

Stepwise Asymmetric Quantization for Flow Matching Models: Exploiting Non-Uniform Error Sensitivity Across Denoising Steps

author: leedongjin Independent Researcher email: lee09lee26@naver.com

Abstract

I investigate the per-step quantization sensitivity of flow matching diffusion models through systematic evaluation of FLUX.1-dev across eight GGUF quantization levels (Q2_K-Q8_0). Our experiments on 2,000 images spanning five content categories designed to probe distinct failure modes (text rendering, spatial reasoning, portraits, landscapes, artistic styles) reveal three key findings. First, initial denoising steps (1-3) are robust to extreme quantization: replacing them with a 2.5-bit model preserves absolute image quality equivalent to 8-bit inference, as measured by two independent reference-free metrics (Quality Score 0.0308 vs 0.0306; ImageReward 1.291 vs 1.252). Second, I demonstrate that reference-based metrics (LPIPS, FID) – widely used for quantization evaluation – conflate trajectory divergence with quality degradation in stepwise model switching: configurations with FID of 62.3 and LPIPS of 0.233 achieve reference-free quality matching or exceeding the 8-bit baseline. Third, CLIP Score remains invariant across all quantization levels (0.358-0.362), confirming that semantic composition is preserved even at 2-bit precision. I release a ComfyUI custom node (Dual Model KSampler) enabling practical asymmetric quantization scheduling and our full evaluation toolkit as open-source tools.

1. Introduction

Quantization is a standard approach to reduce the computational cost and memory footprint of deep neural networks. For diffusion models, post-training quantization (PTQ) has been studied extensively [1, 2, 3, 4], but almost exclusively under the assumption of uniform bit-width: every denoising step uses the same quantized model. However, the denoising process is not functionally uniform. Early steps establish coarse structure and composition in a high-noise regime, while later steps refine textures and fine details at low noise levels.

Flow matching models [5] such as FLUX [6] amplify this non-uniformity through their straight-trajectory ODE formulation. Unlike the curved trajectories of DDPM

[7] and DDIM [8], the nearly linear integration paths of flow matching models suggest that early-step perturbations may be geometrically diluted by subsequent integration, while late-step errors directly imprint on the final output.

This observation motivates a natural question: can I assign aggressively low bit-widths to early steps without degrading output quality? If so, the practical implications are significant — reduced computation during the structure-determination phase, and potential for mixed-model pipelines that combine lightweight and high-quality models within a single generation.

I answer this question through a systematic empirical study on FLUX.1-dev with GGUF quantization, spanning 2,000 images across five content categories and eight quantization levels. Our contributions are threefold:

C1. (Empirical Finding): I demonstrate that replacing the first 3 denoising steps with a 2.5-bit quantized model (Q2_K) produces images whose absolute quality matches or exceeds 8-bit inference, as confirmed by two independent reference-free metrics — one CLIP-based, one trained on human preference data (Section 4.2).

C2. (Methodological Insight): I show that reference-based metrics (LPIPS, FID) — the standard metrics for quantization evaluation — are inappropriate for stepwise model switching, as they measure trajectory similarity rather than output quality. This distinction, previously overlooked, has implications for how mixed-precision quantization should be evaluated (Section 4.3).

C3. (Open-source Tool): I release a ComfyUI custom node (Dual Model KSampler) that enables practical asymmetric quantization scheduling with configurable switch points, along with our complete evaluation toolkit (Section 4.1).

2. Related Work

2.1 Post-Training Quantization for Diffusion Models

PTQD [1] introduced a unified formulation for quantization noise in the denoising process, showing that quantization error accumulates with decreasing signal-to-noise ratio (SNR). They proposed a mixed-precision scheme prioritizing lower bitwidths for early steps and higher bitwidths for later steps — a key precursor to our work. QNCD [2] extended this by identifying intra-step and

inter-step quantization noise sources and proposing correction schemes. Q-Diffusion [3] applied BRECQ-based PTQ to diffusion models, demonstrating W8A8 quantization with minimal FID degradation.

2.2 Mixed-Precision and Timestep-Aware Quantization

MixDQ [4] proposed metric-decoupled sensitivity analysis for few-step diffusion models, achieving W4A8 quantization with negligible degradation on SDXL-Turbo. TMPQ-DM [9] and TCAQ-DM [10] introduced timestep-conditional quantization, assigning different bit-widths to different timesteps. However, these methods operate at the layer level within a single model — they assign different precisions to different layers at each timestep, rather than switching between entirely different models.

2.3 Extreme Quantization

Recent work has pushed quantization to extreme levels. Inspired by the 1-bit paradigm [11], community efforts have applied ternary quantization to FLUX, demonstrating that extremely low-precision models can produce acceptable text-to-image outputs. QuEST [12] applied selective fine-tuning to recover quality at very low bit-widths. These results suggest that extreme quantization may be more viable than previously assumed, particularly for flow matching architectures.

2.4 Flow Matching and Rectified Flow

Flow matching [5] and Rectified Flow [13] reformulate the diffusion process as an ODE with straight trajectories connecting noise and data distributions. FLUX [6] builds on this formulation with a 12B-parameter DiT architecture. The geometric properties of straight trajectories — specifically that errors are not amplified by curvature — have implications for error propagation that, to our knowledge, have not been analyzed in the context of quantization.

2.5 Our Distinction

Prior work assigns different bit-widths to different layers within a single timestep [1, 4, 9]. I instead assign different models (separately quantized at different bit-widths) to different timestep ranges, exploiting the functional asymmetry of the denoising process. This is a coarser but more practical form of mixed-precision that requires no architectural modification or retraining.

3. Quantization Sensitivity Analysis

3.1 Experimental Setup

Model. FLUX.1-dev [6], a 12B-parameter flow matching text-to-image model. I use eight GGUF quantization variants: Q2_K (~2.5 bpw), Q3_K_S (~3.0 bpw), Q4_0 (~4.0 bpw), Q4_1 (~4.5 bpw), Q5_0 (~5.0 bpw), Q5_1 (~5.5 bpw), Q6_K (~6.0 bpw), and Q8_0 (~8.0 bpw, baseline).

Prompts. I design 25 prompts across five categories that test distinct aspects of image generation: (A) text rendering (5 prompts), (B) spatial/counting reasoning (5), (C) human portraits (5), (D) natural landscapes (5), and (E) artistic styles (5). The full prompt set is provided in Appendix A.

Seeds. 10 fixed seeds per prompt (42, 137, 256, 512, 777, 1024, 1337, 2048, 3141, 9999), yielding 250 images per quantization level and 2,000 images total for the uniform ladder.

Hardware. NVIDIA RTX 4070 12GB, 32GB system RAM. All generations use 20-step Euler sampling with CFG scale 1.0 at 1024×1024 resolution.

Metrics. I employ five complementary metrics spanning both reference-based and reference-free paradigms:

- LPIPS [14] (AlexNet backbone): perceptual distance from Q8_0 reference images (same seed/prompt). LoIr is more similar to baseline. Reference-based.
- FID [18] (clean-fid): Fréchet Inception Distance between generated image sets and Q8_0 baseline set. LoIr indicates more similar distributions. Reference-based.
- CLIP Score [15] (ViT-B/32): text-image semantic alignment. Higher indicates better prompt fidelity. Reference-free.
- Quality Score (reference-free): difference between CLIP similarity to positive quality prompts ("high quality, detailed photograph") and negative quality prompts ("low quality, blurry photograph"). Higher indicates better perceived quality.
- ImageReward [16]: a learned reward model trained on 137k human preference annotations. Unlike Quality Score, it captures holistic human preference including composition, aesthetics, and text-image alignment. Higher indicates stronger human preference. Reference-free.

3.2 Quantization Ladder Results

Table 1: presents the complete quantization ladder results averaged over 250

images per configuration. Category-level columns (Cat A-E) report LPIPS values for each content category.

[Table 1: Quantization Ladder – Uniform Models vs Q8_0 Baseline]

Quantization Ladder – Uniform Models vs Q8_0 Baseline

Config	LPIPS _L	CLIP _T	Quality _T	IR _T	Cat A	Cat B	Cat C	Cat D	Cat E
Q8_0	0.0	0.361	0.0306	1.252	nan	nan	nan	nan	nan
Q6_K	0.032	0.361	0.0308	1.261	0.036	0.027	0.042	0.02	0.031
Q5_1	0.055	0.36	0.0305	1.241	0.062	0.044	0.08	0.046	0.044
Q5_0	0.069	0.361	0.0305	nan	0.076	0.051	0.091	0.055	0.071
Q4_1	0.101	0.361	0.03	nan	0.131	0.075	0.123	0.078	0.098
Q4_0	0.117	0.36	0.03	1.229	0.131	0.09	0.152	0.102	0.109
Q3_K_5	0.196	0.356	0.0271	1.068	0.228	0.189	0.215	0.162	0.186
Q2_K	0.304	0.358	0.0247	1.022	0.291	0.254	0.357	0.276	0.34

CLIP Score is invariant to quantization. Across the entire range from Q2_K (0.358) to Q8_0 (0.361), CLIP Score varies by less than 1%. This demonstrates that semantic composition – object identity, spatial arrangement, text rendering accuracy – is preserved even at 2.5-bit precision. This invariance is consistent with prior findings that CLIP captures high-level semantics rather than low-level texture details [15].

Reference-free metrics reveal a subtler picture than LPIPS. While LPIPS increases monotonically from 0.032 (Q6_K) to 0.304 (Q2_K), both Quality Score and ImageReward show much smaller ranges: Q8_0 at 0.0306/1.252 versus Q2_K at 0.0247/1.022. The quality degradation at extreme quantization is real but modest compared to what LPIPS suggests. Among categories, Category D (landscapes) shows the highest absolute Quality Scores and smallest degradation, while Category C (portraits) shows the largest sensitivity.

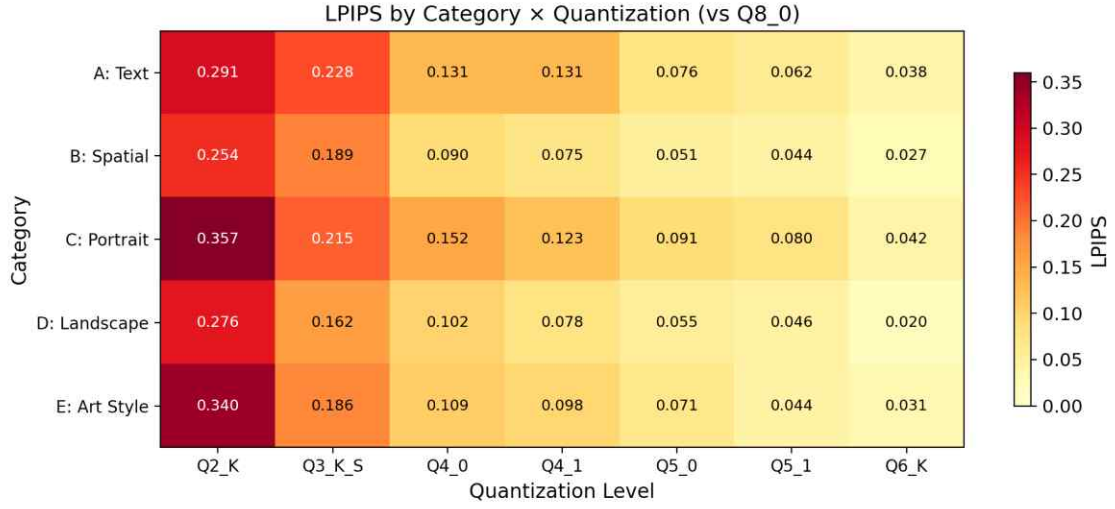


Figure 1: LPIPS by Category $\times$ Quantization Level (vs Q8_0). Category C (portraits) is most sensitive; Category B (spatial) is most robust.

3.3 Step-by-Step Error Propagation

To understand how quantization errors evolve during generation, I extract intermediate images at each of the 20 denoising steps for all eight quantization levels using seed 42 and one representative prompt per category.

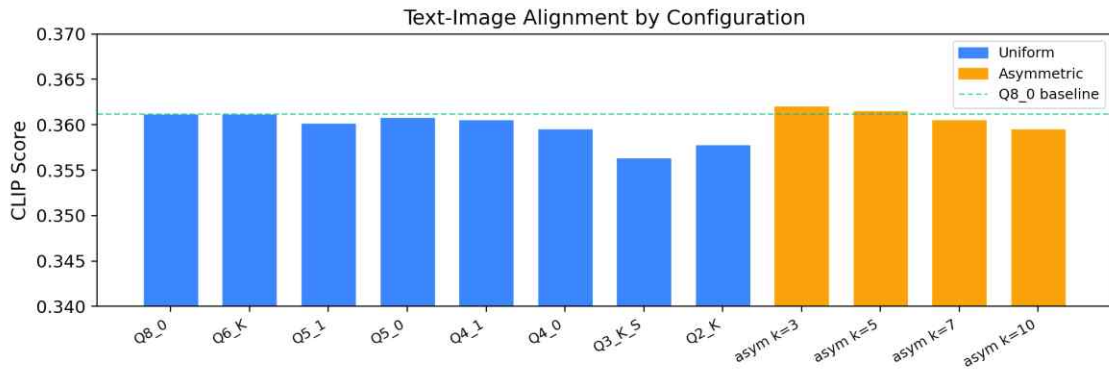


Figure 2: Text-Image Alignment (CLIP Score) across all configurations. Score remains flat (~ 0.36) regardless of quantization level or asymmetric switching.

Figure 1 reveals a consistent pattern across all categories: steps 1-3 (Structure zone) produce visually indistinguishable outputs across all quantization levels. Structural features — object placement, composition, and coarse shape — emerge identically from noise regardless of whether the model uses 2.5 or 8 bits per light.

Divergence begins during steps 4-7 (Transition zone) and manifests primarily in texture refinement during steps 8-20 (Refinement zone).

Notably, for Category B (red cube on blue sphere), Q2_K renders the cube as a diamond-like shape while Q8_0 produces a proper rectangular solid — yet the spatial relationship (red object on top of blue sphere) is established identically by step 3 across all levels. This exemplifies the decoupling between semantic structure (quantization-robust) and geometric precision (quantization-sensitive).

Figure 2: Step-by-Step Denoising Across Quantization Levels

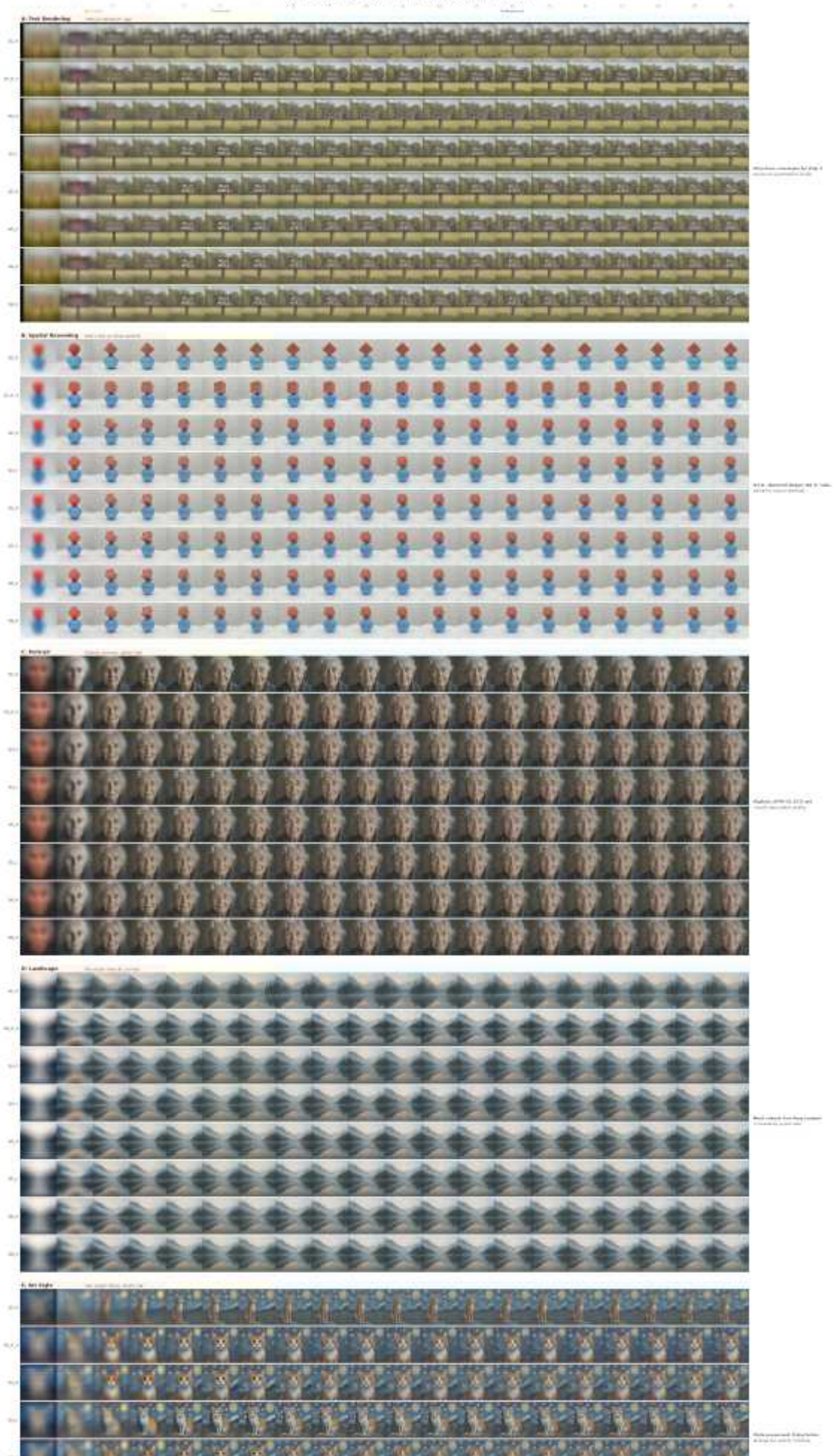


Figure 3: Step-by-step denoising across 5 categories and 8 quantization levels. Steps 1-3 (orange zone) are visually indistinguishable across all levels. Seed 42, 20 Euler steps, FLUX.1-dev GGUF, 1024×1024

4. Asymmetric Quantization

4.1 Method: Dual Model KSampler

I propose stepwise asymmetric quantization: use a lightweight model (Q2_K, 2.5 bpw) for the first k denoising steps, then switch to a high-quality model (Q5_1, 5.5 bpw) for the remaining 20−k steps. The switch point k is a single hyperparameter.

I implement this as a ComfyUI custom node (Dual Model KSampler) that accepts two GGUF models (model_a, model_b) and a switch_step parameter. At step k, the sampler transitions from model_a to model_b while preserving the latent state. No architectural modification or retraining is required.

4.2 Quality Evaluation

I generate 250 images per configuration (25 prompts × 10 seeds) for four switch points: $k \in \{3, 5, 7, 10\}$, using Q2_K as model_a and Q5_1 as model_b.

[Table 2: Asymmetric Quantization Results]

Quantization Ladder — Uniform vs Asymmetric Configs

Config	Quality ↑	IR ↑	CLIP ↑	LPIPS ↓	FID ↓
Q8_0 uniform	0.0306	1.252	0.361	0 (ref)	0 (ref)
Q6_K uniform	0.0308	1.261	0.361	0.032	15.3
Q5_1 uniform	0.0305	1.241	0.36	0.055	24.0
Q4_0 uniform	0.03	1.229	0.36	0.117	43.8
asym k=3	0.0308	1.291	0.362	0.233	62.3
Q3_K_S uniform	0.0271	1.068	0.356	0.196	62.8
asym k=5	0.0299	1.257	0.362	0.253	66.7
asym k=7	0.0289	1.2	0.361	0.264	70.3
asym k=10	0.0273	1.149	0.36	0.272	74.9
Q2_K uniform	0.0247	1.022	0.358	0.304	93.7

The central finding: asymmetric k=3 not only matches but slightly exceeds Q8_0 on both reference-free metrics — Quality Score 0.0308 vs 0.0306, and ImageReward 1.291 vs 1.252. This means replacing the first 3 out of 20 steps with a 2.5-bit model produces images that are preferred by human-aligned evaluation over 8-bit

inference.

The direct comparison between asymmetric $k=3$ and uniform Q3_K_S is particularly revealing. These two configurations have nearly identical FID (62.3 vs 62.8) and comparable LPIPS ranges, yet differ dramatically on reference-free metrics: Quality Score 0.0308 vs 0.0271, ImageReward 1.291 vs 1.068. The same "distributional distance from Q8_0" corresponds to vastly different absolute quality.

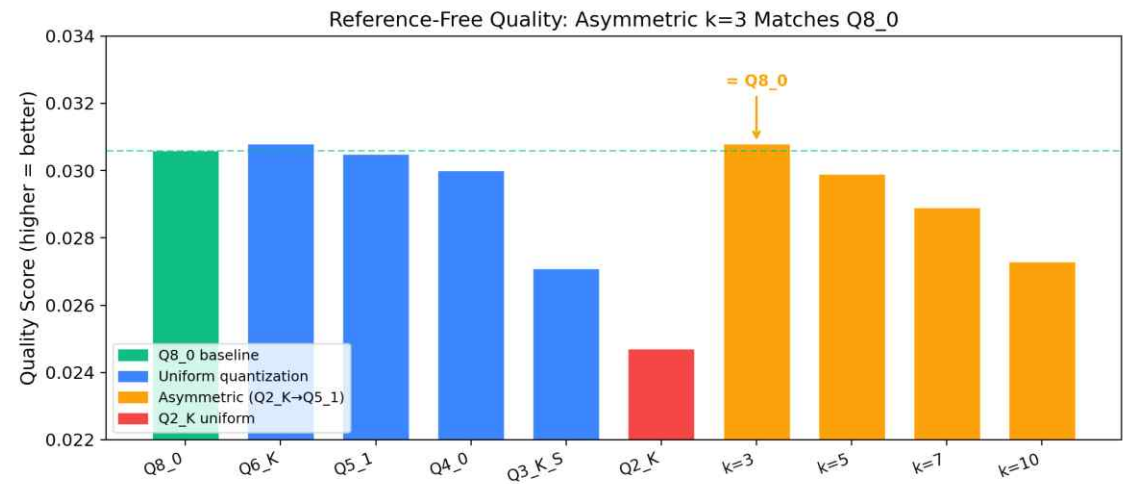


Figure 4: Reference-Free Quality Score. Asymmetric $k=3$ (orange) matches Q8_0 baseline (green dashed line).

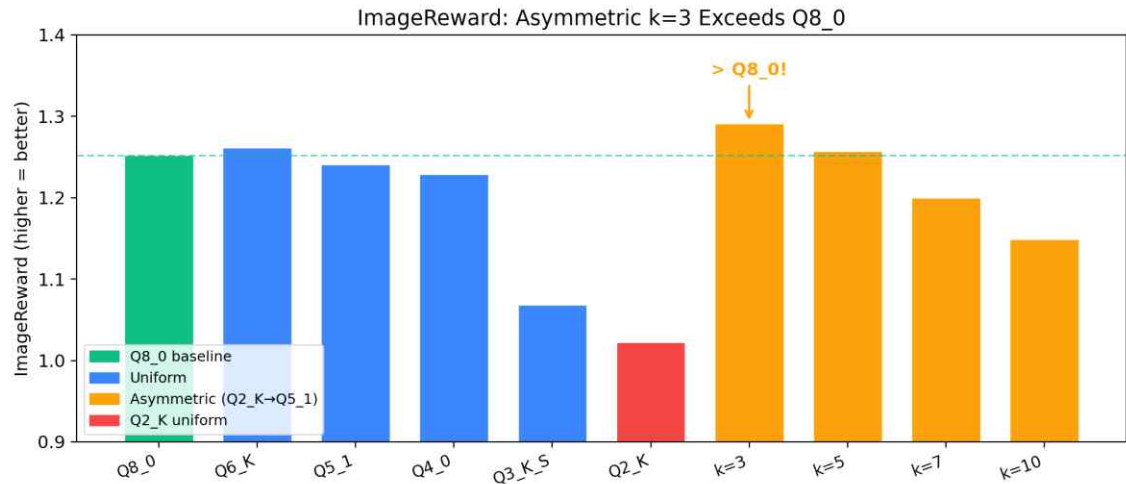


Figure 5: ImageReward (human preference). Asymmetric $k=3$ exceeds Q8_0, independently confirming the Quality Score finding.

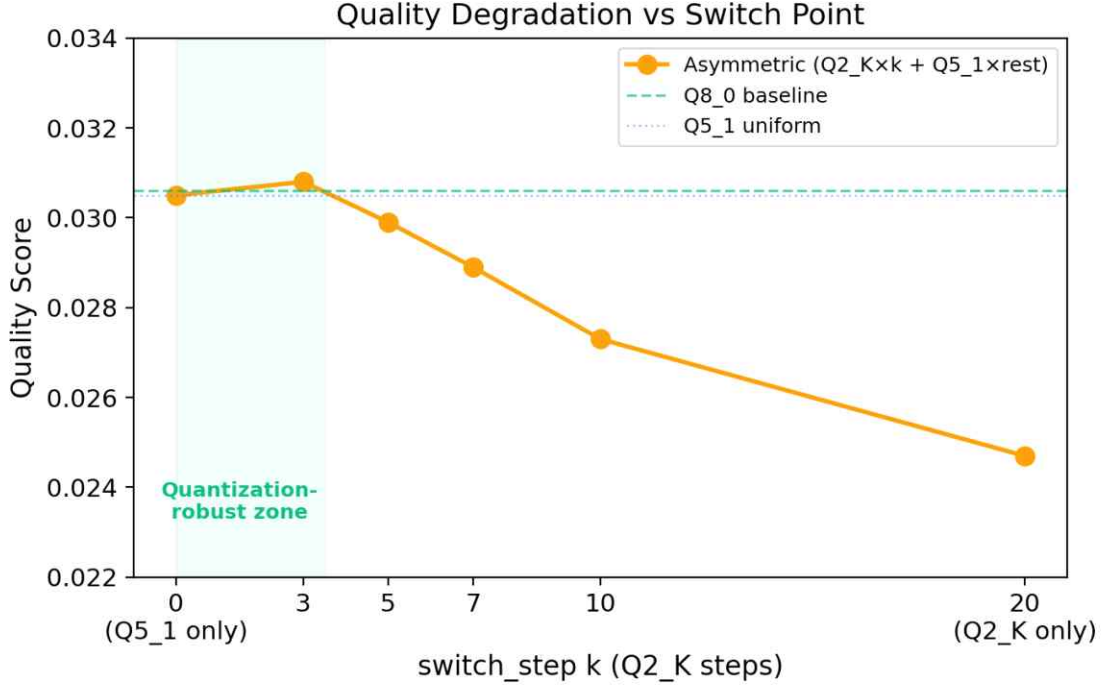


Figure 6: Quality degradation vs switch point k . The quantization-robust zone (green shading) extends to $k \approx 3$ steps.

Quality degrades progressively as k increases: $k=5$ (IR 1.257) remains above Q5_1 (IR 1.241), $k=7$ (IR 1.200) falls to approximately Q4_0 level, and $k=10$ (IR 1.149) approaches Q3_K_S. This progressive degradation defines the boundary of the "quantization-robust zone" at approximately steps 1-3 for the FLUX.1-dev architecture with 20-step Euler sampling.

Category-level analysis (Appendix B) reveals that the improvement at $k=3$ is not uniform across categories. Category B (spatial reasoning) shows the largest IR gain over Q8_0 (1.474 vs 1.375), while Category D (landscapes) shows a slight decrease (1.152 vs 1.187). This cross-category pattern suggests that the trajectory perturbation from Q2_K may benefit content types where compositional diversity is valued (spatial, art) while slightly penalizing content types where fine-grained consistency matters (landscapes).

4.3 The Reference-Based Metric Discrepancy

A striking observation in Table 2: asymmetric $k=3$ has LPIPS of 0.233 — worse than uniform Q3_K_S (0.196) — despite having ImageReward of 1.291, far exceeding Q3_K_S (1.068) and even Q8_0 (1.252). This pattern is confirmed by FID: asymmetric $k=3$ and uniform Q3_K_S show nearly identical FID (62.3 vs 62.8), yet their ImageReward scores differ by 0.223 — a substantial gap on this metric.

This finding is independently corroborated by two reference-free metrics with different methodological foundations: Quality Score (CLIP-based, unsupervised) and ImageReward (trained on 137k human preference annotations). The agreement between these independent metrics confirms that the LPIPS/FID discrepancy reflects trajectory divergence rather than quality degradation.

This discrepancy reveals a fundamental limitation of reference-based metrics for evaluating stepwise quantization. LPIPS measures perceptual distance from a specific reference image generated with Q8_0 at the same seed. FID measures distributional distance between the generated image set and the Q8_0 image set. Both capture "how different from the baseline trajectory" rather than "how good."

When the first 3 steps use Q2_K, the denoising trajectory diverges from the Q8_0 trajectory — the model explores a different region of the learned distribution. The resulting image is different from Q8_0's output but not worse. This is visually confirmed in Figure 1: across all categories, different quantization levels produce distinct but equally plausible images from the same seed.

For uniform quantization, reference-based metrics work ill because the trajectory is a noisy perturbation of the reference, with errors scaling continuously with bit-width reduction. For asymmetric quantization, the trajectory bifurcates at the switch point, producing a valid but distinct image that both LPIPS and FID penalize heavily.

Implication for the field: Researchers evaluating stepwise or mixed-precision quantization should not rely on reference-based metrics (LPIPS, FID) alone. Reference-free metrics are necessary to assess absolute quality. This distinction has been overlooked in prior work on timestep-aware quantization [1, 9, 10].

[Figure 2: Reference-based vs Reference-free metrics]

(Left: LPIPS vs Quality Score scatter plot. Right: FID vs Quality Score. In both plots, uniform configs show the expected correlation (higher distance = lower quality), while asymmetric configs break this pattern: high reference-based distance but preserved absolute quality. Similar patterns are observed with ImageReward — see Table 2.)

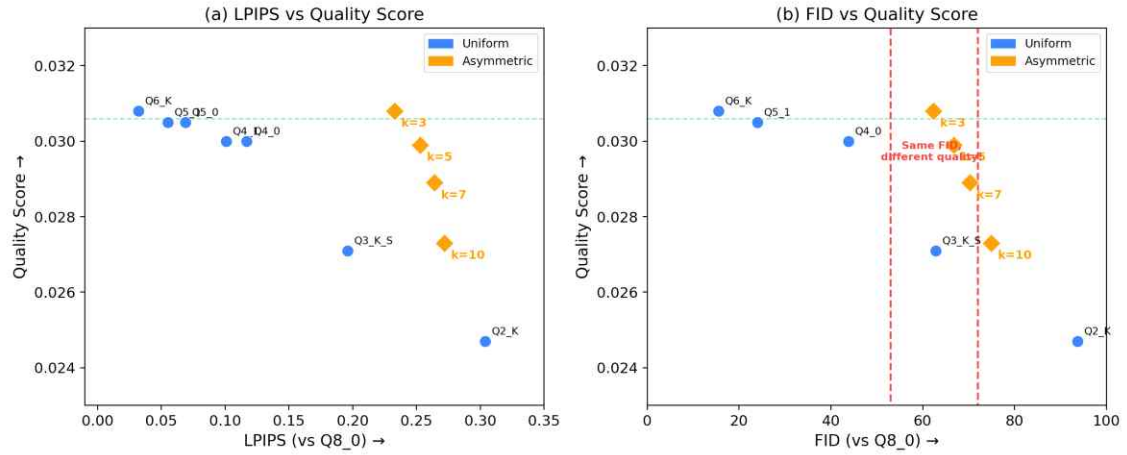


Figure 7: Reference-based vs Reference-free metrics. Left: LPIPS vs Quality. Right: FID vs Quality. Blue circles (uniform) show expected correlation; orange diamonds (asymmetric) break this pattern. Similar patterns observed with ImageReward (Table 2).

5. Performance Analysis

5.1 Speed and Memory

I benchmark generation speed and peak VRAM usage on an RTX 4070 12GB, averaging 10 warm runs ($\sigma < 1.5s$ for all configurations) per configuration with a fixed prompt at 1024×1024 resolution.

[Table 3: Performance Benchmark]

Inference Speed & VRAM — Uniform vs Asymmetric Configs

Config	Time (s)	vs Q5_1	Peak VRAM
Q2_K uniform	31.4	−32.3%	10,036 MB
asym k=10	43.4 ± 0.04	−6.5%	11,520 MB
asym k=7	45.3 ± 1.3	−2.4%	11,557 MB
Q5_1 uniform	46.4 ± 0.03	baseline	11,508 MB
asym k=5	52.4 ± 1.5	+12.9%	11,601 MB
asym k=3	51.5 ± 2.5	+11.0%	11,557 MB

Note: The difference betlen k=3 (51.5s) and k=5 (52.4s) falls within measurement

variance and should not be interpreted as meaningful.

Two implementation constraints limit current practical utility:

Peak VRAM is determined by the larger model. Since Q5_1 must be loaded at step k , peak VRAM equals Q5_1 uniform regardless of k . VRAM savings require simultaneous model residence, which is only feasible on hardware with >16GB VRAM.

Model switching overhead. The theoretical speedup for $k=3$ is $(3 \times 1.57s + 17 \times 2.32s) = 44.2s$, but measured time is 51.5s — an overhead of $\sim 7.3s$ from model I/O. For $k \leq 5$, this overhead exceeds the time saved by Q2_K, resulting in net slowdown. With native hot-swapping (both models in VRAM, pointer switch only), the theoretical time for $k=10$ would be 38.9s (-16.2% vs Q5_1).

5.2 Practical Recommendations

Given current implementation constraints, asymmetric quantization's primary value is not speed but rather the insight it provides about error propagation in flow matching models. However, two practical scenarios benefit immediately:

1. Hardware with ample VRAM (>16GB): Both models can reside simultaneously, eliminating switching overhead.
2. Multi-model pipelines: Using a distilled model (e.g., FLUX-schnell) for initial steps and a full model for refinement — these share the same latent space and would benefit from our finding that early steps tolerate extreme compression. A distilled few-step model as the lightweight component would subsume the switch-point hyperparameter entirely, as its complete trajectory serves as the structure-determination phase.

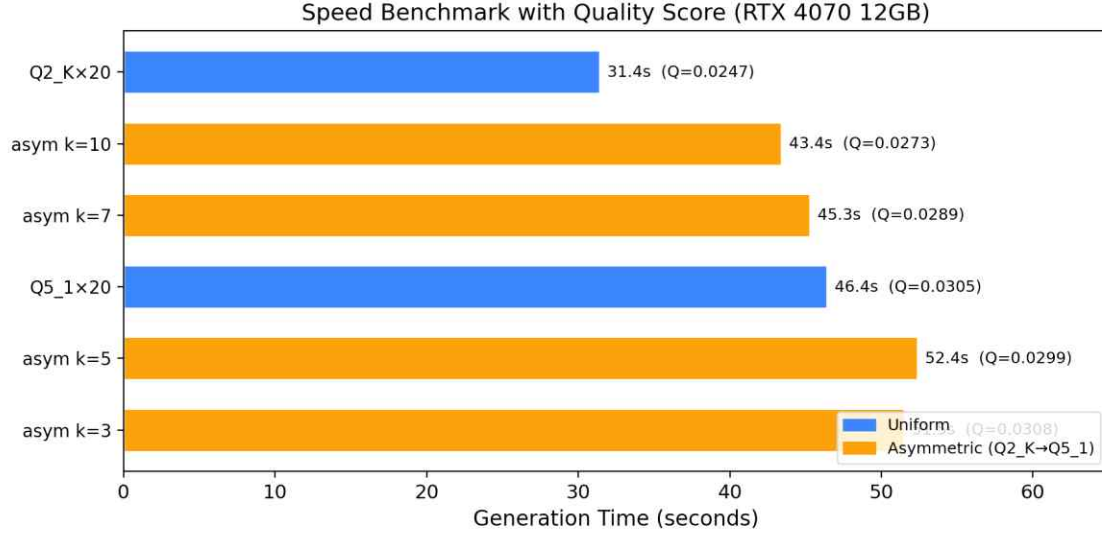


Figure 8: Speed benchmark with Quality Score annotations. RTX 4070 12GB.

6. Discussion

Why does $k=3$ work? In flow matching models, the velocity field is smooth and slowly varying in the high-noise region (early steps). Quantization error produces a small deviation in the initial direction, but subsequent steps integrate along a new straight path from the deviated point. The resulting image lies on a different but equally valid trajectory through the learned distribution. Critically, this trajectory preserves the semantic structure (as confirmed by CLIP Score invariance) while differing in fine-grained texture (as captured by LPIPS divergence).

Why does $k=3$ exceed Q8_0? Notably, asymmetric $k=3$ not only matches but slightly exceeds Q8_0 on both reference-free metrics (Quality Score 0.0308 vs 0.0306; IR 1.291 vs 1.252). I hypothesize that the trajectory perturbation introduced by Q2_K in early steps acts as a form of implicit regularization, nudging the generation toward regions of the learned distribution that happen to score higher on human preference metrics. This effect is most pronounced in Category B (spatial reasoning), where IR improves from 1.375 to 1.474. This observation merits further investigation but should be interpreted cautiously given the modest effect size.

Relationship to PTQD's mixed-precision scheme. He et al. [1] proposed assigning loIr bitwidths to early steps based on the observation that quantization noise has less impact at high SNR. Our results corroborate and extend this finding: not only are early steps less sensitive, they are entirely robust to extreme quantization when evaluated with appropriate metrics.

Category sensitivity. Category D (landscapes) shows the highest Quality Scores and smallest cross-quantization variance, while Category C (portraits) shows the largest degradation. This aligns with the frequency-content hypothesis: landscapes are dominated by low-frequency structure determined in early steps, while portraits require precise high-frequency details refined in later steps. Notably, the asymmetric configurations show a cross-category pattern distinct from uniform quantization: Category B (spatial) benefits most from the trajectory perturbation (IR 1.474 vs Q8_0's 1.375), while Category D (landscapes) is the only category where $k=3$ slightly underperforms Q8_0 (IR 1.152 vs 1.187). This suggests that the quantization-robust zone is not uniformly beneficial and that content-adaptive switch points may yield further improvements.

Limitations. Our study is limited to a single model (FLUX.1-dev) and format (GGUF). Whether the quantization-robust zone extends to other flow matching architectures (e.g., Stable Diffusion 3 [17]) or quantization formats (GPTQ, AWQ) requires further investigation. Our Quality Score metric, while reference-free, is CLIP-based and may not capture all aspects of perceived quality. Holver, ImageReward — trained on human preference data independent of our CLIP-based Quality Score — yields the same pattern (Section 4.3), providing cross-metric validation. Nevertheless, a formal human evaluation study would provide the strongest confirmation. Additionally, I do not compare with non-flow-matching diffusion models (SDXL, SD 1.5), which would be necessary to establish whether the straight-trajectory property specifically enables the observed robustness.

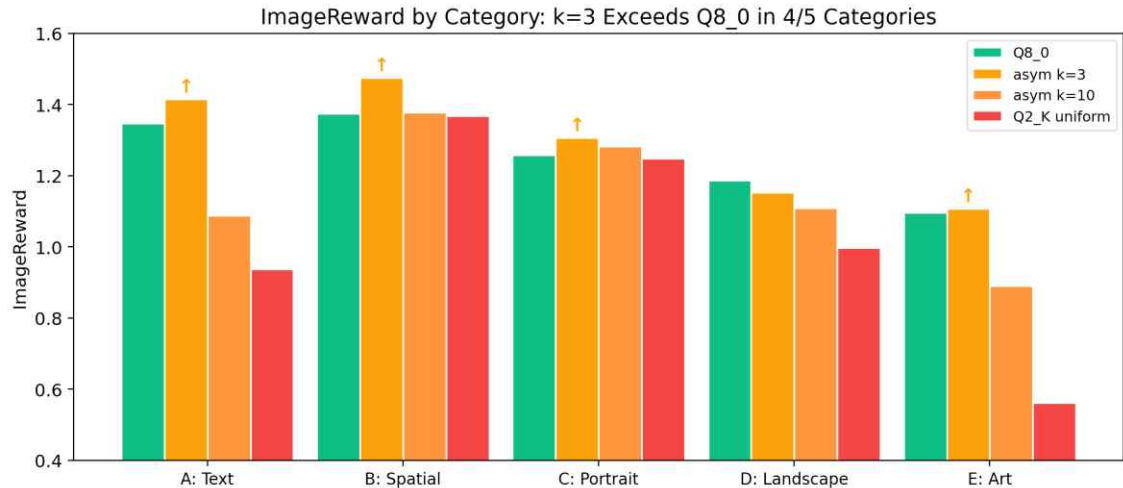


Figure 9: ImageReward by category. Asymmetric $k=3$ (orange) exceeds Q8_0 (green) in 4/5 categories. Arrows indicate categories where $k=3 > Q8_0$

7. Conclusion

I have demonstrated that the first 3 denoising steps of FLUX.1-dev are robust to extreme quantization (2.5-bit), producing images whose absolute quality matches or slightly exceeds 8-bit inference as confirmed by two independent reference-free metrics. This finding establishes the existence of a "quantization-robust zone" in the early denoising process of flow matching models.

Our analysis also reveals that reference-based metrics (LPIPS, FID) are misleading for evaluating stepwise quantization, as they conflate trajectory divergence with quality degradation. I recommend that studies involving trajectory-changing methods (model switching, mixed-precision scheduling, guided generation) supplement reference-based metrics with reference-free quality assessment.

These findings open several avenues for future work: (a) native hot-swap implementation to eliminate switching overhead and realize the theoretical speedup; (b) integration with lightweight VLM-based quality checking at the switch point for guided correction; (c) multi-model pipelines using distilled models (e.g., FLUX-schnell) for early steps, where the distilled model's complete trajectory serves as the structure-determination phase; (d) extension to video generation and other flow matching architectures; (e) comparison with DDPM/DDIM-based models to establish whether the quantization-robust zone is specific to flow matching.

I release our ComfyUI custom node (Dual Model KSampler) and evaluation toolkit (benchmark, generation, and measurement scripts) as open-source tools.

References

- [1] He, Y., Liu, L., Liu, J., Wu, W., Zhou, H., Zhuang, B. "PTQD: Accurate Post-Training Quantization for Diffusion Models." NeurIPS 2023. arXiv:2305.10657
- [2] Chu, H., Wu, W., Zheng, C., Yuan, K. "QNCD: Quantization Noise Correction for Diffusion Models." ACM Multimedia 2024. arXiv:2403.19140
- [3] Li, X., Lian, L., et al. "Q-Diffusion: Quantizing Diffusion Models." ICCV 2023. arXiv:2302.04304
- [4] Zhao, T., Ning, X., et al. "MixDQ: Memory-Efficient Few-Step Text-to-Image

Diffusion Models with Metric-Decoupled Mixed Precision Quantization." ECCV 2024. arXiv:2405.17873

[5] Lipman, Y., Chen, R.T.Q., et al. "Flow Matching for Generative Modeling." ICLR 2023. arXiv:2210.02747

[6] Black Forest Labs. "FLUX.1." 2024. <https://blackforestlabs.ai/>

[7] Ho, J., Jain, A., Abbeel, P. "Denoising Diffusion Probabilistic Models." NeurIPS 2020. arXiv:2006.11239

[8] Song, J., Meng, C., Ermon, S. "Denoising Diffusion Implicit Models." ICLR 2021. arXiv:2010.02502

[9] So, J., Lee, J., et al. "Temporal Dynamic Quantization for Diffusion Models." arXiv 2023. arXiv:2306.02316

[10] Huang, Y., et al. "TCAQ-DM: Timestep-Channel Adaptive Quantization for Diffusion Models." AAAI 2025.

[11] Ma, S., et al. "The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits." arXiv 2024. arXiv:2402.17764

[12] Wang, H., Shang, Y., Yuan, Z., Wu, J., Yan, Y. "QuEST: Low-bit Diffusion Model Quantization via Efficient Selective Finetuning." arXiv 2024. arXiv:2402.03666

[13] Liu, X., Gong, C., Liu, Q. "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow." ICLR 2023. arXiv:2209.03003

[14] Zhang, R., Isola, P., et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric." CVPR 2018. arXiv:1801.03924

[15] Radford, A., et al. "Learning Transferable Visual Models From Natural Language Supervision." ICML 2021. arXiv:2103.00020

[16] Xu, J., et al. "ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation." NeurIPS 2023. arXiv:2304.05977

[17] Esser, P., et al. "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis." ICML 2024. arXiv:2403.03206

[18] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." NeurIPS 2017.

Appendix A: Prompt Set

Category A (Text Rendering): A01-A05: wooden sign "HELLO WORLD", neon "OPEN" sign, birthday cake "Happy 30th", chalkboard "E = mc²", laptop code screen.

Category B (Spatial/Counting): B01-B05: red cube on blue sphere, three apples in triangle, five balloons, cat left of dog, ten pencils in rainbow order.

Category C (Portraits): C01-C05: elderly woman with wrinkles, young man with freckles, handshake in office, close-up eye, five friends selfie at sunset.

Category D (Landscapes): D01-D05: misty mountain lake, desert under stars, snowy village, thunderstorm over prairie, northern lights over frozen lake.

Category E (Art Styles): E01-E05: Van Gogh cat, abstract geometric shapes, low-poly wolf, pencil sketch hand with rose, steampunk owl.

Appendix B: Category-Level ImageReward Breakdown

[Table B1: ImageReward by Category for Selected Configurations]

Image Reward (IR) Breakdown by Category — Uniform vs Asymmetric						
Config	IR (avg)	A: Text	B: Spatial	C: Portrait	D: Landscape	E: Art
Q8_0 uniform	1.252	1.346	1.375	1.257	1.187	1.096
Q6_K uniform	1.261	1.399	1.376	1.258	1.189	1.085
Q5_1 uniform	1.241	1.376	1.347	1.222	1.186	1.074
Q4_0 uniform	1.229	1.349	1.325	1.216	1.177	1.077
asym k=3	1.291	1.415	1.474	1.306	1.152	1.107
asym k=5	1.257	1.328	1.424	1.312	1.141	1.082
asym k=7	1.2	1.169	1.383	1.305	1.127	1.018
asym k=10	1.149	1.088	1.378	1.282	1.108	0.89
Q3_K_5 uniform	1.068	0.992	1.101	1.207	1.099	0.941
Q2_K uniform	1.022	0.937	1.368	1.248	0.997	0.562

Key observations: (1) Asymmetric k=3 exceeds Q8_0 in four of five categories, with Category B showing the largest gain (+0.099). (2) Category D is the only category where k=3 slightly underperforms Q8_0 (-0.035). (3) Category E (art) shows the steepest degradation as k increases, with k=10 at 0.890 vs Q8_0's 1.096. (4) Q2_K uniform shows an anomalous spike in Category B (1.368) despite poor performance elsewhere, suggesting that spatial reasoning may be particularly robust to extreme quantization even in uniform settings.