) of different decoding/sampling strategies on MMMU benchmark (seed=123) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	45.33	41.11	41.22	35.22	32.78
Top-p (nucleus)	45.00	43.11	42.89	37.22	30.56
Min-p	45.33	41.00	39.33	36.56	29.78
ε-Sampling	45.33	40.78	38.56	34.89	33.33
η-Sampling	45.44	40.89	39.89	34.67	29.67
Greedy			45.44		

Table 14: Accuracy (in %) of different decoding/sampling strategies on MMMU benchmark (seed=456) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	43.56	42.56	40.00	35.78	31.89
Top-p (nucleus)	45.44	42.22	41.33	35.78	27.11
Min-p	43.56	44.78	41.56	36.11	29.11
ε-Sampling	43.56	44.89	41.33	35.22	35.44
η-Sampling	43.56	44.67	41.33	37.67	27.56
Greedy			45.44		

Table 15: Accuracy (in %) of different decoding/sampling strategies on MMMU benchmark (seed=789) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	45.44	42.22	41.00	36.22	35.11
Top-p (nucleus)	45.33	43.00	40.67	35.67	26.44
Min-p	45.44	42.22	40.00	34.89	28.11
ε-Sampling	45.44	41.11	40.67	39.11	36.78
η-Sampling	45.33	42.22	40.22	35.67	25.89
Greedy			45.44		

Table 16: Accuracy (in %) of different decoding/sampling strategies on ChartQA benchmark (seed=42) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	82.29	79.53	76.09	65.73	51.98
Top-p (nucleus)	82.34	81.72	79.11	63.75	48.39
Min-p	82.29	79.53	76.72	52.08	44.80
ε-Sampling	81.93	79.48	76.88	67.14	54.84
η-Sampling	82.29	79.53	76.82	65.94	48.80
Greedy			83.12		

Table 17: Accuracy (in %) of different decoding/sampling strategies on ChartQA benchmark (seed=123) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	82.50	79.38	75.78	67.14	51.41
Top-p (nucleus)	82.40	80.68	78.33	62.60	46.88
Min-p	82.50	79.43	76.20	53.02	42.39
ε-Sampling	82.50	79.84	75.99	66.20	51.35
η-Sampling	82.50	79.58	75.89	63.91	48.39
Greedy			83.12		

Table 18: Accuracy (in %) of different decoding/sampling strategies on ChartQA benchmark (seed=456) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	82.60	79.38	76.41	66.88	52.40
Top-p (nucleus)	82.08	81.04	79.43	62.76	46.09
Min-p	82.60	79.38	75.57	51.98	47.97
ε-Sampling	82.60	80.16	76.20	67.81	52.19
η-Sampling	82.60	79.38	76.77	64.95	47.97
Greedy			83.12		

Table 19: Accuracy (in %) of different decoding/sampling strategies on ChartQA benchmark (seed=789) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	82.24	80.16	76.15	65.42	51.77
Top-p (nucleus)	82.97	81.09	78.85	64.17	49.48
Min-p	82.24	80.16	75.89	52.19	47.50
ε-Sampling	82.24	80.16	77.03	66.46	51.46
η-Sampling	82.24	80.16	76.35	62.45	47.50
Greedy			83.12		

Table 20: Accuracy (in %) of different decoding/sampling strategies on BLINK benchmark (seed=42) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	34.68	32.75	32.26	29.33	27.13
Top-p (nucleus)	34.24	29.71	31.44	28.60	26.60
Min-p	35.32	33.07	30.76	29.97	28.63
ε-Sampling	34.28	32.27	30.62	30.74	26.59
η-Sampling	33.13	32.07	31.16	27.63	26.10
Greedy			35.32		

Table 21: Accuracy (in %) of different decoding/sampling strategies on BLINK benchmark (seed=123) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	34.68	33.07	30.40	29.00	25.71
Top-p (nucleus)	34.61	32.37	30.82	29.72	27.81
Min-p	35.32	33.40	30.41	30.76	31.37
ε-Sampling	34.19	31.55	30.66	29.49	25.99
η-Sampling	34.95	30.10	30.70	27.27	24.25
Greedy			35.32		

Table 22: Accuracy (in %) of different decoding/sampling strategies on BLINK benchmark (seed=456) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	34.38	31.01	30.31	29.72	26.17
Top-p (nucleus)	34.29	32.42	32.58	28.34	26.76
Min-p	35.32	32.94	32.08	29.55	30.72
ε-Sampling	34.04	30.67	31.88	28.43	26.62
η-Sampling	34.98	30.58	31.34	28.74	26.37
Greedy			35.32		

Table 23: Accuracy (in %) of different decoding/sampling strategies on BLINK benchmark (seed=789) using Qwen2.5-VL (3B).

Method	τ=0.3	τ=0.7	τ=1.0	τ=1.5	τ=2.0
Top-k	35.32	33.85	30.31	28.99	26.39
Top-p (nucleus)	34.98	31.94	31.85	30.96	27.39
Min-p	35.32	33.78	30.67	31.54	29.00
ε-Sampling	34.15	29.87	31.60	28.22	25.79
η-Sampling	34.56	31.61	30.96	27.46	25.31
Greedy			35.32		

F Dataset Licensing and Intended Use

F.1 Dataset License

BLINK, ChartQA, MMMU, MMHal-Bench, MMLU, and CapArena are intended for research usage.

BLINK It has an Apache license 2.0, allowing research usage.

ChartQA It has a GNU General Public License v3.0, allowing research usage.

MMMU It has an Apache license 2.0, allowing research usage.

MMHal-Bench It has an Apache license 2.0, allowing research usage.

MMLU It has an MIT license, allowing research usage.

CapArena It is publicly released for academic research use (Cheng et al., 2025).

F.2 Intended Use

Our use of the existing artifacts strictly comply with their intended purpose as benchmarks for research usage. We do not introduce any new dataset or artifact in this work.

F.3 Privacy and Offensive Content

We use only established, publicly available benchmarks (MMMU, ChartQA, BLINK, MMHallBench, MMLU, and CapArena) and neither collect, scrape, nor release any new data. These benchmarks are built from charts, diagrams, documents, and natural images paired with factual question–answer items rather than personal records, and their content was curated and vetted by the original dataset creators. To confirm this for our setting, we manually inspected samples during the human verification study (Appendix D) and the error analysis; we did not encounter personally identifying information or offensive content. As we introduce no new data and use each benchmark unmodified under its research license, any anonymization or content filtering is inherited from the original releases, and no additional anonymization was required on our part.

G Potential Risks

This work is foundational research on inference-time decoding strategies for VQA; it introduces no new model, dataset, or deployable system, and evaluates only publicly released MLLMs on public benchmarks. We nonetheless note several potential risks. First, greedy (argmax) decoding deterministically surfaces the model's most probable answer and can therefore reflect or amplify biases present in the underlying model and its training data; calibration-aware selection mitigates but does not eliminate this, and we recommend pairing it with calibration checks or abstention in sensitive applications. Second, reducing output stochasticity lowers answer diversity, which is desirable

for closed-ended VQA but may be undesirable for open-ended or creative generation; our recommendations are scoped to closed-ended VQA and should not be transferred uncritically to such settings. Third, although greedy decoding reduces hallucination-like errors in our experiments, it does not remove them, and a single deterministic answer may convey unwarranted confidence; downstream users should not treat model outputs as authoritative without verification. Finally, all benchmarks used are English-only and research-licensed, so our conclusions may not transfer to other languages or domains. We are not aware of any direct path to malicious use beyond those already inherent to the public MLLMs we evaluate.