UniQ: Modality-Aware Post-Training Quantization for Unified Multimodal Models

Anonymous authors
Paper under double-blind review

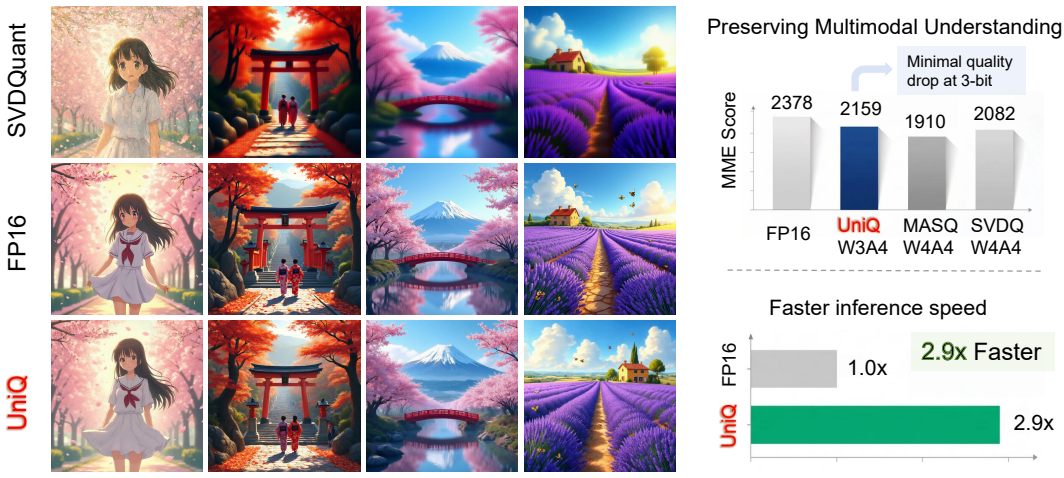

Figure 1: Teaser of UniQ on BAGEL-7B-MoT. UniQ preserves text-to-image quality under W3A4, maintains strong multimodal understanding, and achieves a 2.9× end-to-end speedup over FP16.

Abstract

Unified Multimodal Models (UMMs) integrate visual understanding and generation within a single Transformer backbone, but their large parameter count creates a substantial deployment bottleneck. Post-training quantization (PTQ) is a natural solution to this problem. However, directly applying quantization methods developed for language models, vision-language models, or diffusion models proves suboptimal because these methods do not match the token routing and dual-task structure of UMMs. We find that in UMMs token types and task pathways exhibit distinct statistical behavior, making calibration, precision allocation, and activation quantization unreliable. To address these challenges, we present UniQ, a modality-aware PTQ framework for unified multimodal models. UniQ introduces three innovative designs: dual-modality calibration, pathway-adaptive mixed precision, and token-aware sparse channel rotation. Experiments on BAGEL show that UniQ achieves W3A4 quantization with negligible quality loss and a 2.9× speedup over full precision. Our code and model will be available to the public.

1 Introduction

Unified multimodal models are emerging as an important foundation model paradigm because they integrate visual understanding and visual generation within a single Transformer backbone. Instead of assigning different modalities to separate encoders or task-specific modules, recent models such as BAGEL, Chameleon, Emu3, and Janus process text tokens, visual features, and generation latents in one shared sequence modeling framework Deng et al. (2025); Team (2024); Wang et al. (2024); Wu et al. (2024a); Chen et al. (2025b). This unified architecture enables a single model to answer visual questions, reason over images, and generate visual content. However, the same design also introduces substantial deployment cost. For example, BAGEL contains 14B parameters in total and

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

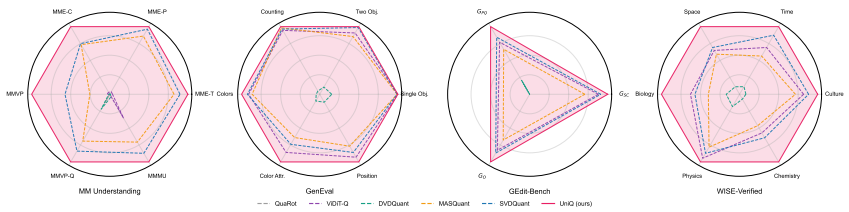

Figure 2: W4A4 radar comparison across multimodal understanding, generation, and editing benchmarks. UniQ preserves a more balanced capability profile than prior quantization baselines.

activates about 7B parameters for each task, resulting in a memory footprint and inference cost that are difficult to support on resource-constrained hardware Deng et al. (2025).

Post-training quantization reduces this cost by compressing a pretrained model without expensive retraining Frantar et al. (2022); Lin et al. (2024); Xiao et al. (2023); Shao et al. (2023). Existing quantization methods, however, do not transfer effectively to unified multimodal models. Methods designed for large language models often assume text-only activations, while methods designed for vision or diffusion models often assume homogeneous visual patch or latent features Li et al. (2023); Shang et al. (2023); Wu et al. (2024b); Chen et al. (2025a); Li* et al. (2025). Unified multimodal models violate both assumptions. In our experiments, uniform W3A4 quantization reduces the BAGEL MME score from 2378 under fp16 to 227, which corresponds to a 62.5% loss in visual understanding performance. This degradation indicates that low-bit deployment of unified multimodal models requires a quantization strategy aligned with their internal modality structure.

Our analysis attributes quantization failure to token-level and pathway-level heterogeneity. First, text, ViT, and VAE latent tokens induce different activation directions and Hessian spectra even when they are processed by the same backbone. This makes single-mode calibration biased because it overemphasizes the directions of one task while under-representing another. Second, the sensitivity order of layers changes across the base and generation pathways. A layer that tolerates low precision during understanding may remain important for generation, so a single global layer ranking is insufficient. Third, activation outliers are sparse and token-dependent, echoing prior observations that multimodal quantization requires modality-aware treatment Li et al. (2025); Hu et al. (2026); Qin et al. (2026). Text tokens often concentrate energy in a few channels, while ViT and VAE latent tokens have smoother channel profiles. Full-block activation rotations can therefore degrade per-group quantization by spreading concentrated outlier energy across group boundaries.

To address these challenges, we propose UniQ, a modality-aware post-training quantization framework for unified multimodal models. UniQ first uses Dual-Modality Hessian Calibration to collect calibration statistics from both understanding and generation data, so each module receives optimization signals from the tokens that actually pass through it. UniQ then applies Pathway-Adaptive Mixed-Precision Allocation, which profiles layer-pathway sensitivity and assigns weight bit-widths under a target compression budget. Finally, UniQ introduces Token-Aware Sparse Channel Rotation for activation quantization. It applies sparse Givens rotations within each quantization group and uses token-specific rotation parameters, which reduces outlier-induced error while avoiding cross-group energy leakage.

Our contributions are summarized as follows:

- **Token- and Pathway-Aware Quantization Perspective:** We identify token-level heterogeneity and pathway-level heterogeneity as the central obstacles to post-training quantization for unified multimodal models, indicating that language-model or vision-model quantization assumptions do not directly apply to unified understanding and generation.
- **Dual-Modality Hessian Calibration:** We propose a calibration strategy that collects activation statistics from both understanding and generation data, allowing Hessian-based weight optimization to match the actual operating distribution of each module.
- **Pathway-Adaptive Mixed-Precision Allocation:** We introduce an independent sensitivity profiling scheme for layers and task pathways, then assign weight bit-widths under a target compression budget.
- **Token-Aware Sparse Channel Rotation:** We design a group-internal sparse Givens rotation method with token-specific parameters, which reduces activation outlier error while avoiding

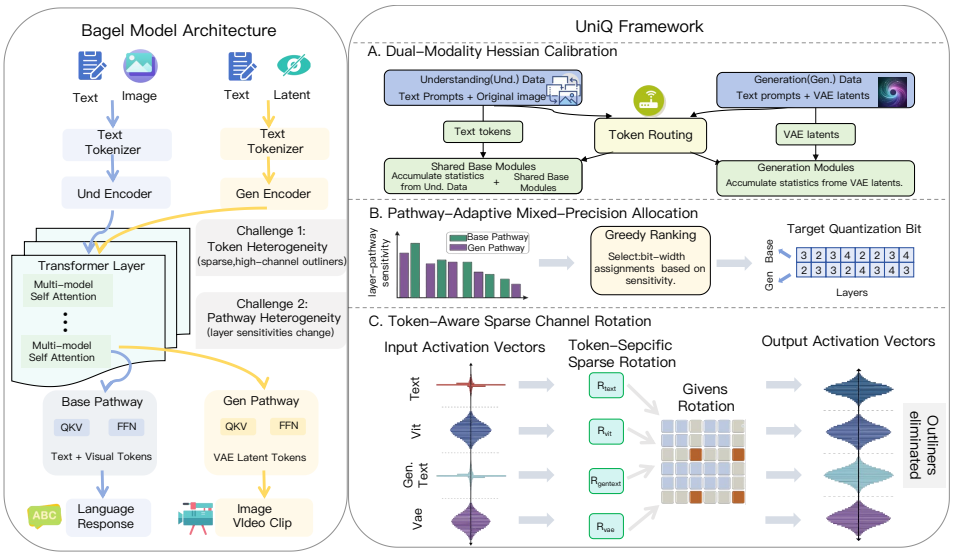

Figure 3: Overview of UniQ. BAGEL routes different token types through base and generation pathways, and UniQ matches this structure with dual-modality calibration, pathway-adaptive mixed precision, and token-aware sparse Givens rotation.

cross-group energy leakage. Experiments on BAGEL demonstrate that UniQ outperforms uniform low-bit quantization with negligible storage and computation overhead.

2 Related Work

2.1 Unified Multimodal Models

Unified multimodal models handle understanding and generation in one model. Chameleon Team (2024) and Emu3 Wang et al. (2024) tokenize images, videos, and text into unified discrete sequences, while Janus Wu et al. (2024a) and Janus-Pro Chen et al. (2025b) decouple visual encoding for understanding and generation within a shared autoregressive backbone. BAGEL Deng et al. (2025) further introduces a Mixture-of-Transformer-Experts architecture for visual understanding, text-to-image generation, and image editing. These works focus mainly on model design and scaling; we instead study post-training quantization, where text tokens, visual understanding tokens, and generation latents have different statistics inside the same backbone.

2.2 Post-Training Quantization

Post-training quantization Li et al. (2024); Lin et al. (2024); Zhang et al. (2025); Yan et al. (2025); Yang et al. (2025) reduces inference cost without retraining. For LLMs, GPTQ Frantar et al. (2022), AWQ Lin et al. (2024), and SmoothQuant Xiao et al. (2023) improve weight or weight-activation quantization, while OmniQuant Shao et al. (2023), QuaRot Ashkboos et al. (2024), and QuIP Chee et al. (2023) optimize parameters or rotate feature spaces. For multimodal language models, MBQ Li et al. (2025) and VEQ Qin et al. (2026) consider modality sensitivity and expert heterogeneity, but still focus on text-dominant or understanding-centric models.

PTQ has also been extended to diffusion models Wu et al. (2025a;b); Li et al. (2023); Chen et al. (2025a): Q-Diffusion Li et al. (2023) and PTQ4DM Shang et al. (2023) study U-Net diffusion, PTQ4DiT Wu et al. (2024b), Q-DiT Chen et al. (2025a), and ViDiT-Q Zhao et al. (2024) target Diffusion Transformers, and SVDQuant Li* et al. (2025), DVD-Quant Li et al. (2026), and Quant-Cache Wu et al. (2025a) handle outliers, temporal redundancy, or adaptive precision. These methods assume either understanding or generation-only pipelines; UniQ instead targets unified multimodal models with routed pathways and mixed token types.

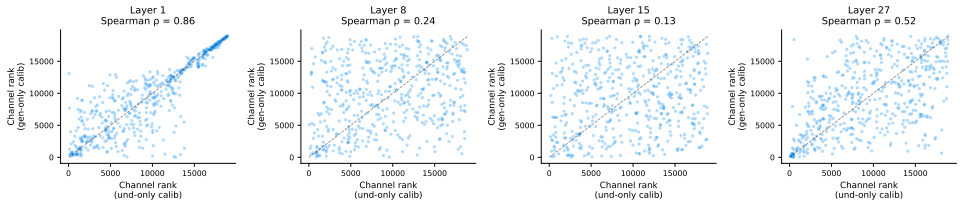

Figure 4: Channel-importance rankings from understanding-only and generation-only calibration often disagree across BAGEL layers, motivating UniQ's routed dual-modality calibration.

3 METHOD

Figure 3 provides an overview of UniQ. The framework follows BAGEL's routed token flow and applies three complementary PTQ components: dual-modality Hessian calibration collects task-aware statistics for the modules that each token type actually visits, pathway-adaptive mixed precision assigns weight bit-widths at the layer-pathway granularity, and token-aware sparse channel rotation reduces activation outlier error within each quantization group before low-bit inference.

3.1 PRELIMINARIES

BAGEL Deng et al. (2025) is a decoder-only unified multimodal model with a Mixture-of-Transformer-Experts backbone. Each Transformer layer contains a base pathway and a generation pathway. In the understanding mode, text tokens $\mathcal{T}_{text}$ and ViT tokens $\mathcal{T}_{vit}$ use the base pathway. In the generation mode, text tokens still use the base pathway, while VAE latent tokens $\mathcal{T}_{vae}$ use the generation pathway. This hard routing activates about 7B parameters in each forward pass out of 14B total parameters, and it exposes quantization to both token-level and pathway-level heterogeneity.

For a linear layer with weight matrix W and input activation matrix X, UniQ quantizes both weights and activations with uniform low-bit quantization. For a tensor value x in a quantization group G, the integer representation is

$$x_{\text{int}} = \text{clamp}\left(\left\lfloor \frac{x}{s_G} \right\rceil + z_G, 0, 2^b - 1\right), \tag{1}$$

where b is the bit-width, s_G is the scale, z_G is the zero point, $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, and $\text{clamp}(\cdot, 0, 2^b - 1)$ restricts the result to the quantized range. For asymmetric min-max quantization, the scale and zero point are

$$s_G = \frac{\max(x_G) - \min(x_G)}{2^b - 1}, \quad z_G = \left\lfloor -\frac{\min(x_G)}{s_G} \right\rceil, \tag{2}$$

where x_G denotes all values in group G. The dequantized value is $\hat{x} = s_G(x_{\text{int}} - z_G)$. The group structure is important in UniQ because activation outliers affect s_G, and any transformation that moves energy across groups can change the quantization range of groups that were originally stable.

Weight quantization uses a bit-width b_w and group size g_w, while activation quantization uses b_a and g_a. UniQ aligns these quantization decisions with modality and pathway structure through dual-modality calibration in Section 3.2, pathway-adaptive mixed precision in Section 3.3, and token-aware sparse rotation in Section 3.4.

3.2 DUAL-MODALITY HESSIAN CALIBRATION

Hessian-based weight quantization depends on calibration activations. Given calibration inputs to a linear layer, the local second-order statistics can be written as

$$H = \sum_n x_n x_n^\top, \tag{3}$$

where x_n is one token activation. Standard single-mode calibration is biased in BAGEL because it cannot cover all routed distributions. We observe that **text, ViT, and VAE latent tokens induce**

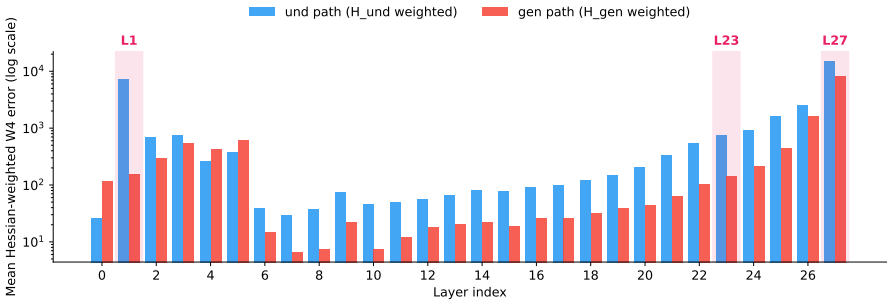

Figure 5: Per-layer W4A4 sensitivity differs between the understanding/base and generation pathways, supporting pathway-specific mixed-precision allocation.

different activation directions and different Hessian spectra even when they are processed by the same backbone. Figure 4 illustrates this mismatch through the channel-importance rankings induced by understanding-only and generation-only calibration. Their agreement varies sharply across layers, showing that a single calibration source can prioritize different channels from the ones used by the other task. Text-heavy calibration therefore overemphasizes language-dominant directions, while generation-only calibration misses the mixed text-vision statistics of the base pathway. This mismatch can make weight rounding preserve one mode while providing insufficient signal for the other.

UniQ constructs a joint calibration set from understanding and generation data,

$$\mathcal{D}_{cal} = \mathcal{D}_{und} \cup \mathcal{D}_{gen}. \tag{4}$$

The understanding subset contains image-text inputs, which activate the base pathway with both text and ViT tokens. The generation subset contains text-to-image prompts, which activate the base pathway with text tokens and the generation pathway with VAE latent tokens. For each module, UniQ accumulates Hessian statistics only from the tokens routed to that module. The resulting statistics are

$$H_m = \sum_{x_n \in \mathcal{A}(m, \mathcal{D}_{und})} x_n x_n^\top + \sum_{x_n \in \mathcal{A}(m, \mathcal{D}_{gen})} x_n x_n^\top, \tag{5}$$

where $\mathcal{A}(m, \mathcal{D})$ denotes the activations observed by module m under dataset $\mathcal{D}$. This formulation differs from naive data mixing because the accumulation follows the model routing. Shared base modules receive signals from text and visual understanding tokens, while generation modules receive signals from VAE latent tokens.

We apply this calibration in a sequential manner across layers. After quantizing one layer, UniQ forwards the calibration samples through the quantized layer and collects updated activations for the next layer. This procedure keeps the calibration distribution close to the actual quantized network and reduces the mismatch caused by error propagation. The objective is not to increase the Hessian magnitude, but to make it faithful to the token and pathway distributions that the quantized model uses at inference time.

3.3 Pathway-Adaptive Mixed-Precision Weight Quantization

Uniform bit-width allocation treats all layers and pathways as equally tolerant to quantization. This assumption is insufficient for BAGEL because the base pathway and the generation pathway process different token distributions and contribute to different tasks. We observe that **the sensitivity order of layers changes across pathways, so a layer that is redundant for understanding can still be critical for generation**. Figure 5 shows that the two pathways have different layer-wise error profiles, with several highly sensitive layers shared across pathways but many intermediate layers changing relative importance. This asymmetry suggests that precision should be assigned to layer-pathway pairs rather than to whole layers. Let l denote a Transformer layer, $p \in \{und, gen\}$ denote a pathway, and $\mathcal{B}$ denote a finite set of candidate bit-widths. For each pair (l, p), UniQ estimates the sensitivity of assigning bit-width $b \in \mathcal{B}$ by quantizing only that pair while keeping the rest of the model at high precision:

$$\Delta_{l,p}(b) = \mathcal{M}_{fp} - \mathcal{M}_{q(l,p,b)}, \tag{6}$$

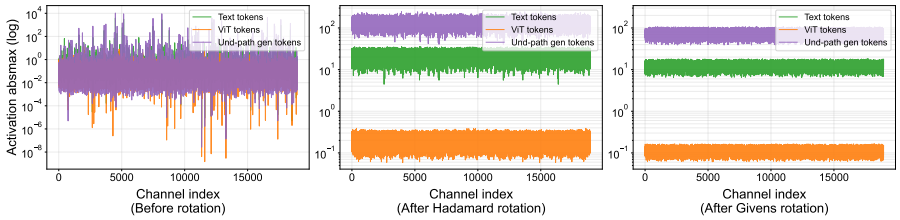

Figure 6: Activation magnitudes before rotation, after Hadamard rotation, and after UniQ's Givens rotation. Givens rotation reduces outliers more effectively while preserving group-local structure.

where $\mathcal{M}_{fp}$ is the full-precision task score and $\mathcal{M}_{q(l,p,b)}$ is the score after quantizing the corresponding layer-pathway pair to b bits. This profiling produces two independent sensitivity tables, one for the base pathway and one for the generation pathway. The separation is necessary because a layer can be robust for understanding but sensitive for generation.

Given a target average bit-width $\bar{b}$, UniQ chooses the bit-width assignment $b_{l,p}$ by balancing compression and sensitivity:

$$\min_{\{b_{l,p}\in\mathcal{B}\}} \sum_{l,p} \Delta_{l,p}(b_{l,p}) \quad \text{s.t.} \quad \frac{\sum_{l,p} N_{l,p} b_{l,p}}{\sum_{l,p} N_{l,p}} \le \bar{b}, \tag{7}$$

where $N_{l,p}$ is the number of quantized weights in layer l and pathway p. In practice, this budgeted assignment is solved with a greedy ranking over the sensitivity increase per saved bit. This formulation avoids committing to a fixed set of precision tiers in the method description. It also matches the MoT structure of BAGEL because the two pathways are implemented by separate linear modules, so pathway-specific bit-widths introduce no additional routing.

3.4 Token-Aware Sparse Channel Rotation

Activation quantization is strongly affected by outlier channels. In per-group min-max quantization, a single large channel can determine the range of an entire group and underutilize many quantization levels. We observe that **activation outliers are sparse and token-dependent: text tokens often concentrate energy in a few channels, while ViT and generation-path tokens have different channel ranges**. Figure 6 compares activations before rotation, after a dense Hadamard rotation, and after our sparse Givens rotation. Dense Hadamard rotation suppresses isolated spikes, but it spreads outlier energy broadly and raises many channel magnitudes, which can enlarge per-group quantization ranges. In contrast, sparse Givens rotation selectively smooths high-energy channels while keeping the transformation inside each quantization group. UniQ therefore applies sparse Givens rotations inside each activation quantization group and uses token-specific rotation parameters.

For a token type τ and group G, UniQ learns a sparse orthogonal matrix $R_{\tau,G}$ composed of a small number of Givens rotations. A single Givens rotation on channels $i, j \in G$ is

$$R(i,j,\theta) = I, \quad R_{ii} = R_{jj} = \cos\theta, \quad R_{ij} = -\sin\theta, \quad R_{ji} = \sin\theta. \tag{8}$$

For an activation vector x, UniQ quantizes the rotated activation and then maps it back:

$$\hat{x}_{\tau,G} = R_{\tau,G}^{\top} Q_a(R_{\tau,G} x_G; b_a, g_a). \tag{9}$$

Since $R_{\tau,G}$ only mixes channels inside G, it preserves the group boundary used by the scale s_G in Section 3.1. The inverse is exact because $R_{\tau,G}^{-1} = R_{\tau,G}^{\top}$, so the only approximation comes from quantization on the rotated activation.

The rotation is selected from calibration activations by minimizing post-quantization error within each token type and group:

$$R^*_{\tau,G} = \arg\min_{R\in\mathcal{R}_K(G)} \sum_{x\in\mathcal{C}_{\tau,G}} \|x_G - R^{\top} Q_a(R x_G; b_a, g_a)\|_2^2, \tag{10}$$

where $\mathcal{C}_{\tau,G}$ is the set of calibration activations for token type τ restricted to group G, and $\mathcal{R}_K(G)$ denotes rotations formed by at most K Givens pairs within the group. UniQ uses a data-driven

sparse search: it pairs high-energy channels with low-energy channels in the same group, evaluates the quantization error after rotation, and retains a pair only when the error decreases. Separate rotations are stored for text tokens, ViT tokens, generation text tokens, and VAE latent tokens. During inference, a token index tracker selects the corresponding $R_{\tau,G}$ before activation quantization, which prevents smooth visual or latent tokens from being unnecessarily transformed by rotations learned for text outliers.

4 Experiments

4.1 Experimental Setup

Model. We evaluate UniQ on BAGEL-7B-MoT Deng et al. (2025), a unified multimodal model for visual understanding, text-to-image generation, and image editing. BAGEL is a suitable testbed because its base and generation pathways process different token types within the same Mixture-of-Transformer-Experts backbone.

Benchmarks. We evaluate visual understanding on MME Fu et al. (2023), MMVP Tong et al. (2024), and MMMU Yue et al. (2024); compositional text-to-image generation on GenEval Ghosh et al. (2023); and image editing on GEdit-Bench Liu et al. (2025). These benchmarks cover perception, reasoning, prompt following, generation fidelity, and editing quality.

Baselines. We compare with QuaRot Ashkboos et al. (2024), ViDiT-Q Zhao et al. (2024), DVDQuant Li et al. (2026), MASQuant Hu et al. (2026), and SVDQuant Li* et al. (2025), using group size 64 for all methods. When possible, each baseline follows the best-performing configuration reported in its original paper.

Table 1: Quantization results on multimodal understanding benchmarks (MME, MMVP, MMMU). The best and the second best results within each quantization-bit group are marked with red and blue, respectively.

Model	Methods	Bits (W/A)	MME Total ↑	MME Perception ↑	MME Cognition ↑	MMVP Pairwise ↑	MMVP Per-Q ↑	MMMU Overall ↑
BAGEL-7B-MoT (28 layers)	FP	16/16	2378	1650	728	16.0	56.7	51.6
	QuaRot	4/4	56	38	18	0.7	26.2	15.1
	ViDiT-Q	4/4	135	98	37	1.3	22.0	28.0
	DVDQuant	4/4	72	51	21	0.9	29.7	16.4
	MASQuant	4/4	1910	1408	502	4.7	46.3	41.2
	SVDQuant	4/4	2082	1570	512	9.7	51.5	47.3
	UniQ (ours)	4/4	2309	1630	678	16.4	57.1	52.0
	SVDQuant	3/4	1570	1152	418	1.4	32.6	42.4
	UniQ (ours)	3/4	2159	1550	608	15.3	53.3	51.0

Quantization protocol. All methods report W4A4 results, and SVDQuant and UniQ additionally report W3A4. We use symmetric per-group quantization for Transformer linear layers, keep non-Transformer input/output projections in high precision, and apply static weight quantization with dynamic activation quantization. UniQ calibrates with 2 understanding images and 32 GenEval prompts, then assigns weight precision by layer-pathway sensitivity and applies token-aware sparse Givens rotations for activation quantization. More protocol details are provided in Appendix B.

4.2 Main Results

Visual understanding. Table 1 shows that UniQ is consistently the strongest quantized model on understanding tasks. At W4A4, UniQ reaches 2309 MME, 16.4 MMVP pairwise accuracy, and 52.0 MMMU, nearly matching or slightly exceeding FP16 on MMVP and MMMU. Under the harder W3A4 setting, UniQ keeps 2159 MME and 51.0 MMMU, while SVDQuant drops to 1570 and 42.4. This gap indicates that protecting both text- and vision-dominant directions is critical for unified multimodal understanding.

Table 2: Quantization results on the GenEval benchmark. The best and the second best results within each quantization-bit group are marked with red and blue, respectively.

Model	Methods	Bits (W/A)	Single Object ↑	Two Object ↑	Counting ↑	Colors ↑	Color Attr. ↑	Position ↑	Overall ↑
BAGEL-7B-MoT (28 layers)	FP	16/16	98.75	98.00	76.74	96.00	58.00	57.00	80.74
	QuaRot	4/4	31.25	21.00	12.00	28.00	8.00	6.00	17.71
	ViDiT-Q	4/4	99.38	89.00	72.09	89.00	51.00	46.00	74.41
	DVDQuant	4/4	41.75	29.00	15.00	31.00	13.00	11.00	23.46
	MASQuant	4/4	98.75	85.00	74.42	86.00	40.00	39.00	70.53
	SVDQuant	4/4	98.75	95.00	73.26	90.00	45.00	43.00	74.17
	UniQ (ours)	4/4	99.38	96.00	75.58	95.00	58.00	49.00	78.82
	SVDQuant	3/4	96.88	74.00	62.79	84.00	28.00	34.00	63.28
	UniQ (ours)	3/4	98.75	90.00	75.58	87.00	56.00	42.00	74.89

Table 3: Quantization results on the GEdit-Bench-EN (Full set). G_{SC}, G_{PQ}, and G_O denote Semantic Consistency, Perceptual Quality, and Overall scores judged by GPT-5 (0–10, higher is better). The best and the second best results within each quantization-bit group are marked with red and blue, respectively.

Model	Methods	Bits (W/A)	G_{SC} ↑	G_{PQ} ↑	G_O ↑
BAGEL-7B-MoT (28 layers)	FP	16/16	6.57	6.02	6.02
	QuaRot	4/4	0.02	1.06	0.04
	ViDiT-Q	4/4	5.73	4.90	5.03
	DVDQuant	4/4	0.02	2.08	0.05
	MASQuant	4/4	4.61	4.38	4.12
	SVDQuant	4/4	5.93	5.27	5.28
	UniQ (ours)	4/4	6.51	6.05	6.07
	SVDQuant	3/4	5.02	4.60	4.43
	UniQ (ours)	3/4	6.12	5.87	5.92

Text-to-image generation. Table 2 reports the GenEval results. UniQ obtains the best W4A4 overall score, 78.82, close to the FP16 score of 80.74 and ahead of ViDiT-Q, MASQuant, and SVDQuant. At W3A4, UniQ keeps 74.89 overall, improving over SVDQuant by 11.61 points. The largest gains appear in counting and attribute-related categories, where pathway-aware precision helps preserve compositional generation.

Image editing. Table 3 further evaluates instruction-guided image editing. UniQ achieves the best W4A4 scores on semantic consistency, perceptual quality, and overall quality, with a 6.07 overall score compared with 6.02 for FP16. At W3A4, UniQ still reaches 5.92 overall, substantially

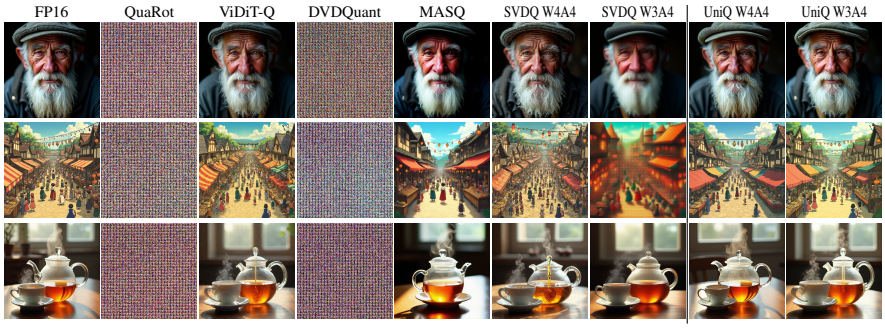

Figure 7: Qualitative text-to-image comparison on representative prompts. Columns follow the table order, with UniQ separated on the right; UniQ better preserves FP16-like structure and prompt fidelity under W4A4 and W3A4.

Table 4: Ablation on three UniQ components (BAGEL-7B, W3A4). A: dual-modality calibration; B: path-aware mixed-precision bit allocation; C: Givens rotation activation quantization. "none" is plain RTN-W3A4.

Config	MME-T	MME-P	MME-C	MMMU	MMVP	GenEval
none	227	154	73	19.3	7.3	10.26
A	1633	1252	381	32.2	34.7	40.73
B	1681	1264	417	35.8	37.2	43.26
C	1580	1121	459	24.9	28.1	39.15
A+B	2029	1428	601	45.2	47.1	65.17
A+C	1824	1368	456	42.8	43.0	54.34
B+C	1703	1302	401	39.1	41.0	50.89
UniQ	**2159**	**1550**	**608**	**51.0**	**53.3**	**74.89**

above SVDQuant's 4.43. This shows that UniQ preserves not only generation quality, but also the instruction-following and visual consistency needed for editing.

4.3 Ablation Study

Component effects. Table 4 ablates UniQ at W3A4. Plain RTN collapses to 227 MME-T and 10.26 GenEval, confirming that direct low-bit rounding is not sufficient. Each component gives a clear recovery: dual-modality calibration (A) reaches 1633 MME-T and 40.73 GenEval, path-aware bit allocation (B) reaches 1681 and 43.26, and Givens rotation (C) reaches 1580 and 39.15. Thus, all three modules address real failure modes, with A and B providing the largest standalone gains.

Combined effects. Combining components further improves performance. A+B is the strongest two-component variant, reaching 2029 MME-T, 45.2 MMMU, and 65.17 GenEval, showing that accurate calibration and pathway-aware bit assignment reinforce each other. Full UniQ performs best on every metric, reaching 2159 MME-T, 51.0 MMMU, 53.3 MMVP, and 74.89 GenEval. The gain over A+B confirms that activation rotation remains useful after weight-side errors are reduced.

4.4 CUDA Kernel Implementation and Efficiency

Kernel implementation. We implement UniQ on top of the open-source Nunchaku engine from SVDQuant Li* et al. (2025). We extend it to support BAGEL's Mixture-of-Transformer-Experts routing, separate base/generation pathways, token-type-specific activation quantization, and fused symmetric per-group dequantization with GEMM.

End-to-end adaptation to BAGEL. BAGEL inference may contain text tokens, ViT features, and VAE latents in one request, so the runtime tracks token type and pathway assignment throughout the forward pass. Static low-bit weights are loaded in packed form, while dynamic activation scales and sparse Givens rotations are applied per group at runtime.

Efficiency. With these extensions, UniQ achieves a 2.9× end-to-end inference speedup over full-precision BAGEL. It reduces model storage by 4.51×, inference memory by 3.97×, and peak memory by 3.72×. These measurements cover the complete BAGEL pipeline, including routing, quantization, dequantization, and multimodal pre/post-processing.

5 Conclusion

We presented UniQ, a modality-aware post-training quantization framework for unified multimodal models. Starting from BAGEL, we showed that quantization failure is closely tied to token-level and pathway-level heterogeneity: different token types induce different calibration statistics, different pathways have different precision sensitivity, and activation outliers are sparse and token-dependent. UniQ addresses these issues through routed dual-modality Hessian calibration, pathway-adaptive mixed-precision allocation, and token-aware sparse Givens rotation. Experiments across visual understanding, text-to-image generation, image editing, and knowledge-intensive generation show that UniQ enables aggressive W3A4 quantization while preserving unified multimodal capability and improving inference efficiency. We hope this work provides a practical foundation for deploying unified understanding-and-generation models under tight memory and latency budgets.

AI USE STATEMENT

We utilized AI tools to assist with English writing and to polish the text. Additionally, we employed AI tools to search for and introduce relevant information.

ETHICS STATEMENT

UniQ reduces the memory requirements and inference cost of unified multimodal models, potentially broadening access to multimodal research and applications. Our experiments use pretrained models and existing benchmarks for visual understanding, image generation, and editing. Lower deployment costs may also facilitate misuse, including the creation of misleading visual content and unauthorized image manipulation. Quantized models may retain biases and harmful behaviors of their pretrained counterparts, and preserving task performance does not establish that safety-related behavior remains unchanged. Our evaluation does not include a dedicated assessment of these risks. We therefore recommend evaluating the safety of quantized checkpoints before deployment, retaining applicable safeguards, and respecting the licenses and usage restrictions of the underlying models and datasets.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *arXiv preprint arXiv:2404.00456*, 2024.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in neural information processing systems*, 36: 4396–4429, 2023.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. In *CVPR*, 2025a.

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*, 2025b.

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

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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.

Lulu Hu, Wenhu Xiao, Xin Chen, Xinhua Xu, Bowen Xu, Kun Li, and Yongliang Tao. Masquant: Modality-aware smoothing quantization for multimodal large language models. *arXiv preprint arXiv:2603.04800*, 2026.

Muyang Li*, Yujun Lin*, Zhekai Zhang*, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, et al. Mbq: Modality-balanced quantization for large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4167–4177, 2025.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17535–17545, October 2023.

Zhiteng Li, Xianglong Yan, Tianao Zhang, Haotong Qin, Dong Xie, Jiang Tian, zhongchao shi, Linghe Kong, Yulun Zhang, and Xiaokang Yang. Arb-llm: Alternating refined binarizations for large language models. *arXiv preprint arXiv:2410.03129*, 2024.

Zhiteng Li, Hanxuan Li, Junyi Wu, Kai Liu, Linghe Kong, Guihai Chen, Yulun Zhang, and Xiaokang Yang. Dvd-quant: Data-free video diffusion transformers quantization. In *ICLR*, 2026.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In *MLSys*, 2024.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Guangshuo Qin, Zhiteng Li, Zheng Chen, Weihang Zhang, Linghe Kong, and Yulun Zhang. Veq: Modality-adaptive quantization for moe vision-language models. *arXiv preprint arxiv:2602.01037*, 2026.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *CVPR*, 2023.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*, 2023.

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

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

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.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024a.

Junyi Wu, Haoxuan Wang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Ptq4dit: Post-training quantization for diffusion transformers. In *NeurIPS*, 2024b.

Junyi Wu, Zhiteng Li, Zheng Hui, Yulun Zhang, Linghe Kong, and Xiaokang Yang. Quantcache: Adaptive importance-guided quantization with hierarchical latent and layer caching for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15035–15044, 2025a.

Junyi Wu, Zhiteng Li, Haotong Qin, Xiaohong Liu, Linghe Kong, Yulun Zhang, and Xiaokang Yang. Flashedit: Decoupling speed, structure, and semantics for precise image editing. *arXiv preprint arXiv:2509.22244*, 2025b.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Xianglong Yan, Chengzhu Bao, Zhiteng Li, Tianao Zhang, Kaicheng Yang, Haotong Qin, Ruobing Xie, Xingwu Sun, and Yulun Zhang. Pt^2-llm: Post-training ternarization for large language models. *arXiv preprint arXiv:2510.03267*, 2025.

Kaicheng Yang, Xun Zhang, Haotong Qin, Yucheng Lin, Kaisen Yang, Xianglong Yan, and Yulun Zhang. Robuq: Pushing dits to w1. 58a2 via robust activation quantization. *arXiv preprint arXiv:2509.23582*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Tianao Zhang, Zhiteng Li, Xianglong Yan, Haotong Qin, Yong Guo, and Yulun Zhang. Quant-dllm: Post-training extreme low-bit quantization for diffusion large language models. *arXiv preprint arXiv:2510.03274*, 2025.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Enshu Liu, Rui Wan, Widyadewi Soedarmadji, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, et al. Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation. *arXiv preprint arXiv:2406.02540*, 2024.

A Limitations and Broader Impact

Limitations. Our study focuses on a single unified multimodal model, BAGEL-7B-MoT, which uses an Mixture-of-Transformer-Experts decoder with explicit understanding and generation pathways. While the three failure modes we identify—routed token-distribution heterogeneity, pathway-dependent layer sensitivity, and sparse token-specific activation outliers—are properties of the unified-modeling paradigm rather than of a particular checkpoint, additional validation on architectures with different routing styles (e.g., Janus-Pro's separate vision encoders or fully tokenized models such as Chameleon) would more thoroughly establish the generality of UniQ. Second, our calibration procedure relies on a small mixed-modality dataset (16 understanding images and 32 generation prompts); how to construct calibration sets that generalize across editing, video, or longer-context tasks remains open. Finally, our pathway-adaptive bit allocation is currently driven by a sensitivity proxy computed under per-layer perturbation; a search over joint per-layer-per-pathway bit assignments could further improve quality at the cost of additional calibration time.

Broader impact. UniQ reduces the deployment cost of unified multimodal models by lowering memory footprint and per-token compute, which makes such models more accessible on commodity GPUs and edge accelerators and reduces the energy required per inference call. As with other techniques that make generative models cheaper to deploy, this lower barrier can also amplify the misuse of underlying capabilities, including the production of misleading visual content or unauthorized image edits. UniQ is a post-training compression method, and our task-performance benchmarks do not establish that quantization preserves the safety-related behavior of the upstream model. We recommend evaluating quantized checkpoints for safety before deployment, retaining applicable safeguards, and following the original model license and usage restrictions. We recommend that practitioners follow the original model license, retain any safety classifiers shipped with the model, and use responsible release practices when deploying quantized variants.

B Detailed Experimental Protocol

Benchmark details. For visual understanding, MME measures both perception and cognition ability, MMVP contains CLIP-blind image pairs with binary multiple-choice questions, and we report pairwise accuracy where a pair is counted as correct only when both questions are answered correctly. MMMU evaluates multi-discipline multimodal reasoning over college-level questions. For generation, GenEval measures compositional text-to-image fidelity across object, counting, color, and spatial-relation categories. For image editing, GEdit-Bench reports semantic consistency, perceptual quality, and overall editing quality. WISE-Verified is used as an additional knowledge-intensive generation benchmark covering cultural, temporal, spatial, biological, physical, and chemical knowledge.

Quantization protocol. All methods report W4A4 results. In addition, SVDQuant and UniQ report W3A4 results to evaluate robustness under a more aggressive weight-compression regime. We use symmetric quantization for both weights and activations because symmetric integer arithmetic is friendlier to fused GPU kernels and hardware acceleration. UniQ applies static weight quantization and dynamic activation quantization: weights are quantized once after calibration, while activation scales are computed online for each inference batch using per-group statistics. UniQ uses a small calibration set: 2 understanding images paired with two short visual-grounding prompts provide text and ViT tokens for the base pathway, while 32 generation prompts drawn from the GenEval prompt distribution, covering all six task categories, provide text tokens for the base pathway and VAE latent tokens for the generation pathway.

Quantized modules. Unless otherwise specified, all methods quantize the Transformer linear layers and leave non-Transformer input/output projection modules in high precision. For mixed precision, UniQ assigns bit-widths at the layer-pathway granularity under the target average weight bit-width. For activation quantization, UniQ applies token-aware sparse Givens rotations inside each quantization group. All methods use per-group granularity with group size 64, so comparisons share the same low-level quantization setting.

C UniQ Quantization Algorithm

Algorithm 1 UniQ post-training quantization

Require: Model M, calibration data $\mathcal{D}_{und}, \mathcal{D}_{gen}$, target weight bit-width $\bar{b}$, activation bit-width b_a, group size g

Ensure: Quantized model $\hat{M}$

$\mathcal{D}_{cal} \leftarrow \mathcal{D}_{und} \cup \mathcal{D}_{gen}$
Record token type τ and pathway p for calibration activations
for each linear module m **do**
    $H_m \leftarrow \sum_{x \in \mathcal{A}(m, \mathcal{D}_{cal})} xx^\top$
end for
for each layer-pathway pair (l, p) **do**
    Profile sensitivity $\Delta_{l,p}(b)$ for candidate bit-widths
end for
Assign $\{b_{l,p}\}$ under average bit-width $\bar{b}$
for each linear module m in Transformer order **do**
    Quantize W_m using H_m, $b_{l,p}$, and group size g
    Update calibration activations through quantized m
end for
for each token type τ and group G **do**
    Learn sparse Givens rotation $R_{\tau,G}$
end for
Pack low-bit weights and store $\{R_{\tau,G}\}$
Use token-aware activation rotation and dynamic activation quantization at inference
return Quantized model $\hat{M}$

D Additional Knowledge-Intensive Generation Results

Table 5: Quantization results on the WISE_Verified benchmark. The best and the second best results within each quantization-bit group are marked with red and blue, respectively.

Model	Methods	Bits (W/A)	Culture ↑	Time ↑	Space ↑	Biology ↑	Physics ↑	Chemistry ↑	WiScore Overall ↑
BAGEL-7B-MoT (28 layers)	FP	16/16	45.00	48.33	57.50	35.00	50.83	27.50	0.443
	QuaRot	4/4	1.75	0.67	3.33	4.17	5.83	2.50	0.027
	ViDiT-Q	4/4	37.00	32.50	32.50	22.50	49.17	15.83	0.331
	DVDQuant	4/4	5.25	5.83	8.33	9.16	14.17	4.17	0.071
	MASQuant	4/4	32.50	26.67	30.00	15.83	41.67	13.33	0.283
	SVDQuant	4/4	39.75	40.83	34.45	20.83	45.83	17.50	0.350
	UniQ (ours)	4/4	44.75	46.67	48.33	33.33	51.67	25.83	0.426
	SVDQuant	3/4	34.25	35.83	30.00	18.33	41.67	12.50	0.303
	UniQ (ours)	3/4	43.00	44.17	46.67	30.83	47.50	22.50	0.420

WISE-Verified. Table 5 reports knowledge-intensive text-to-image generation results on WISE-Verified. At W4A4, UniQ reaches a WiScore of 0.426, close to the FP16 score of 0.443 and clearly higher than SVDQuant's 0.350. UniQ also obtains the best W4A4 result in every individual knowledge category, including culture, time, space, biology, physics, and chemistry. Under W3A4, UniQ keeps a WiScore of 0.420, while SVDQuant drops to 0.303. These results suggest that UniQ better preserves the text-conditioned knowledge and generation pathways needed for knowledge-grounded image synthesis.

E Additional Visual Comparisons

Qualitative generation results. Figures 9–8 provide additional text-to-image comparisons using the same method order as Figure 7. These examples complement the GenEval results in Table 2: methods that collapse numerically also tend to lose object structure, color binding, or spatial layout in the generated images. UniQ remains closer to the FP16 outputs under W4A4 and degrades more gracefully under W3A4, preserving both global scene composition and local prompt details across a broader set of prompts.



Figure 8: Additional visual comparison.

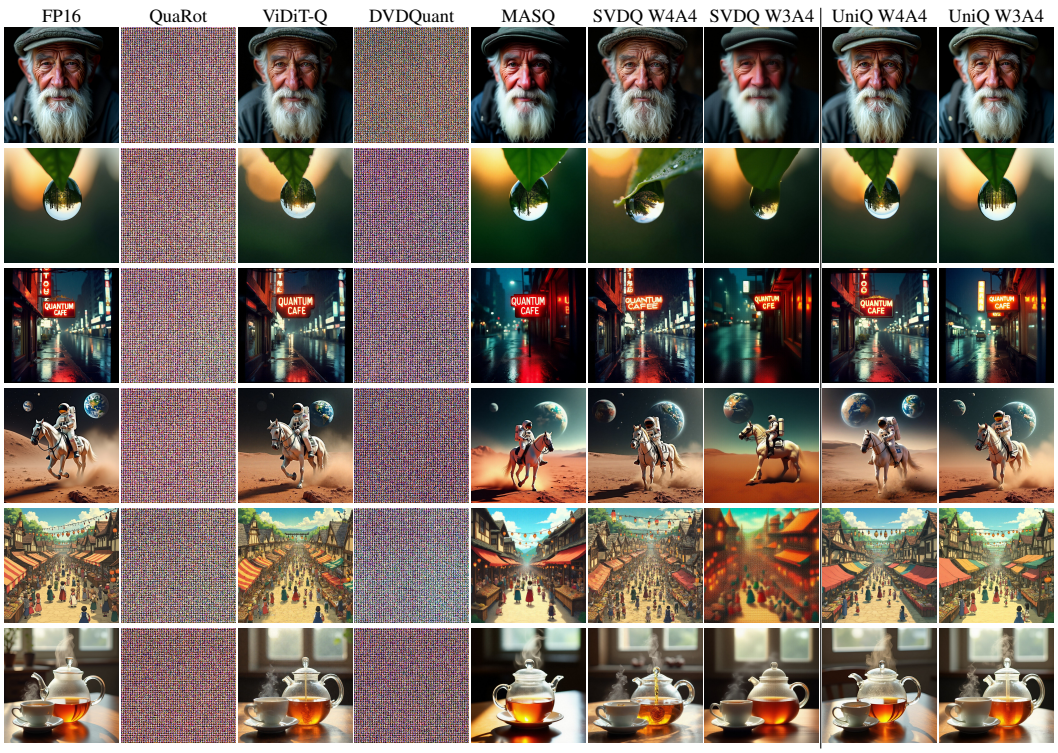

Figure 9: Additional visual comparison.

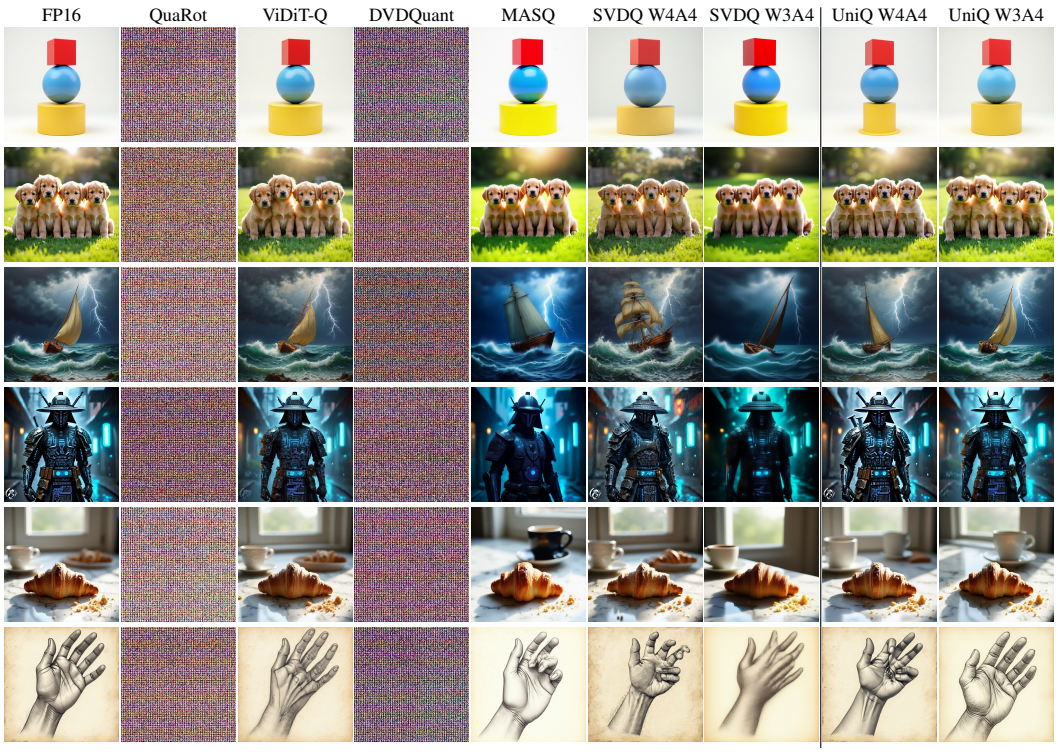

Figure 10: Additional visual comparison.

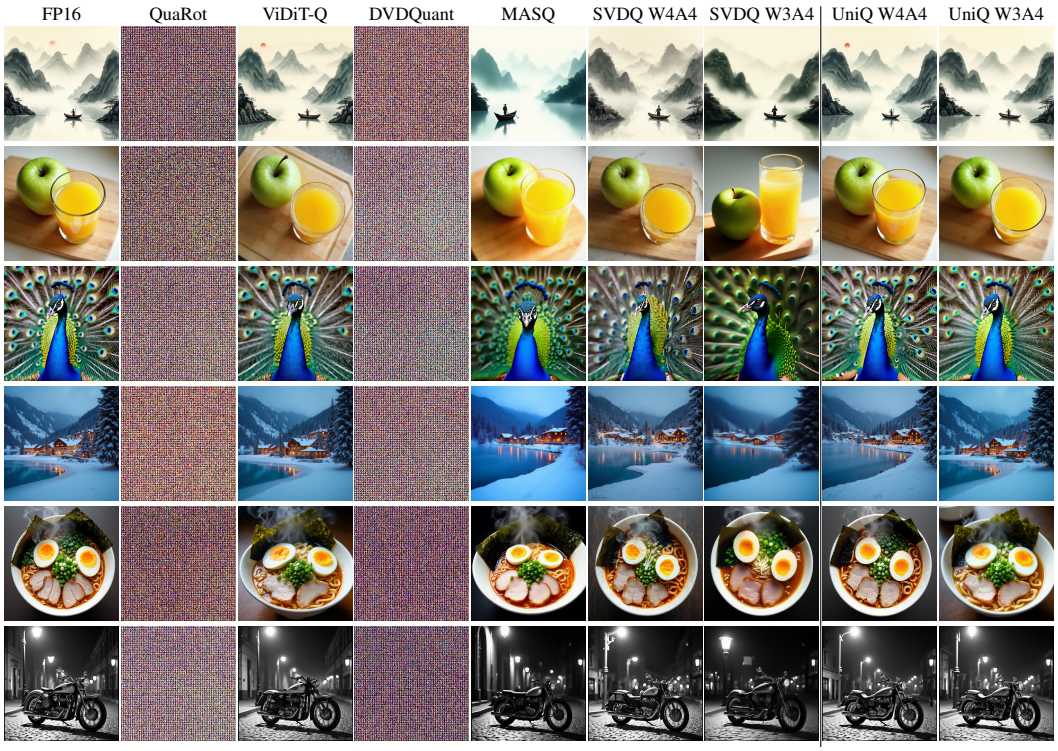

Figure 11: Additional visual comparison.

Method	Bits (W/A)	MME-P ↑	MMMU ↑	GenEval ↑
FP16	16/16	1566.9	40.9	80.34
MASQuant	4/4	1323.6	31.1	70.27
SVDQuant	4/4	1523.3	37.5	75.38
UniQ	4/4	1564.2	40.5	80.13
SVDQuant	3/4	1134.8	29.4	68.72
UniQ	3/4	1527.6	38.1	76.46

Table 6: Additional Janus-Pro-7B results reported for visual understanding and text-to-image generation. GenEval denotes the overall score.

Method	Bits (W/A)	MMB ↑	MMMU ↑	GenEval ↑
FP16	16/16	58.3	31.8	66.04
MASQuant	4/4	47.9	24.7	52.73
SVDQuant	4/4	55.2	29.7	64.55
UniQ	4/4	58.1	31.5	66.36
SVDQuant	3/4	46.5	23.6	53.23
UniQ	3/4	57.3	30.2	64.71

Table 7: Additional Emu3 results. MMB replaces MME-P as the understanding benchmark; GenEval denotes the overall score.

F Additional Cross-Architecture Evaluation

The main paper studies BAGEL-7B-MoT. We additionally evaluate Janus-Pro-7B and Emu3 to examine whether the token-aware calibration and quantization pipeline transfers to other unified multimodal designs. The W4A4 and W3A4 settings use parameter-weighted average weight bit-widths of 4 and 3, respectively, with 4-bit activations. Calibration and allocation follow each model's actual token flow and parameter sharing. In particular, when a parameter is fully shared across tasks, it receives one bit-width based on an aggregate sensitivity signal rather than separate pathway-specific bit maps.

Table 6 reports Janus-Pro-7B results. At W4A4, UniQ remains close to FP16 on all three metrics and exceeds the two quantized baselines shown. At W3A4, it retains 1527.6 MME-P, 38.1 MMMU, and 76.46 GenEval, compared with 1134.8, 29.4, and 68.72 for SVDQuant. These results support transfer to this architecture under the stated adaptation of the calibration and allocation granularity.

Table 7 gives the analogous Emu3 comparison. UniQ preserves scores close to FP16 at W4A4 and retains 57.3 MMB, 30.2 MMMU, and 64.71 GenEval at W3A4, above the reported W3A4 SVDQuant scores. Small differences near FP16, including the 0.32 GenEval difference at W4A4, should not be interpreted as reliable improvements without repeated evaluations. The results broaden the empirical evidence to three model families, but they do not establish performance on arbitrary unified architectures or tasks.

G Additional Method and Evaluation Details

Calibration and evaluation independence. The BAGEL calibration set contains 2 understanding images and 32 generation prompts. The generation calibration prompts are disjoint from the official GenEval test prompts at the prompt level. For a given physical module, Hessian statistics are accumulated only from the tokens actually routed to that module. Calibration proceeds in Transformer order: after a layer is quantized, the next layer's statistics are collected from activations produced by the partially quantized network. These details distinguish routed calibration from simply pooling two input datasets.

Pathway sensitivity and bit allocation. Section 3.3 describes task-score sensitivity conceptually, while Figure 5 visualizes Hessian-weighted W4 error. The implemented allocation uses an offline, single-layer, single-pathway perturbation-based robustness ordering under the target mixed-precision

budget. Official evaluation examples, labels, GenEval test prompts, and judge scores are not used for bit allocation or hyperparameter selection. A fully specified perturbation proxy, candidate bit set, data source, aggregation rule, and final bit map are required for exact reproduction and should accompany a code release.

Sparse Givens selection. For modules with activation excess kurtosis of at least 50, the search retains at most three channel pairs per activation-quantization group. It greedily pairs the highest- and lowest-absolute-maximum channels within the same group, selects an initial angle from seven candidates by quantization RMSE, and refines angles with 200 straight-through-estimator steps. A rotation is retained only if the complete rotate–quantize–inverse-rotate reconstruction RMSE is strictly lower than direct quantization. Parameters are stored by layer, token type, and module; a token tracker selects the corresponding parameters at inference. The component-level ablation in Table 4 supports complementarity of the three components, but does not independently isolate every routed, pathway-specific, and token-specific sub-mechanism.

Baseline scope. The reported comparisons use weight group size 64 and activation group size 32, with the same Transformer-linear quantization scope. Non-Transformer `vae2llm`, `llm2vae`, `time_embedder`, and `patch_embedding` projections remain in high precision. Method-specific calibration inputs, adaptation to BAGEL routing, effective checkpoint bit-width, and complete resource measurements should be reported alongside the released implementation; they are not inferred from the benchmark scores alone.

End-to-end timing protocol. Understanding and generation are measured separately at batch size 1 with three end-to-end repetitions. Model loading and kernel materialization precede timing, and the current protocol uses no additional warm-up. CUDA is synchronized before stopping each timer. Per-run duration, mean and minimum duration, and PyTorch CUDA peak allocated and reserved memory are recorded. An understanding request contains one image and one text query and generates at most eight think tokens; a generation request contains one text query and uses 30 sampling steps. Timing includes token tracking, dynamic activation scaling, token-aware rotation, low-bit matrix multiplication, routing, and multimodal pre- and post-processing. The software stack uses CUDA 12.8, PyTorch 2.8.0+cu128, and Nunchaku commit `8f41840596bd516d434a1f88ac16c86fdb64e74f`. Exact hardware configuration, input resolution, and absolute latency should be reported with the final measurements before interpreting the 2.9× speedup beyond this setup.

H CUDA Kernel Implementation and Efficiency

Kernel implementation. We implement UniQ on top of the open-source Nunchaku engine from SVDQuant Li* et al. (2025). We extend it to support BAGEL's Mixture-of-Transformer-Experts routing, separate base/generation pathways, token-type-specific activation quantization, and fused symmetric per-group dequantization with GEMM.

End-to-end adaptation to BAGEL. BAGEL inference may contain text tokens, ViT features, and VAE latents in one request, so the runtime tracks token type and pathway assignment throughout the forward pass. Static low-bit weights are loaded in packed form, while dynamic activation scales and sparse Givens rotations are applied per group at runtime.

Efficiency. With these extensions, UniQ achieves a 2.9× end-to-end inference speedup over full-precision BAGEL. It reduces model storage by 4.51×, inference memory by 3.97×, and peak memory by 3.72×. These measurements cover the complete BAGEL pipeline, including routing, quantization, dequantization, and multimodal pre/post-processing.

I Sparse Rotation Search Details

The rotation is selected from calibration activations by minimizing post-quantization error within each token type and group:

$$R^*_{\tau,G} = \arg\min_{R \in \mathcal{R}_K(G)} \sum_{x \in \mathcal{C}_{\tau,G}} \|x_G - R^\top Q_a(R x_G; b_a, g_a)\|_2^2, \tag{11}$$

where $\mathcal{C}_{\tau,G}$ is the set of calibration activations for token type τ restricted to group G, and $\mathcal{R}_K(G)$ denotes rotations formed by at most K Givens pairs within the group. UniQ uses a data-driven sparse search: it pairs high-energy channels with low-energy channels in the same group, evaluates the quantization error after rotation, and retains a pair only when the error decreases. Separate rotations are stored for text tokens, ViT tokens, generation text tokens, and VAE latent tokens. During inference, a token index tracker selects the corresponding $R_{\tau,G}$ before activation quantization, which prevents smooth visual or latent tokens from being unnecessarily transformed by rotations learned for text outliers.