READOUT FIDELITY BOTTLENECKS SELF-VERIFICATION IN UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

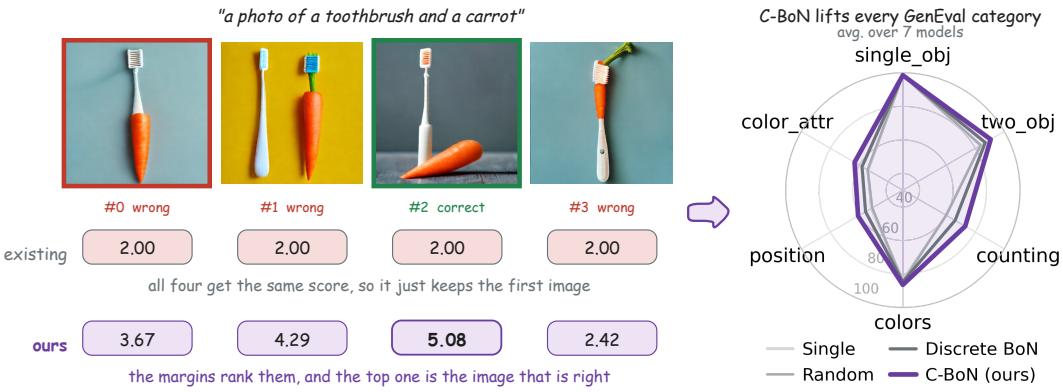

Figure 1: Left: Discrete self-verification collapses distinct candidates to the same score, causing deterministic tie-breaking to return an incorrect candidate. C-BoN instead reads the pre-decoding Yes-vs.-No logit margin, which preserves candidate distinctions and selects the correct image. Right: C-BoN improves GenEval performance across categories.

ABSTRACT

Unified multimodal models combine image generation and visual understanding, allowing them to evaluate and select among their own generations without an external verifier. Yet self-verified Best-of-N recovers only a small fraction of the quality available in candidate pools. We identify readout fidelity as an independent bottleneck: useful candidate preferences can already exist in verifier evidence but are compressed by discrete Yes/No decoding. Across seven models, discrete readout produces extensive score collisions and can even underperform random selection despite substantial oracle headroom. We model discrete verification as finite observations of an underlying Yes/No preference, with a formulation agnostic to how the verifier evidence is produced. In the direct-answer setting, repeated stochastic observations converge to the pre-decoding logit margin, while full stochastic decoding closely tracks the predicted finite-observation behavior. We use Continuous Best-of-N (C-BoN) as a controlled intervention that directly exposes this limiting preference while keeping the remaining selection pipeline fixed. Across seven models and four benchmarks, C-BoN improves all 28 settings without additional models, training, generation, or verifier calls, and enables stronger Best-of-N scaling. Larger continuous margins also yield more reliable ordering within discrete ties, while recovered ordering tracks downstream gains. These results establish readout fidelity as an information bottleneck in self-verification and show that verifier evidence can contain richer preferences than discrete verdicts reveal.

1 INTRODUCTION

Unified multimodal models integrate image generation and visual understanding within a single model Xie et al. (2025); Chen et al. (2025); Deng et al. (2025); Xie et al. (2026). This enables

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.

a distinctive capability: a model can generate multiple candidate images, evaluate them with its own understanding, and select the best one without an external reward model or vision-language evaluator. Recent methods exploit this capability for test-time scaling and Best-of-N selection Tian et al. (2026); Xin et al. (2026); Wang et al. (2025), as well as self-rewarding and post-training Jin et al. (2025); Pan et al. (2026); Thawkar et al. (2026).

A common strategy decomposes the prompt into verification questions, evaluates each candidate, and aggregates the judgments into a score Tian et al. (2026); Kim et al. (2026a); Jin et al. (2025). Prior work has mainly enriched this evidence through better question decomposition, reasoning, evaluation signals, or search Tian et al. (2026); Kim et al. (2026a); Jin et al. (2025); Xin et al. (2026). This leaves a separate factor largely unisolated: whether the interface used to read verifier evidence preserves useful candidate preferences it already contains.

We study this factor across seven unified multimodal models spanning autoregressive VQ, masked discrete diffusion, and continuous diffusion or flow-based generation. On GenEval Ghosh et al. (2023), standard self-verified Best-of-4 ranges from 0.36 points below to 6.42 points above random selection, while an oracle over the same candidates lies 10 to 19 points above a single sample. Thus, substantial quality exists in the candidate pool but is often lost during selection. We find that this failure does not necessarily imply weak verifier evidence. Standard self-verification reduces each answer to a discrete Yes/No verdict, producing a low resolution score that collapses candidate distinctions despite differing evidence. All four candidates receive the same score on 19% to 83% of prompts across models and benchmarks. Resolving these ties recovers much, but not all, of the lost quality, while continuous preference remains informative within discrete ties. Together, these results identify readout fidelity as an independent bottleneck between verifier evidence and final selection.

We formalize this bottleneck through an observation model that allows arbitrary latent verification history before a terminal Yes, No, or neither outcome. Repeated observations estimate the marginal Yes/No preference, while an exact finite observation analysis characterizes uncertainty and candidate ordering. For our direct-answer experiments, this limiting preference reduces exactly to the pre-decoding Yes/No logit margin. Full stochastic decoding closely tracks the predicted finite-m recovery behavior, connecting selection failure to information discarded at the answer readout.

To test this explanation, we use Continuous Best-of-N (C-BoN) as a controlled intervention that reads the Yes/No logit margin while keeping the candidates, questions, aggregation, and selection rule fixed. It requires no additional model, training, generation, or verifier call. Across seven models, preserving this preference improves Best-of-4 by $+1.45$ to $+7.41$ points on GenEval and yields consistent gains on T2I-CompBench Huang et al. (2023), GenAI-Bench Jiang et al. (2024), and DPG-Bench Hu et al. (2024). It also enables stronger Best-of-N scaling. Within discrete ties, larger continuous margins yield more reliable ordering across all four benchmarks, while improvements in recovered ordering track downstream gains.

Our contributions are threefold:

- We identify readout fidelity as an independent bottleneck in self-verification. Holding candidate pools, verification evidence, aggregation, and selection fixed, we show that discrete readout can discard useful ordering already present in the verifier.
- We characterize this bottleneck through a finite observation view of discrete verification. Repeated observations converge to a marginal Yes/No preference, which reduces to the pre-decoding logit margin for direct-answer verification and predicts stochastic decoding.
- We validate the finding through controlled readout interventions across seven models and four benchmarks. Preserving continuous preference improves all 28 settings, enables stronger Best-of-N scaling, and reveals ranking information hidden by discrete verdicts.

2 Related Work

Unified multimodal models. Unified multimodal models combine visual understanding and generation within a single architecture. Early systems such as DreamLLM, Chameleon, VILA-U, Emu3, Show-o, and Janus established this paradigm Dong et al. (2024); Team (2024); Wu et al. (2025); Wang et al. (2024); Xie et al. (2025); Chen et al. (2025). More recent models include BAGEL Deng et al. (2025), MMaDA Yang et al. (2026), Show-o2 Xie et al. (2026), and LaViDa-O Li et al.

(2026). Their shared capabilities enable intrinsic evaluation of generated outputs. We study whether self-selection is limited not only by verifier evidence, but also by how it is read.

Test-time scaling and self-verification. Test-time scaling improves image generation through sampling, verification, or search Ma et al. (2025); Zhang et al. (2026); Sundaresha et al. (2025). Candidate selection often relies on external models such as ImageReward, PickScore, HPSv2, and VQAScore Xu et al. (2023); Kirstain et al. (2023); Wu et al. (2023); Lin et al. (2024). Unified models instead enable intrinsic verification. UniGen uses Chain-of-Thought Verification that aggregates decomposed judgments into a discrete score Tian et al. (2026), while dMLLM-TTS uses Yes confidence Xin et al. (2026). Related methods use self-evaluation, refinement, or search Wang et al. (2025); Chen et al. (2026); Li et al. (2026); Kim et al. (2026b). These works enrich verifier evidence, while we isolate how much information survives its conversion into a selection score. We hold the remaining pipeline fixed to isolate this readout factor.

Multimodal evaluation and self-rewarding. VQA-based metrics provide fine-grained text-image alignment signals. TIFA uses decomposed visual questions Hu et al. (2023), while VQAScore uses the probability of answering Yes Lin et al. (2024). Unified models also use understanding as a training signal, as in SRUM Jin et al. (2025) and Ask, Solve, Generate Thawkar et al. (2026). Our contribution is not a new continuous Yes score. We identify readout fidelity as a distinct bottleneck in self-verified selection. Controlled comparisons isolate information discarded at this interface, while a finite-observation model connects discrete judgments to continuous preference.

3 Readout Fidelity in Self-Verification

3.1 Preliminaries

A unified model provides a generator G and an understanding path U that share weights. Given a prompt p, Best-of-N self-verification draws $x_n = G(p; \xi_n)$ for $n \in [N]$, where ξ_n is sampling noise. Each candidate is scored by U, and the final output is

$$\hat{x} = x_{n^\star}, \qquad n^\star = \arg\max_{n \in [N]} s(x_n, p), \tag{1}$$

with ties resolved by the smallest index. The score aggregates K verification questions, each paired with a negated counterpart $\bar{q}_k$ to reduce Yes bias:

$$s(x, p) = \frac{1}{K} \sum_{k=1}^{K} \left[a\big(x, q_k(p)\big) - a\big(x, \bar{q}_k(p)\big) \right], \tag{2}$$

where $a(x, q)$ is the answer signal. We use $K = 3$ throughout (Appendix A.4). To isolate readout fidelity, we keep the candidates, questions, verifier computation, aggregation, and selection fixed, varying only a.

Following Tian et al. (2026), the baseline reads a from the decoded response y:

$$a_{\text{disc}}(x, q) = \begin{cases} +1, & \text{if } y \text{ begins with yes,} \\ -1, & \text{if } y \text{ begins with no,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

We call this the discrete verdict. The baseline uses greedy decoding, so repeating it reveals no additional information.

3.2 A Readout Bottleneck in Self-Verification

Figure 2 diagnoses where self-verified selection loses candidate quality on GenEval. Across seven models, the oracle over the same four candidates lies 10.3 to 19.0 points above a single sample, while Discrete Best-of-4 recovers only 0% to 40% of this headroom, and none on Janus-Pro-7B. Thus, substantial quality is available but lost during selection.

Discrete readout has limited resolution. Under equation 3, each term $a(x, q_k) - a(x, \bar{q}_k)$ lies in $\{-2, -1, 0, 1, 2\}$, so with $K = 3$, s_{disc} takes at most 13 values, of which only 6 to 10 appear in

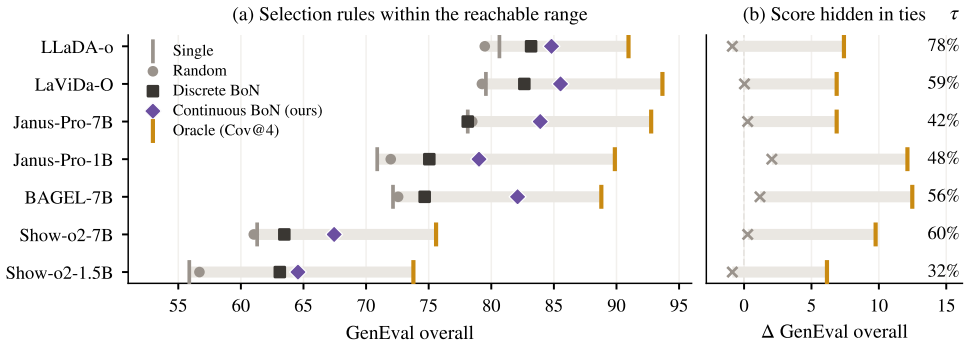

Figure 2: GenEval diagnosis on the same four candidate pools. (a) Selection performance across rules. (b) Recoverable performance within score ties, where τ denotes the four-candidate tie rate.

practice. The four-candidate tie rate $\tau = \Pr_p[s(x_1, p) = \cdots = s(x_N, p)]$ ranges from 32.4% to 77.9% on GenEval and reaches 83.4% on DPG-Bench. Under the baseline tie rule, these prompts reduce Best-of-N to Best-of-1.

Score collisions are only one symptom of the bottleneck. Random tie breaking recovers little headroom, and Janus-Pro-7B recovers none despite a 42.5% tie rate. Even when discrete scores differ, their ordering can remain weak. On Janus-Pro-7B, reading the continuous margin raises the advantage over chance from 2.5 to 16.2 percentage points. Improving the candidate pool raises coverage@4 by 0.7 to 2.0 points but selection by only 1.8 to 2.4, less than the 2.9 to 7.4 obtained by changing the readout alone (Appendix A.7). Together, these observations indicate that discrete verdicts compress richer preferences already present in verifier evidence.

3.3 Finite Observations Reveal Continuous Preference

We model discrete readout without restricting how verifier evidence is produced. Let H denote the verification history before the final judgment, including decomposition, reasoning, or other computation. Conditional on H, a stochastic verification run yields Yes, No, or neither with probabilities $p_Y(H)$, $p_N(H)$, and $p_0(H)$. Marginalizing over H gives

$$\bar{p}_Y = \mathbb{E}_H[p_Y(H)], \qquad \bar{p}_N = \mathbb{E}_H[p_N(H)], \qquad \bar{p}_0 = 1 - \bar{p}_Y - \bar{p}_N. \tag{4}$$

For m independent observations, $(n_Y^{(m)}, n_N^{(m)}, n_0^{(m)})$ is multinomial. With $M^{(m)} = n_Y^{(m)} + n_N^{(m)}$,

$$\begin{aligned} M^{(m)} &\sim \mathrm{Binomial}(m, \bar{p}_Y + \bar{p}_N), \\ n_Y^{(m)} \mid M^{(m)} &\sim \mathrm{Binomial}\Big(M^{(m)}, P_Y\Big), \qquad P_Y = \frac{\bar{p}_Y}{\bar{p}_Y + \bar{p}_N}, \qquad P_N = 1 - P_Y. \end{aligned} \tag{5}$$

We estimate the underlying preference by

$$a_{\mathrm{rep}}^{(m)}(x, q) = \log \frac{n_Y^{(m)} + \alpha}{n_N^{(m)} + \alpha}, \tag{6}$$

where $\alpha > 0$ provides smoothing. By the strong law of large numbers,

$$a_{\mathrm{rep}}^{(m)}(x, q) \longrightarrow \log \frac{\bar{p}_Y}{\bar{p}_N} \equiv a_\star(x, q) \tag{7}$$

almost surely as $m \to \infty$. Thus, repeated discrete observations converge to a continuous preference regardless of the preceding verification process.

In our direct answer setting, the verification history is fixed and the Yes/No probabilities are available from the model logits. Let $z(x, q) \in \mathbb{R}^{|V|}$ denote these logits, and $\mathcal{Y}, \mathcal{N}$ the Yes and No token sets. Then $p_Y = \sum_{t \in \mathcal{Y}} \mathrm{softmax}(z)_t$ and $p_N = \sum_{t \in \mathcal{N}} \mathrm{softmax}(z)_t$, giving

$$a_\star(x, q) = \log \frac{p_Y}{p_N} = \underbrace{\log \sum_{t \in \mathcal{Y}} e^{z_t(x,q)} - \log \sum_{t \in \mathcal{N}} e^{z_t(x,q)}}_{a_{\mathrm{cont}}(x,q)}. \tag{8}$$

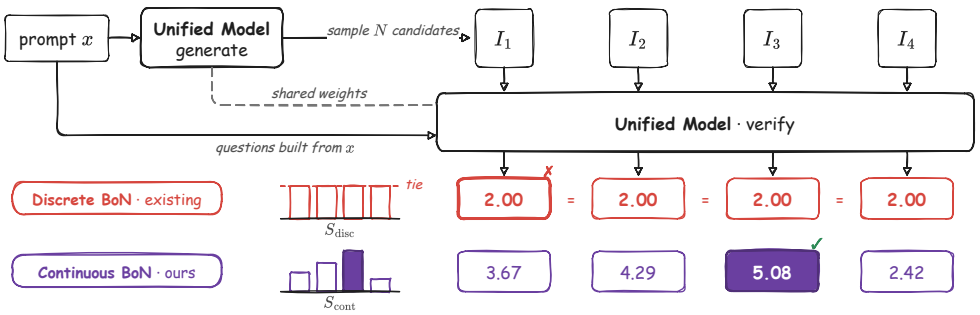

Figure 3: Discrete verdicts compress verifier evidence, while continuous readout preserves the underlying Yes versus No preference.

The pre-decoding margin is therefore the accessible infinite-observation preference. The candidate score converges to s_{cont}, and with a unique maximizer, selection converges to C-BoN.

Equations equation 5 and equation 6 characterize finite-observation uncertainty and determine finite-m scoring and selection across questions. §4.3 shows that full stochastic decoding closely tracks the predicted recovery behavior.

This view separates verifier preference from its observed resolution. C-BoN reads a_{cont} while keeping the questions, aggregation, and selection unchanged. Algorithm 1 in Appendix A.5 gives both readouts. On pairs with $a_{\text{disc}} \neq 0$, the sign of a_{cont} agrees with the discrete verdict on 97.2% to 100% of pairs. The main difference is the preference information retained rather than a systematic reversal of the verdict. §4.6 tests whether this strength predicts candidate quality, while §4.5 isolates score collisions.

3.4 Scope of the Readout Bottleneck

The observation model applies after verifier evidence is produced and is not restricted to direct answer verification. The history H may include decomposition, reasoning, or other computation. We use direct answers because the limiting preference is available from the Yes/No logits, while stochastic histories may require sampling over verification paths. Higher fidelity readout can recover existing distinctions, but cannot create missing evidence or correct wrong Yes/No ordering. §4.6 tests when recovered preference improves ordering.

4 Experiments and Analysis

4.1 Setup

Models and benchmarks. We evaluate seven unified image understanding and generation models: BAGEL-7B-MoT (Deng et al., 2025), Janus-Pro-7B/1B (Chen et al., 2025), Show-o2-7B/1.5B (Xie et al., 2026), LaViDa-O (Li et al., 2026), and LLaDA-o (You et al., 2026). They span autoregressive VQ, masked discrete diffusion, and continuous diffusion or flow-based generation. All use released checkpoints, inference settings, and native understanding pathways. We evaluate on GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023), GenAI-Bench (Jiang et al., 2024), and DPG-Bench (Hu et al., 2024) with official evaluators.

Baselines and evaluation. For each prompt, we generate $N = 4$ candidates and compare a single sample, random selection, Discrete BoN, Qwen3-VL-2B/8B-Instruct as external verifiers, and C-BoN. Discrete BoN and C-BoN use identical candidates, questions, verifier evaluations, aggregation, and selection, differing only in the discrete verdict versus the predecoding Yes versus No margin. On GenEval, we also use stochastic decoding to test the finite observation model in §3.3.

Table 1: Main results on four T2I benchmarks. Gray numbers show gains over Discrete BoN.

Model	Single	Random	Discrete BoN	Qwen BoN 2B	Qwen BoN 8B	C-BoN (ours)
			GenEval			
BAGEL-7B-MoT	72.15	72.56	74.68	71.43	76.49	**82.10** (+7.41)
Show-o2-7B	61.30	61.03	63.47	61.66	67.09	**67.45** (+3.98)
LaViDa-O	79.57	79.25	82.64	80.29	81.74	**85.53** (+2.89)
Show-o2-1.5B	55.88	56.69	63.11	59.86	62.93	**64.56** (+1.45)
Janus-Pro-7B	78.12	78.48	78.12	80.47	83.73	**83.91** (+5.79)
Janus-Pro-1B	70.89	71.97	75.05	71.61	77.94	**79.02** (+3.98)
LLaDA-o	80.65	79.48	83.18	80.47	82.82	**84.81** (+1.63)
			T2I-CompBench			
BAGEL-7B-MoT	51.61	51.35	53.59	52.18	53.58	**54.70** (+1.11)
Show-o2-7B	48.55	48.32	50.75	49.91	51.22	**52.61** (+1.87)
LaViDa-O	54.20	54.27	55.31	54.98	55.57	**55.76** (+0.46)
Show-o2-1.5B	46.24	46.00	47.83	47.81	49.35	**50.73** (+2.90)
Janus-Pro-7B	35.14	35.40	37.18	38.17	38.73	**40.51** (+3.33)
Janus-Pro-1B	30.96	31.31	34.37	33.69	34.88	**35.76** (+1.39)
LLaDA-o	54.44	54.15	55.82	54.52	55.29	**56.54** (+0.72)
			GenAI-Bench			
BAGEL-7B-MoT	77.58	77.62	80.00	79.02	79.86	**81.65** (+1.65)
Show-o2-7B	72.79	73.17	75.14	74.44	75.48	**76.99** (+1.85)
LaViDa-O	76.59	76.56	77.94	77.65	78.48	**78.70** (+0.76)
Show-o2-1.5B	71.66	71.67	72.96	72.77	74.37	**75.53** (+2.57)
Janus-Pro-7B	75.21	75.34	75.97	76.63	77.49	**78.54** (+2.57)
Janus-Pro-1B	70.59	70.65	72.14	72.54	73.59	**74.20** (+2.06)
LLaDA-o	80.00	80.18	81.76	81.33	82.14	**82.80** (+1.04)
			DPG-Bench			
BAGEL-7B-MoT	85.44	85.57	85.74	86.15	85.97	**86.73** (+0.99)
Show-o2-7B	84.82	85.23	84.64	85.55	86.51	**86.97** (+2.33)
LaViDa-O	82.95	82.91	83.09	83.39	83.48	**83.74** (+0.65)
Show-o2-1.5B	83.46	83.87	83.92	84.18	84.76	**85.07** (+1.15)
Janus-Pro-7B	83.59	83.87	84.01	84.82	84.68	**85.08** (+1.08)
Janus-Pro-1B	82.32	82.23	82.01	83.32	**83.82**	82.88 (+0.88)
LLaDA-o	87.25	87.22	87.36	87.75	87.71	**87.83** (+0.48)

4.2 Readout Fidelity Consistently Improves Selection

Table 1 shows that continuous readout improves all 28 settings over discrete readout by 0.46 to 7.41 points and is best in 27. On GenEval, C-BoN also outperforms Qwen3-VL-8B with discrete readout for all seven models. With matched continuous readout, the native verifier remains within 0.50 points of Qwen3-VL-8B in benchmark means (Appendix A.1). Human evaluation and significance tests are in Appendices A.12 and A.9.

4.3 Does the Finite Observation Model Predict Stochastic Decoding?

Table 2: Predicted and observed recovery on GenEval. Each entry reports Pred./Obs. from the finite-observation model and full stochastic decoding.

Model	$m = 1$	$m = 2$	$m = 4$	$m = 8$	$m = 16$	$m = 32$
BAGEL-7B-MoT	76.52/75.77	77.79/76.85	78.58/77.94	79.64/78.84	80.30/79.75	80.90/80.47
Show-o2-7B	64.70/64.20	65.33/65.82	66.05/66.00	66.41/66.37	66.68/66.73	66.75/67.09
LaViDa-O	81.88/82.64	82.17/84.63	82.92/84.81	83.58/84.99	83.98/85.53	84.35/85.35
Show-o2-1.5B	59.35/60.22	60.46/60.94	61.65/62.21	62.46/63.65	63.16/63.65	63.94/64.38
Janus-Pro-7B	79.73/77.76	80.12/76.85	80.15/78.66	80.33/79.02	80.98/80.47	81.46/81.01
Janus-Pro-1B	75.27/73.42	76.17/74.68	77.32/77.03	77.79/79.39	78.28/79.75	78.70/79.39
LLaDA-o	83.66/82.46	83.85/82.82	84.01/82.64	84.10/83.18	84.16/83.36	84.49/83.73

For our direct answer setting, Section 3.3 instantiates the observation model using the Yes/No/other distribution available from the verifier logits. We test its finite observation prediction against full

stochastic decoding. The predicted condition samples m observations from the cached Yes/No/other distribution, while the observed condition generates m complete responses and applies the same Yes/No/otherwise parser. All other components are fixed. Table 2 shows close overall agreement across 42 settings, with a mean absolute error of 0.98 GenEval points and Pearson $r = 0.9894$. After centering within each model, the correlation remains $r = 0.9019$. As m increases, predicted and observed performance generally move toward continuous readout despite model specific deviations. This agreement supports the finite observation model as a quantitative account of stochastic recovery.

Finding 1. The finite observation model predicts stochastic recovery. Predicted performance closely tracks full stochastic decoding across models and observation counts.

4.4 Does Self-Verified Best-of-N Actually Scale?

Table 3: Best-of-N scaling on GenEval. C-BoN scales while Discrete BoN saturates.

	Discrete				C-BoN (ours)			
Model	N=1	N=2	N=4	N=8	N=1	N=2	N=4	N=8
BAGEL-7B-MoT	72.40	75.24	75.56	75.23	72.40	78.50	82.42	**84.45**
Show-o2-7B	61.28	62.84	63.87	64.01	61.28	65.35	68.69	**71.07**
LaViDa-O	79.02	81.16	82.15	82.82	79.02	82.47	84.28	**86.26**
Show-o2-1.5B	56.87	59.80	61.75	63.29	56.87	61.47	64.77	**67.09**
Janus-Pro-7B	79.18	79.45	78.34	76.85	79.18	82.34	83.42	**83.54**
Janus-Pro-1B	71.88	75.31	77.15	77.40	71.88	76.34	79.62	**81.37**
LLaDA-o	79.62	81.92	82.80	83.36	79.62	83.25	85.47	**87.52**
Mean	71.46	73.67	74.52	74.71	71.46	75.67	78.38	**80.19**

We generate four additional candidates per prompt, pool all eight, and average each size N rule over 200 random subsets (Table 3). Averaged over seven models, Discrete BoN rises from 71.46 at N=1 to only 74.71 at N=8, while continuous readout reaches 80.19. Janus-Pro-7B even falls below its single sample baseline at N=8 under Discrete BoN, while C-BoN improves monotonically. Additional sampling therefore helps only when the selector preserves candidate ordering.

Finding 2. Readout fidelity limits Best-of-N scaling. Larger pools create more opportunity, but discrete readout cannot reliably exploit it. Preserving verifier preference converts sampling into gains.

4.5 What Information Does the Readout Discard?

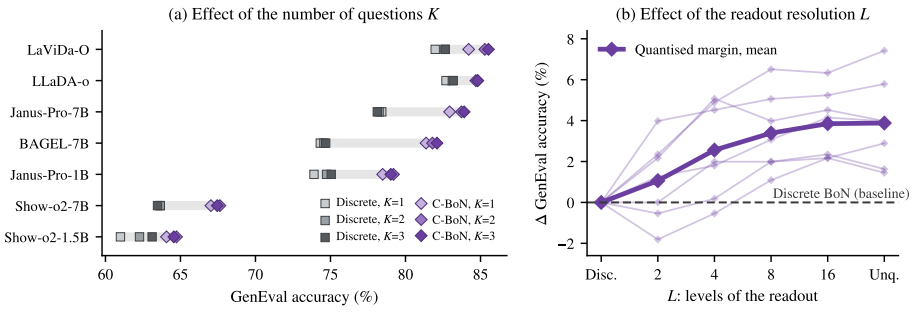

Figure 4: Ablations on question count and readout quantization. (a) C-BoN remains effective with fewer questions. (b) Quantized margins outperform the discrete readout even at low cardinality.

Figure 4 varies question count K and margin quantization levels L. We quantize a_{cont} into L equal width bins over each model's global range with fixed edges across prompts. Continuous readout remains effective with fewer questions, and even two levels outperform the discrete verdict.

Thus, the effect depends on which preference distinctions survive, not score cardinality. Alternative continuous forms show the same pattern in Appendix A.3.

Table 4: Tie resolution versus full continuous ranking. Gray numbers show row-wise gains.

Readout	GenEval	T2I-CompBench	GenAI-Bench	DPG-Bench
Discrete BoN	74.32	47.83	76.56	84.39
Margin tie-break	77.37 (+3.05)	48.93 (+1.09)	77.70 (+1.14)	85.26 (+0.87)
C-BoN (ours)	**78.20** (+0.83)	**49.52** (+0.59)	**78.34** (+0.64)	**85.47** (+0.21)

Beyond tie breaking. Using the margin only among candidates tied for the maximum discrete score explains 64% to 81% of the total gain (Table 4). Full continuous ranking still improves all four benchmarks, with 95% bootstrap CIs excluding zero. Score collisions are therefore a major symptom, but not the entire readout bottleneck.

4.6 Does Recovered Preference Predict Candidate Quality?

If discrete verdicts discard useful preference strength, larger recovered margin gaps should yield more reliable candidate ordering. We test this among candidate pairs with identical discrete scores but different benchmark scores.

Table 5: GenEval ranking reliability within discrete ties.

	Q1	Q2	Q3	Q4	Q5
Ranking acc. (%)	54.5	57.9	59.9	69.5	**84.1**
Median $\lvert\Delta s_{\text{cont}}\rvert$	0.08	0.29	0.50	0.92	2.04

Table 5 shows a strong monotonic relation on GenEval. As the median margin gap grows from 0.08 to 2.04, ranking accuracy rises from 54.5% to 84.1%. The same trend holds on T2I-CompBench ($51.4\% \to 63.2\%$), GenAI-Bench ($50.8\% \to 70.8\%$), and DPG-Bench ($50.4\% \to 63.0\%$). Preference strength discarded by discrete readout therefore carries systematic information about candidate quality across all four benchmarks.

We next test whether recovered ordering tracks downstream gains using pairwise ranking accuracy:

$$\rho_r = \Pr\big[s_r(x^+) > s_r(x^-)\big] + \frac{1}{2}\Pr\big[s_r(x^+) = s_r(x^-)\big], \tag{9}$$

where x^+ and x^- denote candidates with higher and lower benchmark scores, and $r \in \{\text{disc}, \text{cont}\}$. Benchmark ties are excluded and readout ties receive half credit, so $\rho_r = \frac{1}{2}$ corresponds to random pairwise ranking.

Figure 5 shows the broader pattern across 98 model and category cells from GenEval and T2I-CompBench. The continuous margin improves ρ in 90 cells, 83 of which also improve in benchmark score. Seven cells move from below chance to above chance, compared with two in the opposite direction. Aggregated across models, ρ increases from 0.624 to 0.726 on GenEval and from 0.535 to 0.580 on T2I-CompBench. Moreover, $\Delta\rho = \rho_{\text{cont}} - \rho_{\text{disc}}$ tracks downstream score gain, with Pearson $r = 0.45$ and Spearman $r_s = 0.62$. Bootstrap intervals are $[0.27, 0.62]$ and $[0.40, 0.79]$, respectively. Full per model results are in Appendix A.2.

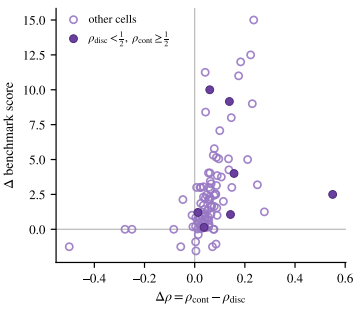

Figure 5: Recovered ordering tracks downstream gains across 98 cells. Filled markers cross above chance.

Finding 3. Recovered preference predicts candidate quality. Larger continuous margins yield more reliable ordering within discrete ties, while recovered ordering tracks downstream gains.

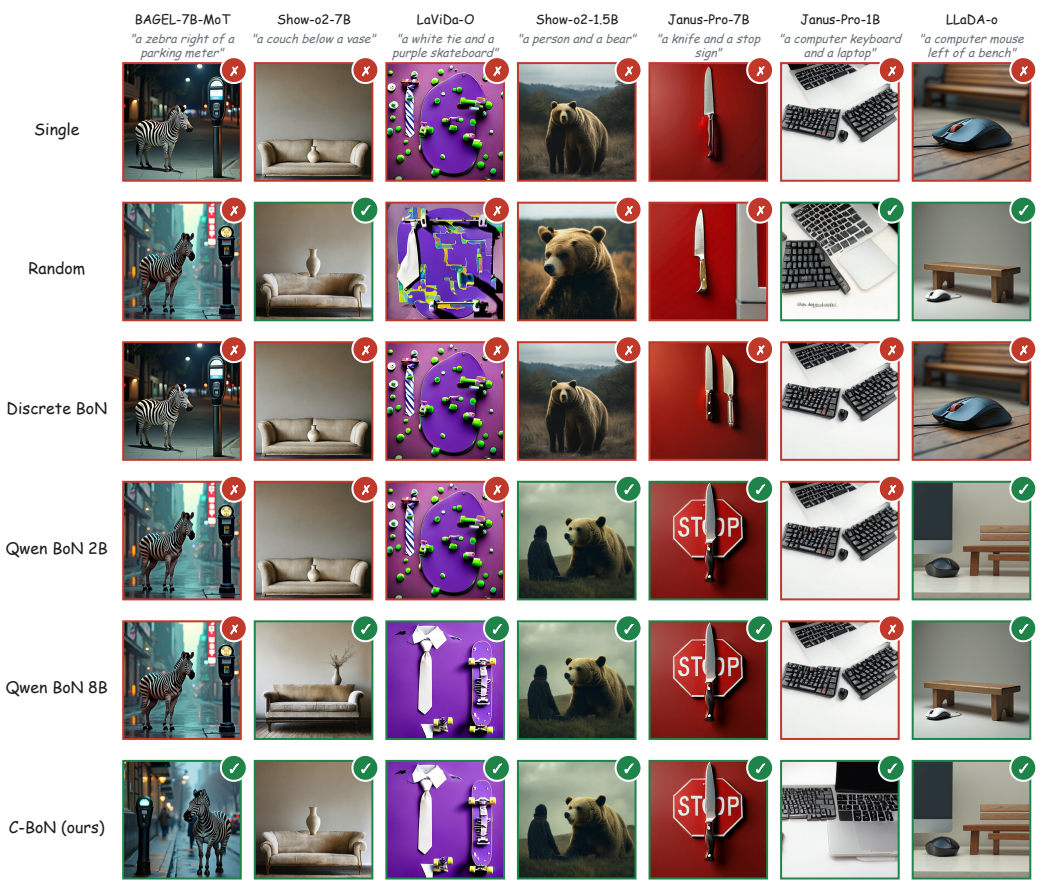

Figure 6: Each column shows one model and prompt, while each row shows the candidate selected by a different rule. Green and red markers denote correct and incorrect selections.

4.7 Qualitative Analysis

Figure 6 illustrates the same readout failure across all seven models. In six of seven examples, Discrete BoN returns the single sample candidate because all four candidates receive the same score despite substantial visual differences. Qwen BoN 2B and 8B succeed in three and five cases, respectively, while C-BoN succeeds in all seven. The examples cover object count, attributes, spatial relations, and object composition, indicating that the hidden preference is not restricted to one type of generation error. Useful candidate distinctions can therefore remain in verifier evidence even when decoded scores collapse.

5 Conclusion

We study why self-verified Best-of-N fails to fully exploit the quality already available in candidate pools. We identify readout fidelity as a key bottleneck and show that discrete verdicts can discard useful candidate preferences already present in verifier evidence. Under a general observation model, repeated stochastic observations converge to a marginal Yes/No preference, which reduces to the pre-decoding logit margin for direct-answer verification. Full stochastic decoding closely follows the predicted finite-observation behavior. A controlled continuous readout that exposes this preference consistently improves selection across seven models and four benchmarks without additional training, generation, or verifier calls. Larger recovered margins also yield more reliable ordering among candidates tied by discrete readout, showing that the discarded preference strength is informative. More broadly, our results separate verifier evidence quality from readout fidelity and show that preserving candidate ordering is critical for effective Best-of-N scaling.

AI USE STATEMENT

In this work, we used generative AI tools to assist with mathematical derivations, coding, and improving the clarity and readability of the manuscript. All AI-assisted mathematical derivations were independently checked by the authors, all AI-assisted code was verified and tested for correctness, and all AI-assisted text was reviewed and revised by the authors. We did not use generative AI tools to generate research hypotheses, design the methodology or experiments, or analyze and interpret experimental results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Leon Liangyu Chen, Haoyu Ma, Zhipeng Fan, Ziqi Huang, Animesh Sinha, Xiaoliang Dai, Jialiang Wang, Zecheng He, Jianwei Yang, Chunyuan Li, et al. Unit: Unified multimodal chain-of-thought test-time scaling. *arXiv preprint arXiv:2602.12279*, 2026.

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.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Runpei Dong, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *International Conference on Learning Representations*, volume 2024, pp. 6666–6702, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

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 *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

Weiyang Jin, Yuwei Niu, Jiaqi Liao, Chengqi Duan, Aoxue Li, Shenghua Gao, and Xihui Liu. Srum: Fine-grained self-rewarding for unified multimodal models. *arXiv preprint arXiv:2510.12784*, 2025.

Yongjin Kim, Yoonjin Oh, Yerin Kim, Hyomin Kim, Jeeyoung Yun, Yujung Heo, Minjun Kim, and Sungwoong Kim. Enhanced text-to-image generation by fine-grained multimodal reasoning. *arXiv e-prints*, pp. arXiv–2604, 2026a.

Yongjin Kim, Yoonjin Oh, Yerin Kim, Hyomin Kim, Jeeyoung Yun, Yujung Heo, Minjun Kim, and Sungwoong Kim. Fire: Fine-grained multimodal reasoning for enhanced image generation. *arXiv preprint arXiv:2604.13491*, 2026b.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2026, pp. 50571–50603, 2026.

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*, pp. 366–384. Springer, 2024.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Jiadong Pan, Liang Li, Yuxin Peng, Yu-Ming Tang, Shuohuan Wang, Yu Sun, Hua Wu, Qingming Huang, and Haifeng Wang. Learning to generate via understanding: Understanding-driven intrinsic rewarding for unified multimodal models. *arXiv preprint arXiv:2603.06043*, 2026.

Vignesh Sundaresha, Akash Haridas, Vikram Appia, and Lav R Varshney. An efficient test-time scaling approach for image generation. *arXiv e-prints*, pp. arXiv–2512, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Ritesh Thawkar, Shravan Venkatraman, Omkar Thawakar, Abdelrahman Shaker, Fahad Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. Ask, solve, generate: Self-evolving unified multimodal understanding and generation via self-consistency rewards. *arXiv preprint arXiv:2606.27376*, 2026.

Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *Advances in Neural Information Processing Systems*, 38:152386–152415, 2026.

Kaishen Wang, Ruibo Chen, Tong Zheng, and Heng Huang. Imagent: A unified multimodal agent framework for test-time scalable image generation. *arXiv preprint arXiv:2511.11483*, 2025.

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.

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.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 93620–93638, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

Yi Xin, Siqi Luo, Tianxiang Xu, Qi Qin, Haoxing Chen, Kaiwen Zhu, Zhiwei Zhang, Yangfan He, Rongchao Zhang, Jinbin Bai, et al. dmllm-tts: Self-verified and efficient test-time scaling for diffusion multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35726–35735, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026.

Zebin You, Xiaolu Zhang, Jun Zhou, Chongxuan Li, and Ji-Rong Wen. Llada-o: An effective and length-adaptive omni diffusion model. *arXiv preprint arXiv:2603.01068*, 2026.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Y Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. In *International Conference on Learning Representations*, volume 2026, pp. 39133–39166, 2026.

A Appendix

A.1 Readout Fidelity with External Verifiers

Our main experiments ask whether a unified model can better exploit the verification signal already present in its own understanding pathway. As a control, we test whether the same readout bottleneck appears with independent external verifiers. Table 6 applies continuous readout to Qwen3-VL-2B and Qwen3-VL-8B while keeping the candidate images, verification questions, and verifier forward passes fixed. For each Qwen verifier, the decoded and continuous variants differ only in the readout.

Continuous readout improves both external verifiers in every model and benchmark setting. Averaged over seven candidate generators, replacing decoded answers with continuous margins improves Qwen3-VL-2B by 5.71, 2.40, 2.32, and 0.65 points on GenEval, T2I-CompBench, GenAI-Bench, and DPG-Bench, respectively. The corresponding gains for Qwen3-VL-8B are 1.91, 1.47, 1.50, and 0.53 points. Thus, information loss at the discrete readout is not specific to the native understanding pathway of a unified model.

The native verifier also remains competitive when external verifiers receive the same continuous readout. Averaged across models, C-BoN scores 78.20, 49.52, 78.34, and 85.47 on the four benchmarks, compared with 78.02, 49.85, 78.84, and 85.80 for continuous Qwen3-VL-8B. The absolute difference is at most 0.50 points on every benchmark, and the native verifier achieves the highest mean score on GenEval. Against continuous Qwen3-VL-2B, the benchmark level differences are only 0.23, 0.21, 0.32, and 0.20 points. These results show that preserving candidate ordering can make the verification signal already available within a unified model competitive with strong external verifiers.

A.2 Additional Pairwise Ranking Analysis

Margin strength and ranking reliability. To isolate information unavailable to the discrete readout, we consider candidate pairs with identical discrete scores but different benchmark scores. For each pair, we compute the absolute continuous score gap $|\Delta s_{\text{cont}}|$, divide pairs into quintiles within each benchmark, and measure whether the continuous score ranks the better candidate higher. Every analyzed pair is tied under Discrete BoN, so this analysis directly tests whether preference strength discarded by discrete verdicts contains useful ordering information.

Table 7 shows a monotonic increase from Q1 to Q5 on all four benchmarks. On GenEval, the analysis contains 2,197 tied pairs, and ranking accuracy rises from 54.5% to 84.1% as the median margin gap increases from 0.08 to 2.04. The same monotonic pattern holds on T2I-CompBench, GenAI-Bench, and DPG-Bench, where accuracy rises from 51.4% to 63.2%, 50.8% to 70.8%, and 50.4% to 63.0%, respectively. Preference strength discarded by discrete readout therefore contains systematic information about candidate quality.

Per model ranking accuracy. Table 8 provides the per model ranking results underlying the analysis in Section 4.6. The pairwise ranking accuracy ρ is computed over candidate pairs with different benchmark scores. Benchmark ties are excluded and readout ties receive half credit. The discrete

Table 6: Continuous vs. discrete readout for Qwen3-VL verifiers. Gray numbers show gains from continuous readout over its discrete counterpart, and bold marks the best in each row.

		Qwen BoN 2B		Qwen BoN 8B		
Model	Discrete BoN	Discrete	Continuous	Discrete	Continuous	C-BoN (ours)
			GenEval			
BAGEL-7B-MoT	74.68	71.43	80.65 (+9.22)	76.49	80.29 (+3.80)	**82.10** (+7.41)
Show-o2-7B	63.47	61.66	67.09 (+5.43)	67.09	**68.17** (+1.08)	67.45 (+3.98)
LaViDa-O	82.64	80.29	82.64 (+2.35)	81.74	83.18 (+1.44)	**85.53** (+2.89)
Show-o2-1.5B	63.11	59.86	**66.73** (+6.87)	62.93	64.92 (+1.99)	64.56 (+1.45)
Janus-Pro-7B	78.12	80.47	**84.99** (+4.52)	83.73	84.09 (+0.36)	83.91 (+5.79)
Janus-Pro-1B	75.05	71.61	79.39 (+7.78)	77.94	**82.10** (+4.16)	79.02 (+3.98)
LLaDA-o	83.18	80.47	84.27 (+3.80)	82.82	83.36 (+0.54)	**84.81** (+1.63)
			T2I-CompBench			
BAGEL-7B-MoT	53.59	52.18	54.65 (+2.47)	53.58	**54.78** (+1.20)	54.70 (+1.11)
Show-o2-7B	50.75	49.91	52.41 (+2.50)	51.22	**52.63** (+1.41)	52.61 (+1.87)
LaViDa-O	55.31	54.98	56.20 (+1.22)	55.57	**56.43** (+0.86)	55.76 (+0.46)
Show-o2-1.5B	47.83	47.81	50.41 (+2.60)	49.35	**50.85** (+1.50)	50.73 (+2.90)
Janus-Pro-7B	37.18	38.17	**40.96** (+2.79)	38.73	40.85 (+2.12)	40.51 (+3.33)
Janus-Pro-1B	34.37	33.69	37.14 (+3.45)	34.88	**37.23** (+2.35)	35.76 (+1.39)
LLaDA-o	55.82	54.52	56.30 (+1.78)	55.29	56.16 (+0.87)	**56.54** (+0.72)
			GenAI-Bench			
BAGEL-7B-MoT	80.00	79.02	81.60 (+2.58)	79.86	81.44 (+1.58)	**81.65** (+1.65)
Show-o2-7B	75.14	74.44	76.48 (+2.04)	75.48	76.85 (+1.37)	**76.99** (+1.85)
LaViDa-O	77.94	77.65	79.48 (+1.83)	78.48	**79.57** (+1.09)	78.70 (+0.76)
Show-o2-1.5B	72.96	72.77	75.32 (+2.55)	74.37	**75.96** (+1.59)	75.53 (+2.57)
Janus-Pro-7B	75.97	76.63	78.84 (+2.21)	77.49	**78.95** (+1.46)	78.54 (+2.57)
Janus-Pro-1B	72.14	72.54	75.99 (+3.45)	73.59	**76.01** (+2.42)	74.20 (+2.06)
LLaDA-o	81.76	81.33	82.91 (+1.58)	82.14	**83.12** (+0.98)	82.80 (+1.04)
			DPG-Bench			
BAGEL-7B-MoT	85.74	86.15	86.53 (+0.38)	85.97	86.41 (+0.44)	**86.73** (+0.99)
Show-o2-7B	84.64	85.55	86.85 (+1.30)	86.51	86.71 (+0.20)	**86.97** (+2.33)
LaViDa-O	83.09	83.39	83.64 (+0.25)	83.48	**83.91** (+0.43)	83.74 (+0.65)
Show-o2-1.5B	83.92	84.18	85.21 (+1.03)	84.76	**85.43** (+0.67)	85.07 (+1.15)
Janus-Pro-7B	84.01	84.82	85.43 (+0.61)	84.68	**85.74** (+1.06)	85.08 (+1.08)
Janus-Pro-1B	82.01	83.32	84.21 (+0.89)	83.82	**84.54** (+0.72)	82.88 (+0.88)
LLaDA-o	87.36	87.75	87.82 (+0.07)	87.71	**87.89** (+0.18)	87.83 (+0.48)

Table 7: Ranking accuracy (%) among candidate pairs tied by the discrete readout, grouped by continuous score gap.

Benchmark	Q1	Q2	Q3	Q4	Q5
GenEval	54.5	57.9	59.9	69.5	**84.1**
T2I-CompBench	51.4	52.5	53.1	56.4	**63.2**
GenAI-Bench	50.8	54.2	56.5	60.3	**70.8**
DPG-Bench	50.4	51.5	54.6	55.5	**63.0**

and continuous columns use the same candidate pools, verification questions, and verifier evaluations, differing only in the answer readout. Continuous readout improves ranking accuracy for every model on both benchmarks, showing that the recovered ordering is consistent across model families rather than driven by one model.

Correlation uncertainty. The correlations reported in Section 4.6 are computed over the 98 model and category cells of GenEval and T2I-CompBench. Because cells sharing the same model are not independent, we bootstrap at the model level. Each of 20,000 replicates samples seven models with replacement and retains all associated category cells. The 95% percentile intervals are $[0.27, 0.62]$ for Pearson $r = 0.45$ and $[0.40, 0.79]$ for Spearman $r_s = 0.62$.

Table 8: Pairwise ranking accuracy ρ under discrete and continuous readouts. Continuous readout improves ranking for every model on both benchmarks.

	GenEval		T2I-CompBench	
Model	Disc.	Continuous	Disc.	Continuous
BAGEL-7B-MoT	0.623	**0.786**	0.544	**0.592**
Show-o2-7B	0.609	**0.757**	0.539	**0.594**
LaViDa-O	0.640	**0.716**	0.519	**0.546**
Show-o2-1.5B	0.659	**0.737**	0.534	**0.591**
Janus-Pro-7B	0.525	**0.662**	0.538	**0.591**
Janus-Pro-1B	0.654	**0.696**	0.549	**0.581**
LLaDA-o	0.654	**0.729**	0.524	**0.563**
Mean	0.624	**0.726**	0.535	**0.580**

Chance corresponds to $\rho = \frac{1}{2}$ under equation 9. The mean row averages over the seven models.

A.3 Robustness to Continuous Scoring Form

We test whether the readout effect depends on the particular continuous score used by C-BoN. We separate two choices: whether the answer signal uses the normalized Yes probability or its log odds, and whether the score uses only positive questions or also their negated counterparts. All variants use the same GenEval candidate pools and verifier outputs.

Table 9: Continuous readout variants on GenEval, averaged over seven models. All continuous variants substantially outperform Discrete BoN.

Readout	GenEval	ρ
Discrete BoN	74.32	0.624
VQAScore style	77.47	0.703
Positive log odds	**78.51**	**0.734**
Paired probability	78.46	0.732
C-BoN	78.20	0.726

The VQAScore style variant averages P_Y over the positive questions. Positive log odds instead averages $\log(P_Y/P_N)$ over the same questions. Paired probability uses the difference between P_Y for each positive and negated question, while C-BoN uses the corresponding difference in log odds. Under conditional Yes/No normalization, probability and log odds are invertible transformations of the same per question evidence. Their difference therefore concerns how preference strength is aggregated rather than how much information each contains.

All four continuous variants substantially outperform Discrete BoN. The GenEval gain ranges from 3.15 to 4.19 points, while ρ increases by 0.079 to 0.110. A positive only VQAScore style readout already recovers much of the improvement, and the remaining continuous variants perform similarly. Positive log odds achieves the highest mean score and ranking accuracy. The differences among continuous formulations are smaller than their common gain over discrete readout. This supports the finding that preserving predecoding preference information matters more than the particular continuous scoring form. We retain C-BoN in the main analysis because it changes only the answer readout while preserving the original questions, negation pairing, aggregation, and selection rule.

A.4 Verification Questions

The three phrasings and their negations are used unchanged for every model, benchmark, and candidate. Here a is the prompt from which the image was generated, inserted verbatim.

k	Positive q_k	Negative $\bar{q}_k$
1	`Does this image show {a}?`	`Is {a} missing from this image?`
2	`Is this a correct picture of {a}?`	`Is this picture wrong for {a}?`
3	`Does the image match the description: {a}?`	`Does the image contradict: {a}?`

Each question is followed by `Answer Yes or No.` on a new line. Pairing each question with its negation makes the score a difference and is designed to cancel a uniform preference for answering Yes. Under the discrete verdict, 71.8% of question pairs yield the $+2$ pattern and only 0.5% yield the -2 pattern, while 26.4% return the same answer to a question and its negation and therefore contribute zero.

A.5 Controlled Readout Comparison

Algorithm 1 makes explicit that the discrete and continuous conditions share the same candidates, questions, aggregation, and final selection rule. They differ only in how the answer signal is read.

Algorithm 1 Discrete and Continuous Self Verification

Require: Prompt p, candidates $x_1, \ldots, x_N$, readout r

for $n = 1, \ldots, N$ **do**
  for $k = 1, \ldots, K$ **do**
    for $q \in \{q_k(p), \bar{q}_k(p)\}$ **do**
      if $r = \text{disc}$ **then**
        $y \leftarrow$ greedy decoded response of $U(x_n, q)$
        $a[q] \leftarrow a_{\text{disc}}(x_n, q)$ from equation 3
      else
        $z \leftarrow$ Yes/No logits of $U(x_n, q)$ before decoding
        $a[q] \leftarrow \log \sum_{t \in \mathcal{Y}} e^{z_t} - \log \sum_{t \in \mathcal{N}} e^{z_t}$
      end if
    end for
  end for
  $s_n \leftarrow \frac{1}{K} \sum_{k=1}^{K} \left(a[q_k(p)] - a[\bar{q}_k(p)]\right)$
end for
$n^\star \leftarrow \arg\max_n s_n$, ties broken toward the lowest n
return $x_{n^\star}$

Repeated stochastic decoding. The experiment in Section 4.3 uses the same candidate pools, questions, negation pairing, aggregation, and selection rule. For each candidate and question, we independently generate m complete stochastic responses and parse each response with the same Yes/No/otherwise rule in equation 3. Stochastic decoding uses temperature 1.0 with no top p or top k truncation and a maximum of six generated tokens. The prompt requests an immediate Yes/No answer, so the verdict distribution matches the full softmax used in §3.3. Let $n_Y^{(m)}$, $n_N^{(m)}$, and $n_0^{(m)}$ denote the resulting counts. We convert $n_Y^{(m)}$ and $n_N^{(m)}$ to $a_{\text{rep}}^{(m)}$ using equation 6, with $\alpha = 0.5$ for every model and value of m. The resulting answer signals replace a in equation 2. Each stochastic response uses an independent sampling seed. The greedy Discrete BoN baseline is reported separately and is not identified with $m = 1$.

As a controlled reference, we also sample observations directly from the Yes/No/other distribution in §3.3 using (p_Y, p_N, p_0). Its finite m recovery curve closely tracks that obtained from full stochastic decoding across the reported values of m. This agreement links the finite observation model to the recovery behavior of stochastic verification.

A.6 Readout Fidelity under Chain of Thought Verification

Our main experiments use direct answer verification to isolate the readout factor cleanly. We additionally test whether the same bottleneck persists when verifier evidence is produced through Chain

of Thought reasoning. We use UniGen's CoT-V procedure (Tian et al., 2026), which decomposes the prompt into atomic visual judgments and evaluates them through a reasoning trace.

For a candidate x, CoT-V produces J atomic Yes/No judgments $A_1, \ldots, A_J$. Its discrete readout aggregates these judgments as

$$s_{\mathrm{cot,disc}}(x, p) = \frac{1}{J} \sum_{j=1}^{J} \mathbf{1}[A_j = \mathrm{Yes}]. \tag{10}$$

To isolate the readout, we keep the same decomposition and CoT trace and record the Yes/No logits z_j associated with each judgment. We replace the binary verdict with its continuous preference:

$$s_{\mathrm{cot,cont}}(x, p) = \frac{1}{J} \sum_{j=1}^{J} \left[\log \sum_{t \in \mathcal{Y}} e^{z_{j,t}} - \log \sum_{t \in \mathcal{N}} e^{z_{j,t}} \right]. \tag{11}$$

The two conditions therefore use identical CoT evidence and differ only in whether each judgment is discretized before candidate selection.

Table 10: Discrete and continuous readout under CoT-V. Both conditions use the same CoT evidence and differ only in the readout of each judgment.

Benchmark	Single	Discrete CoT-V	Continuous CoT-V	Gain	Tie rate
T2I-CompBench	65.36	66.42	67.08	+0.66	18.0%
GenAI-Bench	77.61	78.38	79.46	+1.08	29.1%
GenEval	71.61	73.60	75.41	+1.81	6.1%
DPG-Bench	85.43	85.36	86.08	+0.72	88.2%

Table 10 reports the controlled comparison. Both conditions use exactly the same CoT evidence and candidate images; they differ only in how each judgment is read out. Continuous readout improves CoT-V from 66.42 to 67.08 on T2I-CompBench (+0.66), from 78.38 to 79.46 on GenAI-Bench (+1.08), from 73.60 to 75.41 on GenEval (+1.81), and from 85.36 to 86.08 on DPG-Bench (+0.72). The corresponding discrete tie rates are 18.0%, 29.1%, 6.1%, and 88.2%, respectively. These results show that richer reasoning does not eliminate information loss at the final readout. The readout bottleneck therefore extends beyond direct answer verification to CoT-based self-verification: even when the model produces the same reasoning evidence, replacing a discrete Yes/No readout with a continuous pre-decoding signal consistently recovers additional information for candidate selection.

A.7 Intervention on Candidate Quality

We test a natural alternative explanation by improving the candidate pool. This intervention increases the available sampling opportunity without changing the readout and therefore tests whether limited candidate quality alone explains the observed selection gap.

A better candidate pool. We draw the N candidates from N logically equivalent prompt restatements instead of N sampling seeds, varying the order in which objects and relations are expressed. On BAGEL-7B-MoT, the oracle rises from 88.79 to 90.78, confirming that the restated pool contains additional correct candidates. However, under continuous readout the selected score changes from 82.10 to 81.56. Across the models for which we generated restated pools, coverage@4 rises by 0.7 to 2.0 points and Discrete BoN by 1.8 to 2.4 points, less than the 2.9 to 7.4 points obtained by changing the readout alone on the same models. Increased pool quality therefore does not by itself translate into better selected quality.

Interpretation. This intervention improves sampling opportunity while leaving the same low fidelity scoring interface in place. A better pool can contain more correct candidates, but exploiting those candidates still depends on preserving their ordering. Together with §4.6, this control separates candidate availability from readout fidelity and shows that improving coverage alone does not explain the observed selection bottleneck.

A.8 Readout Cost

Continuous readout changes only the answer interface, requires no additional verifier call, and reduces scoring latency in our implementation because it reads the Yes/No logits without decoding a response. Discrete BoN instead continues decoding to produce the response consumed by the Yes/No parser.

We measure both readouts on the same images and prompts on one device while alternating runs. On Janus-Pro-7B, a single question takes 85.8 ms with the continuous margin and 170.7 ms with the discrete verdict. Over six questions, this is 0.51 s versus 1.02 s per candidate. The saving follows from requiring fewer decoding steps, while the exact speedup depends on the model, implementation, and hardware.

Generation dominates total latency in either case, ranging from 0.4 s per candidate for Janus-Pro-7B at 384×384 to 23 s for LLaDA-o at 1024×1024. The readout therefore has a modest effect on total Best-of-N cost, but preserving the continuous preference introduces no additional scoring overhead and is faster in this measurement.

A.9 Significance Testing

Every comparison is paired at the prompt level because all rules select from the same candidate images. We compare continuous and discrete readout with McNemar's test on GenEval, whose labels are binary, and the Wilcoxon signed rank test on T2I-CompBench, GenAI-Bench, and DPG-Bench, whose scores are continuous. Table 11 reports the results for each setting.

Twenty three of the twenty eight gains reach $p < 0.05$ before correction for multiple testing. After Holm correction across all twenty eight tests, 21 remain significant. The seven that do not are Show-o2-1.5B, LaViDa-O, and LLaDA-o on GenEval, LaViDa-O on T2I-CompBench, and LaViDa-O, Janus-Pro-1B, and LLaDA-o on DPG-Bench. All seven gains remain positive.

Tie break bootstrap. For the tie break analysis in Table 4, we bootstrap models with replacement within each benchmark and recompute the mean difference between full continuous readout and the margin tie break variant. The 95% percentile intervals are $[+0.15, +1.94]$ on GenEval, $[+0.15, +1.10]$ on T2I-CompBench, $[+0.27, +1.05]$ on GenAI-Bench, and $[+0.08, +0.35]$ on DPG-Bench.

Table 11: Gain of continuous over discrete readout and its p value for each model and benchmark. McNemar is used on GenEval and Wilcoxon signed rank on the other three benchmarks. Bold marks uncorrected $p < 0.05$.

	GenEval		T2I-CompBench		GenAI-Bench		DPG-Bench	
Model	Δ	p	Δ	p	Δ	p	Δ	p
BAGEL-7B-MoT	+7.41	< 0.0001	+1.11	< 0.0001	+1.65	< 0.0001	+0.99	**0.0046**
Show-o2-7B	+3.98	**0.0026**	+1.87	< 0.0001	+1.85	< 0.0001	+2.33	< 0.0001
LaViDa-O	+2.89	**0.0226**	+0.46	**0.0384**	+0.76	< 0.0001	+0.65	0.1907
Show-o2-1.5B	+1.45	0.3222	+2.90	< 0.0001	+2.57	< 0.0001	+1.15	**0.0007**
Janus-Pro-7B	+5.79	< 0.0001	+3.33	< 0.0001	+2.57	< 0.0001	+1.08	**0.0003**
Janus-Pro-1B	+3.98	**0.0054**	+1.39	**0.0001**	+2.06	< 0.0001	+0.88	0.1545
LLaDA-o	+1.63	0.2529	+0.72	**0.0004**	+1.04	< 0.0001	+0.48	0.1058

A.10 Inference Configurations

Each model uses its released inference settings, listed in Table 12. Candidates differ only in sampling seed, set to $5000c + i$ for candidate c of prompt i. Candidate 0 therefore uses the same seed as the single sample baseline and is identical to it.

Table 12: Released inference settings used unchanged. Resolution and step count determine most of the candidate generation cost.

Model	Generation	Resolution	Steps	Guidance
BAGEL-7B-MoT	rectified flow	512×512	20	4.0
Janus-Pro-7B	autoregressive	384×384	N/A	5.0
Janus-Pro-1B	autoregressive	384×384	N/A	5.0
Show-o2-7B	flow matching	432×432	30	repository default
Show-o2-1.5B	flow matching	432×432	30	repository default
LaViDa-O	masked diffusion	1024×1024	24	4.0
LLaDA-o	continuous diffusion	1024×1024	50	4.0

A.11 Official Evaluators

Every image is scored using the benchmark's official evaluator implementation rather than a reimplementation. This matters because seemingly minor preprocessing and evaluation choices can materially affect absolute scores.

GenEval uses `evaluate_images.py` with the Mask2Former detector and CLIP ViT-L/14 colour and attribute classifiers, following the official object matching, counting, position rules, and thresholds. T2I-CompBench uses the metric assigned by its repository to each category. It uses BLIP-VQA at 480 pixel input for colour, shape, and texture, UniDet for spatial, 3D spatial, and numeracy, CLIPScore for nonspatial, and the 3-in-1 combination for complex. Using a 384 pixel rather than 480 pixel BLIP-VQA processor changes the score by 0.043 on the same images, larger than several gains reported here. DPG-Bench uses mPLUG-large with the official dependency zeroing rule, scored at the question level and averaged per image. GenAI-Bench uses VQAScore computed by CLIP-FlanT5-XXL, the metric reported by the benchmark. The image and prompt are placed in the template `Does this figure show {a}? Please answer yes or no.` and the score is the probability assigned to Yes.

This last evaluator itself uses a continuous readout of a Yes/No question, so its agreement with continuous verification is not independent evidence that the recovered ordering is correct. However, it uses a different model from every verifier under study and therefore cannot reward a candidate merely because the generator assigned it a high verification score. Appendix A.12 provides an independent human preference evaluation.

A.12 Human Preference Evaluation

To provide evidence independent of the GenAI-Bench evaluator, we conduct a human preference evaluation on 301 image pairs selected by continuous and discrete readout. Each pair is evaluated twice with the left and right positions exchanged to control for presentation order. Each judgment records whether the image selected by continuous readout is preferred, tied, or disfavored relative to the image selected by discrete readout.

Table 13: Human preference between continuous and discrete readout on 301 GenAI-Bench image pairs. Each pair is evaluated in both left and right orders. Win rate excludes ties.

Evaluation	Win	Tie	Loss	Win rate
All 602 judgments	252	200	150	62.7%
Position consistent	101	79	52	**66.0%**

Across all 602 judgments, the continuous readout selection is preferred 252 times and the discrete readout selection 150 times, with 200 ties, giving a 62.7% win rate among judgments without ties. More importantly, 232 pairs receive the same outcome after the left and right positions are exchanged. Within this position consistent subset, continuous readout wins 101 pairs, ties 79, and loses 52, corresponding to a 66.0% win rate among pairs without ties. A two sided binomial test on the 153 such pairs gives $p = 9.2 \times 10^{-5}$. These results provide human evidence that the recovered

ordering is useful and that the GenAI-Bench improvement is not explained solely by agreement with its continuous VQAScore evaluator.

A.13 Local Exceptions to the Readout Effect

Of the 98 model and category cells from GenEval and T2I-CompBench, seven score lower with continuous than discrete readout. On GenEval, these are single object for Show-o2-1.5B and LLaDA-o, colors for Janus-Pro-1B, and counting for Show-o2-1.5B. Two occur where Discrete BoN has already saturated at 98.75 and 100.00, leaving no room to improve. The other two are small categories. Counting contains 80 prompts and colors 94, so one flip changes the score by 1.25 and 1.06 points, respectively, and each cell loses exactly one. On T2I-CompBench, the negative cells are 2D spatial for BAGEL-7B-MoT, 3D spatial for Show-o2-7B, and numeracy for LLaDA-o.

At the whole benchmark level, all 28 model and benchmark combinations improve, with the smallest gain being $+0.46$ for LaViDa-O on T2I-CompBench. Uniform random selection also remains below continuous readout in every combination. These exceptions are consistent with the scope discussed in §3.4. Higher fidelity readout can recover ordering already present in the verifier, but it does not guarantee that every local preference is correct.