PROMPTING WHAT MATTERS: ADAPTIVE CALIBRATION FOR DATA-FREE QUANTIZATION OF LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-Free Quantization (DFQ) provides a promising solution for compressing Large Vision-Language Models (LVLMs) without real calibration data. Existing DFQ methods mainly focus on global statistical matching, visual semantic preservation, or generic contrastive alignment, overlooking the *heterogeneous quantization sensitivity* of multimodal capabilities and model components. As a result, semantically plausible synthetic samples may fail to activate capability-specific quantization errors that dominate performance degradation. In this work, we reveal that LVLM quantization vulnerability is highly capability-module dependent, raising two fundamental questions: *which capabilities and modules are most sensitive to quantization* and *how can calibration samples be synthesized to recover them?* To address these challenges, we propose **PromptQ**, Prompt-Guided Data-Free Quantization framework that optimizes calibration distribution according to capability-module quantization vulnerability. PromptQ first introduces *Sensitivity-Driven Prompt Construction (SDPC)*, which estimates capability-module sensitivity through behavioral degradation and representation discrepancies under controlled quantization. Based on the estimated sensitivity, SDPC allocates calibration prompts across vulnerable capabilities while maintaining semantic diversity, and generates *module-aware prompts* that specify the target capability and quantization-sensitive component. Guided by these prompts, *Quantization-Sensitive Image Synthesis (QSIS)* synthesizes semantically valid calibration samples that expose informative quantization errors by increasing the discrepancy between full-precision and quantized representations under semantic constraints. By coupling sensitivity-aware capability allocation with module-targeted sample synthesis, PromptQ transforms data-free calibration from distribution matching into vulnerability-aware error discovery. Extensive experiments on diverse multimodal benchmarks demonstrate that our PromptQ consistently improves quantization performance without requiring real calibration data. *Our code is available in supplementary material package*.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) Liu et al. (2023); Li et al. (2023a); Dai et al. (2023); Liu et al. (2024a) have recently achieved remarkable progress in visual perception, multimodal understanding, and complex reasoning by integrating powerful vision encoders with large language models Bai et al. (2025). Meanwhile, the rapidly increasing model scale introduces substantial memory and computational overhead Li et al. (2025); Wang et al. (2024), limiting their deployment on resource-constrained devices. Post-training quantization (PTQ) Nagel et al. (2020); Frantar et al. (2022); Lin et al. (2024) has emerged as an effective solution by compressing pre-trained models into low-bit representations without expensive retraining, thereby reducing both memory footprint and inference latency. Nevertheless, PTQ critically depends on representative calibration data to estimate quantization parameters and preserve model behaviors Nagel et al. (2020); Shao et al. (2024). In practical scenarios, such calibration data may be inaccessible due to privacy constraints Liu et al. (2021) and deployment restrictions Nagel et al. (2019).

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

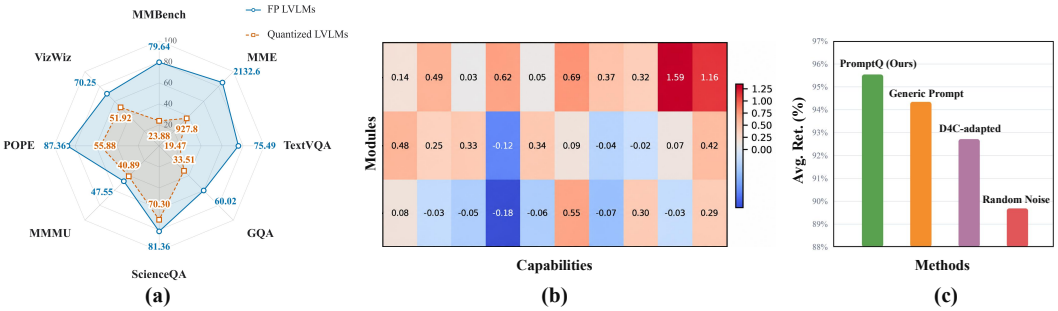

Figure 1: (a) Quantization introduces heterogeneous capability degradation between full-precision LVLMs and their quantized counterparts, revealing uneven sensitivity across multimodal tasks. (b) Capability-module pairs exhibit distinct quantization sensitivity, highlighting the necessity of sensitivity-aware calibration budget allocation. (c) Our PromptQ addresses the challenges by identifying quantization-critical capabilities and synthesizing targeted calibration samples, enabling effective data-free LVLM quantization. Avg. Ret. denotes average performance retention relative to the full-precision model across nine benchmarks. The results are obtained from Qwen2.5-VL-3B.

As a compelling alternative, Data-Free Quantization (DFQ) Cai et al. (2020); Xu et al. (2020); Choi et al. (2021); Zhong et al. (2022); Qian et al. (2023); Dung et al. (2024); Kim et al. (2025) aims to construct effective synthetic calibration data without accessing original data. Early DFQ methods synthesize images by matching activation or normalization statistics of convolutional networks Qian et al. (2023); Dung et al. (2024); Kim et al. (2025); Cai et al. (2020). Subsequent approaches for Vision Transformers (ViTs) exploit patch-level relationships Ramachandran et al. (2024), attention structure Choi et al. (2025); Zhong et al. (2025) and teacher-student alignment Qian et al. (2026). Generative approaches further leverage text-conditioned image synthesis Li et al. (2024) and learn multi-mode prompts Li et al. (2026) to improve semantic fidelity and sample diversity. These methods are primarily developed for vision-only tasks. More recently, D4C Zhang et al. (2026) extends DFQ to contrastive vision-language models by incorporating prompt-guided semantics and structural diversity into image synthesis. However, preserving contrastive image-text alignment alone cannot ensure that synthetic samples adequately support the calibration of instruction-conditioned LVLM behavior. To understand this issue, we investigate the quantization behavior of LVLMs and identify two key challenges:

1) Heterogeneous quantization sensitivity across multimodal capabilities. Unlike conventional models Kim et al. (2025); Qian et al. (2026); Zhang et al. (2026), LVLMs involve tightly coupled interactions among visual encoding, multimodal projection, and language reasoning. As a result, quantization errors are not uniformly distributed, but vary across different capabilities and model components, leading to heterogeneous capability degradation and distinct sensitivity patterns across capability-module pairs; see Fig.1(a)(b). This heterogeneity makes semantic coverage alone insufficient: under a limited calibration budget, prompt allocation should account for both vulnerable capabilities and the modules responsible for their degradation.

2) Capability-agnostic synthetic calibration. Such heterogeneous sensitivity challenges existing DFQ strategies, which typically optimize synthetic samples based on realism Li et al. (2024), diversity Qian et al. (2026), or image-text alignment Zhang et al. (2026) while overlooking capability-level vulnerability. Consequently, synthesized samples may fail to activate quantization-sensitive multimodal behaviors and provide insufficient signals for capability recovery, leading to suboptimal quantization performance; see Fig.1(c).

Motivated by these findings, we propose **PromptQ**, a novel Prompt-Guided Data-Free Quantization approach for LVLMs, as illustrated in Fig.2. Our basic idea is to introduce a sensitivity-aware calibration paradigm that explicitly considers *which capabilities are vulnerable to quantization* and *how calibration samples should be synthesized for their recovery*, which are connected by decomposing the synthetic calibration distribution into semantic prompt allocation and conditional visual synthesis; refer to Eq.(5). Specifically, PromptQ consists of two primary components: first, we present *Sensitivity-Driven Prompt Construction (SDPC)*, which reformulates prompt generation as quantization-aware semantic allocation. SDPC estimates capability-module sensitivity from controlled module-wise quantization through behavioral degradation and representation discrepancy,

then allocates more prompts to vulnerable capabilities while retaining a base semantic prior for coverage. Each prompt is further conditioned on the selected capability, its most sensitive module, and the target bit-width, directing calibration toward specific quantization vulnerabilities. Second, we devise *Quantization-Sensitive Image Synthesis (QSIS)*, which uses adaptive prompts and module-level sensitivity signals to optimize the conditional image distribution. Under semantic constraints, QSIS actively increases the normalized representation discrepancy between full-precision and quantized versions of the selected module, synthesizing samples that expose quantization errors for subsequent calibration. Finally, SDPC and QSIS employ quantization sensitivity to guide both *where calibration budget is allocated* and *which visual samples are synthesized*. Extensive experiments on diverse multimodal benchmarks demonstrate that PromptQ consistently improves the performance of quantized LVLMs across different architectures and bit-width settings without requiring real calibration data..

2 Adaptive Calibration for Data-Free Quantization of LVLMs

2.1 Preliminaries

2.1.1 Model Quantization for LVLMs

A pre-trained (LVLM) jointly models visual perception, cross-modal alignment, and language reasoning. Given an image x and an instruction prompt q, a LVLM generates responses by

$$F_\theta(x, q) = F_{\mathrm{LM}}\left(F_{\mathrm{proj}}(F_{\mathrm{vis}}(x)), q\right), \tag{1}$$

where F_{vis}, F_{proj}, and F_{LM} denote the visual encoder, multimodal projector, and language model. θ represents the full-precision parameters of the entire LVLM. Quantization compresses the model by mapping full-precision parameters and intermediate activations into low-bit representations. Given a bit-width b, for a tensor u, the quantization and reconstruction processes are defined as

$$u_q = \mathrm{clamp}\left(\mathcal{R}(u/s_u) + z_u, q_{\min}, q_{\max}\right), \quad \hat{u} = s_u(u_q - z_u), \tag{2}$$

where s_u and z_u denote the scale factor and zero-point of tensor u; $q_{\min}$ and $q_{\max}$ are the quantization bounds; $\mathcal{R}(\cdot)$ denotes the rounding operation. Accordingly, the overall quantization procedure under bit-width b can be formulated as $\theta^Q = Q_b(\theta)$, where $Q_b(\cdot)$ includes the quantization and reconstruction operations. For the resulting low-bit model F_{θ^Q}, the optimization objective is to preserve the functional behavior of the original model:

$$\min_{\theta^Q} D(F_\theta, F_{\theta^Q}), \tag{3}$$

where $D(\cdot, \cdot)$ measures the behavioral discrepancy between the full-precision and quantized models. However, directly minimizing Eq.(3) requires access to the original calibration distribution $p_{real}(x, q)$, which is unavailable in the data-free setting. Therefore, constructing informative synthetic calibration samples becomes essential for recovering the quantized VLM behavior.

2.1.2 Data Synthesis

In data-free quantization (DFQ), a synthetic calibration distribution $p_{syn}(x, q)$ is constructed to approximate the informative regions required for quantization calibration. Given a synthetic calibration set $\mathcal{D}_{syn} = \{(x_i, q_i)\}_{i=1}^N$, the DFQ objective is formulated as

$$\min_{\theta^Q} \mathbb{E}_{(x,q)\sim p_{syn}}\left[D(F_\theta(x, q), F_{\theta^Q}(x, q))\right]. \tag{4}$$

Therefore, the effectiveness of DFQ is largely determined by the quality of the synthetic calibration distribution. We decompose it into semantic prompt allocation and conditional visual synthesis:

$$p_{syn}(x, q) = p_{syn}(x|q)p_{syn}(q), \tag{5}$$

where $p_{syn}(q)$ determines which semantic capabilities are covered during calibration, while $p_{syn}(x|q)$ controls the visual samples generated for each prompt. Eq.(5) reveals two complementary challenges for DFQ of VLMs: optimizing the semantic distribution $p_{\mathrm{syn}}(q)$ to determine *which capabilities should be calibrated*, and optimizing the conditional visual distribution $p_{\mathrm{syn}}(x|q)$ to determine *what informative samples should be synthesized*. PromptQ addresses these two factors separately: SDPC optimizes $p_{\mathrm{syn}}(q)$ through quantization-driven capability allocation, while QSIS optimizes $p_{\mathrm{syn}}(x|q)$ through quantization-sensitive image synthesis.

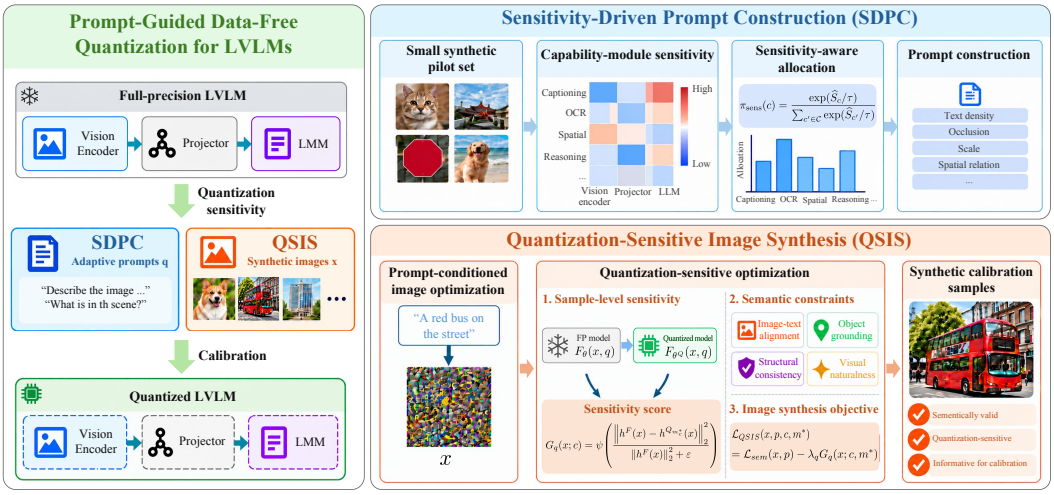

Figure 2: (a) Illustration of the proposed PromptQ. PromptQ first identifies quantization-sensitive capabilities and constructs adaptive prompts through (b) sensitivity-driven prompt construction (SDPC), and then synthesizes capability-aware calibration samples via (c) quantization-sensitive image synthesis (QSIS). The generated samples are finally used for calibration to preserve multimodal capabilities without access to real data.

2.2 Prompting What Matters: Sensitivity-Driven Prompt Construction

As aforementioned, effective data-free quantization requires synthetic calibration samples to cover quantization-sensitive behaviors. However, existing prompt construction strategies typically rely on fixed semantic priors and uniformly sample concepts, ignoring the heterogeneous quantization sensitivity across capability-module pairs. To address this limitation, we propose *Sensitivity-Driven Prompt Construction* (SDPC), which formulates prompt generation as a quantization-aware semantic allocation problem, as shown in Fig. 2(b). Specifically, SDPC identifies quantization-sensitive capability-module pairs and allocates more calibration prompts to vulnerable regions instead of uniformly covering all concepts. Therefore, SDPC directly optimizes the semantic component of the synthetic calibration distribution $p_{\text{syn}}(q)$ in Eq. (5), producing more informative prompts for subsequent quantization-sensitive image synthesis, while improving calibration budget allocation toward quantization-critical capabilities.

Quantization Sensitivity. To characterize the vulnerability of different capability-module pairs, we first introduce the following quantization sensitivity:

Definition 2.1 (Quantization Sensitivity)**.** *For a visual-language capability $c \in \mathcal{C}$ and a quantized module $m \in \mathcal{M}$, the quantization sensitivity $S(c, m)$ quantifies how much the capability performance is affected by quantizing module m, defined as*

$$S(c, m) = \mathbb{E}_{\xi \sim \mathcal{D}_c} \left[\Delta_c \left(F_\theta(\xi), F_{\theta^{Q_m}}(\xi) \right) \right], \tag{6}$$

where $\mathcal{C}$ denotes the set of visual-language capabilities and $\mathcal{M}$ denotes the set of quantized modules. Capabilities are instantiated according to common multimodal instruction behaviors covered by downstream benchmarks, including perception, OCR, spatial understanding, reasoning, and captioning. Each capability corresponds to a semantic category that can be represented by capability specific prompt templates. $\mathcal{D}_c$ represents the data distribution of capability c, and θ^{Q_m} denotes the parameters obtained by quantizing only module m. $\Delta_c(\cdot, \cdot)$ measures the capability-specific behavioral difference between full-precision model F_θ and its partially quantized counterpart $F_{\theta^{Q_m}}$.

Definition 2.1 indicates that measuring only global model sensitivity is insufficient for VLM calibration, since it cannot identify which semantic regions require additional calibration coverage. In practice, directly evaluating Eq.(6) over all capability-module pairs is computationally expensive; hence we approximate the sensitivity using a small synthetic pilot set $\mathcal{D}_c^{\text{pilot}}$ for each capability. Specifically, we combine output-level behavioral degradation and representation-level discrepancy:

$$\widehat{S}_{c,m} = \text{Agg}_{\xi \in \mathcal{D}_c^{\text{pilot}}} \left[D_{\text{beh}}(\xi; m) + D_{\text{rep}}(\xi; m) \right], \tag{7}$$

where $D_{\text{beh}}(\xi; m)$ measures the output degradation after quantizing module m, while $D_{\text{rep}}(\xi; m)$ measures the representation discrepancy between full-precision and quantized models. $\text{Agg}(\cdot)$ aggregates the sensitivity measurements over the pilot samples. Since different modules contribute differently to the overall LVLM behavior, we further aggregate module-level sensitivity into capability-level sensitivity:

$$\widehat{S}_c = \sum_{m \in \mathcal{M}} w_m \widehat{S}_{c,m}, \tag{8}$$

where w_m denotes the importance weight of module m. A larger $\widehat{S}_c$ indicates that capability c is more vulnerable to quantization and should receive more calibration budget, thereby SDPC improves the utilization of a limited calibration budget (refer to Theorem 3.1). The sensitivity estimation is only performed during the lightweight pilot stage. Given N_p pilot samples and $|\mathcal{M}|$ quantized modules, the additional complexity is $O(N_p|\mathcal{M}|)$, and the estimated sensitivity scores are reused throughout the subsequent synthesis process. Therefore, SDPC introduces limited overhead compared with synthetic image generation.

Sensitivity-Aware Prompt Allocation. Based on Eq.(8), SDPC further constructs an adaptive prompt allocation distribution as

$$\pi_{\text{sens}}(c) = \frac{\exp(\widehat{S}_c/\tau)}{\sum_{c' \in \mathcal{C}} \exp(\widehat{S}_{c'}/\tau)}, \tag{9}$$

where τ controls the concentration of the allocation distribution. Capabilities with larger sensitivity scores receive higher sampling probabilities. Although sensitivity-aware allocation improves calibration efficiency, over-emphasizing vulnerable capabilities may reduce semantic diversity. Therefore, we combine the sensitivity-driven distribution with the original capability prior:

$$\pi_{\text{SDPC}}(c) = (1-\rho)\pi_{\text{sens}}(c) + \rho\pi_{\text{base}}(c), \tag{10}$$

where $\pi_{\text{base}}(c)$ denotes the original semantic prior and ρ controls the trade-off between sensitivity awareness and semantic coverage. During prompt construction, we first samples a capability $c_i \sim \pi_{\text{SDPC}}(c)$ and identifies its most sensitive module $m_i^* = \arg\max_{m \in \mathcal{M}} \widehat{S}_{c_i,m}$. The corresponding prompt is generated by a sensitivity-aware prompt distribution $p_i \sim P_{\text{SDPC}}(p|c_i, m_i^*, b)$, where b denotes the target bit-width. Consequently, the induced prompt distribution can be formulated as

$$p_{\text{syn}}(q) = \pi_{\text{SDPC}}(c) P_{\text{SDPC}}(q|c, m^*, b), \tag{11}$$

which jointly considers capability-level allocation and module-level quantization sensitivity.

2.3 Synthesizing Calibration Samples: Quantization-Sensitive Image Synthesis

The above SDPC determines *which semantic capabilities* should be emphasized by optimizing the prompt distribution $p_{\text{syn}}(q)$. However, informative calibration also requires synthesizing visual samples that can effectively expose quantization-induced errors. To this end, we propose ***Quantization-Sensitive Image Synthesis (QSIS)***, as depicted in Fig.2(c), which optimizes the conditional image distribution $p_{\text{syn}}(x|q)$ in Eq.(5). The key idea of QSIS is to search for calibration samples that not only match the semantic prompts but also activate quantization-sensitive behaviors of the LVLM. Specifically, given a capability c and its sensitive module m^* identified by SDPC, QSIS first measures the sample-level quantization sensitivity by evaluating the discrepancy between the full-precision and quantized representations:

$$G_q(x; c, m^*) = \psi\left(\left\| h_{m^*}^F(x) - h_{m^*}^{Q_{m^*}}(x) \right\|_2^2 / (\left\| h_{m^*}^F(x) \right\|_2^2 + \epsilon)\right), \tag{12}$$

where $h_{m^*}^F(x)$ and $h_{m^*}^{Q_{m^*}}(x)$ denote the intermediate representations of module m^* from the full-precision and quantized models. ϵ is a small constant for numerical stability, and $\psi(\cdot)$ normalizes the representation discrepancy into a sensitivity score. A larger $G_q(x; c, m^*)$ indicates that the synthesized sample activates stronger quantization-sensitive behaviors. Unlike adversarial generation that alters semantic labels or maximize prediction errors, QSIS searches within a semantically valid region for samples that expose quantization-induced representation deviations, thereby providing informative calibration signals. To avoid unrealistic or semantically irrelevant samples, QSIS further incorporates semantic constraints, thus the overall image synthesis objective is formulated as

$$\mathcal{L}_{QSIS}(x, p, c, m^*) = \mathcal{L}_{sem}(x, p) - \lambda_q G_q(x; c, m^*), \tag{13}$$

where λ_q is the balancing parameter, and $\mathcal{L}_{sem}(x, p)$ enforces to preserve meaningful visual-language correspondence, *e.g.*, image-text alignment, object grounding, structural consistency and visual naturalness. With Eq.(13) to be minimized, the calibration image x_i^* can be obtained, which further forms the resulting calibration dataset $\mathcal{D}_{syn} = \{(x_i^*, q_i, y_i)\}_{i=1}^N$. $\mathcal{D}_{syn}$ is finally used to calibrate the quantized LVLM through the standard post-training quantization objective. The whole procedure is summarized in Algorithm 1, with detailed optimization steps provided in Appendix A.3.2.

3 Theoretical Analysis

One may wonder why PromptQ benefits data-free calibration of LVLMs. We provide further insights into PromptQ by establishing a finite-budget calibration bound. Formally, let $\xi = (x, q)$ denote a calibration instance, where x and q represent the visual input and capability-aware instruction. Each instance is associated with a latent capability variable $c \in \mathcal{C}$ and a quantization target module $m \in \mathcal{M}$. For each capability c, let $\mathcal{D}_c$ denote the corresponding real-data conditional distribution. The deployment environment is assumed to follow an unknown capability mixture characterized by capability weights $\{\mu_c\}_{c=1}^K$, where $\mu_c \geq 0$, $\sum_{c=1}^K \mu_c = 1$ and $K = |\mathcal{C}|$ be the number of capability categories. Under this case, μ_c represents the probability of observing capability c in the target deployment scenario. Although μ_c determines the real calibration demand, it is unavailable without access to deployment data. To approximate this unknown distribution, PromptQ generates capability-conditioned synthetic samples. Let $\mathcal{D}_c^{\text{syn}}$ denote the synthetic distribution constructed by SDPC and QSIS under capability c. The synthetic calibration allocation is determined by $\pi_{\text{SDPC}}(c)$, which specifies the probability of sampling capability c during calibration.

For a quantized model $\theta^Q \in \mathcal{Q}_b$, where $\mathcal{Q}_b$ denotes the family of models obtained under the target bit-width setting b, we measure the quantization-induced discrepancy by $d_{\theta^Q}(\xi) = D\left(F_\theta(\xi), F_{\theta^Q}(\xi)\right)$ (refer to Eq.(4)). Thus, we define the deployment-level quantization risk as

$$R_{\text{real}}(\theta^Q) = \sum_{c=1}^{K} \mu_c \mathbb{E}_{\xi \sim \mathcal{D}_c}[d_{\theta^Q}(\xi)]. \tag{14}$$

Eq.(14) measures the expected discrepancy between the full-precision and quantized models under the unknown deployment capability distribution. To characterize the discrepancy caused by replacing real calibration data with synthetic samples, we define the capability-wise distribution gap as

$$\text{disc}_c = \sup_{\theta^Q \in \mathcal{Q}_b} \left| \mathbb{E}_{\xi \sim \mathcal{D}_c} d_{\theta^Q}(\xi) - \mathbb{E}_{\xi \sim \mathcal{D}_c^{\text{syn}}} d_{\theta^Q}(\xi) \right|. \tag{15}$$

Eq.(15) directly measures whether synthetic samples preserve the quantization-sensitive behavior of real samples for capability c. Assume that SDPC assigns n_c synthetic samples to capability c, with the total calibration budget $N = \sum_{c=1}^K n_c$. Given the synthetic samples $\{\xi_i^c\}_{i=1}^{n_c}$ drawn from $\mathcal{D}_c^{\text{syn}}$, we define the empirical discrepancy as

$$\widehat{R}_c(\theta^Q) = \frac{1}{n_c} \sum_{i=1}^{n_c} d_{\theta^Q}(\xi_i^c). \tag{16}$$

Eq.(16) represents the calibration error estimated from synthetic samples for capability c.

Theorem 3.1 (Capability-Conditioned Calibration Bound)**.** *Condition on the pilot stage and the frozen synthesis procedure. Assume that samples within each capability are independently drawn from $\mathcal{D}_c^{\text{syn}}$, that $\mathcal{Q}_b$ is fixed before sampling, and that $0 \leq d_{\theta^Q}(\xi) \leq B_c$, where B_c is the maximum discrepancy range for capability c. Let $t_\delta = \log \frac{2K|\mathcal{Q}_b|}{\delta}$. Then, with probability at least $1 - \delta$, for all $\theta^Q \in \mathcal{Q}_b$, we can obtain*

$$R_{\text{real}}(\theta^Q) \leq \underbrace{\sum_c \mu_c \widehat{R}_c(\theta^Q)}_{\textit{Synthetic empirical discrepancy}} + \underbrace{\sum_c \mu_c \, \text{disc}_c}_{\textit{Capability-wise distribution gap}} + \underbrace{\sum_c \mu_c B_c \sqrt{\frac{t_\delta}{2n_c}}}_{\textit{Finite-budget estimation error}}. \tag{17}$$

The proof is deferred to Appendix A.2.1. Theorem 3.1 implies that the real-world quantization discrepancy is governed by three factors: the informativeness of synthetic calibration samples, the

capability-wise mismatch between synthetic and real distributions, and the uncertainty induced by limited calibration budgets.

We *claim* the above analysis to provide theoretical support for our proposal: ***first***, SDPC and QSIS synthesize semantically valid samples that expose quantization-sensitive behaviors, improving the synthetic empirical discrepancy $\sum_c \mu_c \widehat{R}_c(\theta^Q)$. ***Second***, SDPC allocates more calibration samples to sensitivity-critical capabilities, mitigating the finite-budget estimation error $\sum_c \mu_c B_c \sqrt{\frac{t_\delta}{2n_c}}$. ***Third***, QSIS preserves capability-specific quantization patterns, reducing the distribution gap $\sum_c \mu_c \operatorname{disc}_c$. Therefore, PromptQ jointly optimizes sample informativeness, distribution alignment, and budget allocation, to benefit data-free calibration of LVLMs.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS AND DETAILS

Models and Benchmarks. We evaluate PromptQ on Qwen2.5-VL-3B-Instruct Bai et al. (2025) and LLaVA-1.5-7B Liu et al. (2023). Qwen2.5-VL employs a native dynamic-resolution vision encoder with an MLP-based patch merger, whereas LLaVA-1.5 connects a CLIP vision encoder to its language backbone through a multimodal projector. We use the official pre-trained checkpoints without additional task-specific fine-tuning and adopt FP16 models as the reference. Synthetic calibration images are generated at 336×336 resolution for both models. We evaluate quantized models on nine multimodal benchmarks, grouped into three categories: 1) Multimodal Understanding: MMBench Liu et al. (2024b) and MME Fu et al. (2026) measure overall visual-language understanding and multimodal instruction-following ability; 2) Perception & Reasoning: TextVQA Singh et al. (2019), GQA Hudson & Manning (2019), VQAv2 Goyal et al. (2017), ScienceQA Lu et al. (2022) and MMMU Yue et al. (2024) assess visual representation, fine-grained perception, and complex multimodal reasoning capabilities; 3) Hallucination & Reliability: POPE Li et al. (2023b) and VizWiz Gurari et al. (2018) evaluate visual grounding, hallucination robustness, and model reliability under challenging real-world scenarios.

Implementation Details. Our primary configuration quantizes the visual path, comprising the vision encoder and vision-language connector, to 4w8a, 6w6a and 5w5a, while applying 8-bit weight-only quantization to the LLM backbone. The connector corresponds to the patch merger in Qwen2.5-VL and the multimodal projector in LLaVA-1.5. For visual-path quantization, activations use asymmetric per-tensor quantization implemented with an LSQ-style fake quantizer Esser et al. (2019). Weights use asymmetric per-output-channel quantization, with rounding decisions optimized by AdaRound Nagel et al. (2020). Quantization ranges are initialized using an MSE-based observer that minimizes the discrepancy between original and fake-quantized tensors. PromptQ supplies synthetic image–instruction pairs for calibration without accessing real calibration data. For the hyper-parameters, we set ρ in Eq.(10) and λ_q in Eq.(13) to 0.2; the calibration budget N to 128. All experiments are conducted using the PyTorch framework Paszke et al. (2019) and run on an NVIDIA GeForce RTX 5090 GPU. *More details and experimental results are accessible in Appendix A.3 – A.6.*

4.2 MAIN RESULTS

To verify the superiority of our PromptQ, we compare it with representative calibration strategies, including: 1) **Real-Data Calibration** utilizes a budget-matched subset from ShareGPT4V-enhanced COCO Chen et al. (2024a); 2) **Random Noise** calibration removes semantic guidance and image optimization; 3) D4C Zhang et al. (2026), adapted to the LVLM visual pathway by retaining its prompt-guided synthesis and diversity modeling while excluding our sensitivity-aware designs, denoted as **D4C-adapted**; and 4) **Generic Prompt** adopts the same synthesis pipeline as PromptQ but replaces sensitivity-driven prompts with task-agnostic prompts. All these methods uses identical calibration budgets, random seeds, and PTQ procedures for fair comparison.

Table 1 summarizes our findings. *First*, PromptQ consistently achieves the best performance among synthetic-data calibration methods across different models and bit widths. Impressively, PromptQ obtains at most 5.86%, 4.44%, and 4.07% gains of average retention under 4w8a, 6w6a, and 5w5a settings, respectively, confirming the *advantages of targeting the recovery of quantization-vulnerable capabilities* for LVLMs (Sec.1). *Second*, PromptQ still exhibits large improvements under lower-

Table 1: Comparison results on multimodal benchmarks. Hall. & Reli.: Hallucination & Reliability; R: Real data; S: Synthetic samples. *mwna* indicates that weights and activations are quantized to *m*-bit and *n*-bit, respectively. Avg. Ret. denotes the average performance retention relative to the corresponding full-precision model (FP) across nine benchmarks. The best synthetic-data result are reported in **boldface** and green.

Method	Data	Bit Width	Understanding		Perception & Reasoning					Hall. & Reli.		Avg. Ret. (%)
			MMBench	MME	TextVQA	GQA	VQAv2	ScienceQA	MMMU	POPE	VizWiz	
Qwen2.5-VL-3B												
Baseline	-	**FP**	79.64	2132.6	75.49	60.02	80.38	81.36	47.55	87.36	70.25	100.00
Real Data	R		77.49	2104.5	73.97	59.81	78.65	79.77	47.22	87.15	68.98	98.53
Random Noise	S		75.77	2091.6	71.75	59.13	66.79	**79.82**	46.11	85.17	66.79	95.28
D4C-adapted	S	***4w8a***	**77.66**	2096.6	72.40	59.34	77.22	79.62	**47.22**	86.23	67.15	97.57
Generic Prompt	S		75.77	2102.8	**73.74**	59.29	78.36	78.92	46.88	86.49	67.74	97.64
PromptQ (Ours)	S		77.32	**2116.2**	73.72	**59.53**	**78.43**	79.72	46.77	**86.54**	**67.80**	**98.07**
Real Data	R		73.28	2016.3	67.63	58.53	76.69	78.03	45.33	85.67	65.98	94.70
Random Noise	S		72.07	2016.4	67.43	56.74	73.37	76.54	45.66	82.72	64.66	93.00
D4C-adapted	S	***6w6a***	72.50	2004.5	67.64	57.45	74.65	77.44	**46.88**	84.56	64.19	93.91
Generic Prompt	S		73.08	2020.4	69.62	57.57	75.84	77.39	45.11	85.30	65.58	94.44
PromptQ (Ours)	S		**73.88**	**2069.9**	**69.68**	**58.07**	**76.07**	**78.18**	46.11	**85.62**	**65.76**	**95.36**
Real Data	R		69.50	1772.2	59.28	57.43	74.01	76.25	44.33	84.75	62.02	89.88
Random Noise	S		65.80	1773.9	59.78	52.15	66.47	75.06	45.66	80.22	60.06	86.69
D4C-adapted	S	***5w5a***	65.46	1774.1	60.77	54.33	70.51	**75.55**	45.11	83.69	61.05	88.28
Generic Prompt	S		68.64	1839.8	63.16	55.82	71.48	75.21	45.55	84.64	62.44	90.23
PromptQ (Ours)	S		**68.81**	**1899.6**	**63.40**	**55.91**	**71.60**	75.41	**45.77**	**84.88**	**62.56**	**90.76**
LLaVA-1.5-7B												
Baseline	-	FP	62.88	1784.5	57.71	60.41	75.82	65.89	34.56	85.68	54.46	100.00
Real Data	R		60.22	1682.6	55.28	59.89	73.66	64.10	34.33	84.34	55.74	97.73
Random Noise	S		46.39	1416.8	49.45	55.40	69.91	63.70	34.00	75.11	55.36	89.68
D4C-adapted	S	***4w8a***	52.32	1461.6	51.05	57.10	70.66	63.61	**34.11**	82.64	55.25	92.71
Generic Prompt	S		55.67	1583.3	53.09	57.63	70.90	63.36	33.77	82.29	54.99	94.34
PromptQ (Ours)	S		**56.44**	**1595.6**	**54.10**	**58.65**	**71.81**	**63.96**	34.00	**83.32**	**55.81**	**95.54**
Real Data	R		56.35	1561.5	49.85	58.54	70.43	64.15	35.22	83.03	55.27	94.55
Random Noise	S		42.78	1164.7	42.99	49.97	57.14	**65.84**	32.77	71.71	53.65	82.54
D4C-adapted	S	***6w6a***	45.61	1343.2	44.79	52.22	58.66	64.65	**33.77**	73.45	**54.63**	85.68
Generic Prompt	S		41.92	1323.8	45.20	50.04	58.84	65.49	32.22	72.32	52.97	83.77
PromptQ (Ours)	S		**46.90**	**1375.4**	**47.19**	**52.70**	**62.35**	64.99	32.77	**74.59**	54.12	**86.98**
Real Data	R		20.61	931.6	40.23	37.68	39.14	63.56	32.11	47.16	50.76	67.37
Random Noise	S		20.47	822.3	39.16	36.46	37.20	63.51	31.33	49.49	50.16	65.87
D4C-adapted	S	***5w5a***	**20.96**	918.0	39.17	**37.26**	38.09	62.17	31.11	49.49	50.34	66.57
Generic Prompt	S		20.53	897.3	40.18	37.25	38.12	**63.90**	31.67	**49.53**	50.04	66.98
PromptQ (Ours)	S		20.44	**928.0**	**40.48**	37.25	**38.92**	63.56	**32.11**	49.49	**50.98**	**67.60**

bit settings, highlighting the importance of sensitivity-aware calibration when quantization errors become more severe. For example, under Qwen2.5-VL-3B 5w5a quantization, PromptQ improves the retention from 90.23% of Generic Prompt to 90.76%, and consistently improves perception, reasoning, and hallucination-related benchmarks. This indicates that *capability-aware prompt allocation* and *targeted sample synthesis* provide more informative calibration signals than generic semantic coverage (Sec.2.2 and Sec.2.3). *Finally*, PromptQ generalizes well across different LVLM architectures. On LLaVA-1.5-7B, PromptQ achieves at least 1.20%, 1.30%, and 0.62% retention gains under 4w8a, 6w6a, and 5w5a settings, respectively, surpassing existing DFQ baselines by a clear margin. Such fact demonstrates that PromptQ is not tied to a specific model architecture, but *effectively adapts to the heterogeneous sensitivity patterns* inherent in different LVLMs (Sec.1).

4.3 Ablation Studies

4.3.1 Discussion on Each Component of PromptQ

To verify the effectiveness of varied components constituting PromptQ, we perform ablation experiments with Qwen2.5-VL-3B-Instruct under 5w5a from the following variants: *w/o* SDPC; *w/o* QSIS; *w/o* SDPC and QSIS; and our PromptQ. Table 2 suggests that removing both SDPC and QSIS causes the largest degradation (90.97% to 88.77%), indicating that capability-agnostic calibration is insufficient for preserving diverse multimodal abilities under quantization. Besides, *w/o* SDPC

Table 2: Ablation study of different components of PromptQ. FP denotes full precision. Avg. Ret. (%) is the mean performance retention relative to FP across benchmarks. The best quantized results are reported in **boldface** and green.

Variant	MMBench	MME	TextVQA	GQA	ScienceQA	MMMU	POPE	VizWiz	Avg. Ret. (%)
FP	79.64	2132.6	75.49	60.02	81.36	47.55	87.36	70.25	100.00
w/o SDPC and QSIS	68.13	1721.4	59.62	55.51	75.26	45.56	84.01	61.81	88.77
w/o QSIS	68.64	1824.3	62.89	55.73	75.30	45.66	83.85	61.99	90.09
w/o SDPC	68.60	1856.7	63.34	55.70	75.30	45.68	84.05	62.37	90.44
PromptQ (Ours)	**68.81**	**1899.6**	**63.40**	**55.91**	**75.41**	**45.77**	**84.88**	**62.56**	**90.97**

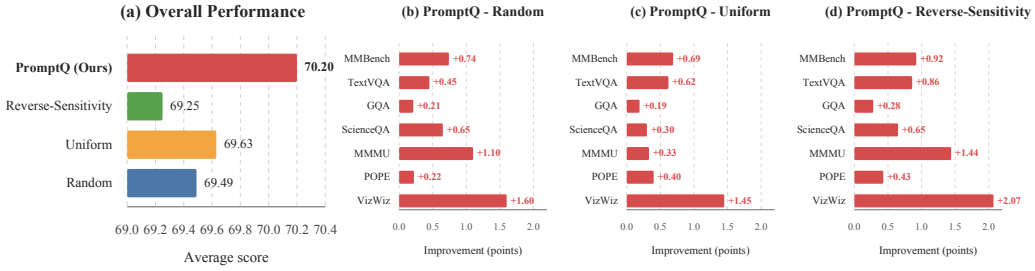

Figure 3: Discussion sensitivity-guided allocation. (a) Overall average performance is the unweighted average of the reported scores on seven benchmarks. Absolute improvements of PromptQ are reported over (b) Random, (c) Uniform, and (d) Reverse-Sensitivity.

decreases the retention to 90.44%, demonstrating the importance of sensitivity-aware prompt allocation for prioritizing quantization-critical capabilities. Without QSIS, the retention drops to 90.09%, showing that sensitive capability identification alone cannot guarantee effective calibration without corresponding visual synthesis. Notably, PromptQ achieves the best retention of 90.97%, verifying the complementary roles of capability-aware allocation (SDPC) and module-targeted synthesis (QSIS), enabling more effective data-free calibration.

4.3.2 Discussion on Sensitivity-Guided Allocation

We further investigate the effectiveness of capability-level sensitivity-guided allocation under a fixed calibration budget. We perform the experiments with Qwen2.5-VL-3B-Instruct by involving several allocation strategies: 1) *Uniform* assigns equal budgets to all capabilities; 2) *Random* samples a non-uniform allocation from $\mathbf{p}^{\text{rand}} \sim \text{Dirichlet}(\alpha\mathbf{1})$ with $\alpha = 1$; 3) *Reverse-Sensitivity* utilizes the same allocation probabilities as our PromptQ but reverses their assignment to capabilities; and 4) *our PromptQ*. Fig.3 illustrates that our PromptQ consistently outperforms Random, Uniform, and Reverse-Sensitivity by 0.71%, 0.57%, and 0.95%, which are attributed to the quality of the effective allocation strategy. The gain over Uniform shows that quantization vulnerability varies across capabilities and equal allocation is suboptimal under limited budgets. The gap with Random indicates that non-uniform allocation alone is insufficient, while the degradation of Reverse-Sensitivity confirms that the importance of correctly matching calibration resources with quantization-sensitive capabilities. The above results further support the *sensitivity-aware allocation principle* of our PromptQ.

5 Conclusion

In this paper, we propose PromptQ, a prompt-guided data-free quantization framework for LVLMs that addresses heterogeneous quantization sensitivity across multimodal capabilities and model components. PromptQ connects two complementary questions: which capabilities are vulnerable to quantization, and how synthetic calibration samples can support their recovery. Specifically, SDPC translates capability-module sensitivity into adaptive prompt allocation and construction, while QSIS synthesizes semantically consistent images that expose errors in the selected modules. By coupling these stages, PromptQ aligns calibration coverage and sample informativeness with quantization vulnerability. Experiments on diverse benchmarks verify the effectiveness of PromptQ, highlighting the importance of capability-aware, model-guided synthesis for DFQ of LVLMs.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with language polishing, improving clarity and organization of the manuscript, and providing suggestions for code debugging and visualization design. We have not used generative AI tools to generate experimental results, create or process datasets, perform the reported evaluations, or replace human judgment in the development and validation of the proposed method.

Additionally, generative AI tools were used for brainstorming and improving the presentation of technical descriptions during manuscript preparation. All AI-assisted content was carefully reviewed and revised by the authors. The authors independently verified all experimental settings, implementation details, mathematical derivations, and scientific claims through manual inspection and experimental validation. Any AI-assisted suggestions were treated only as writing or development aids and were not accepted without author verification.

We take full responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the assistance of generative AI tools.

ETHICS STATEMENT

Our work studies data-free quantization of large vision-language models (LVLMs) and does not involve human subjects, private user data, or the collection of new real-world datasets. The proposed PromptQ framework generates synthetic calibration samples solely for the purpose of improving quantization robustness and does not aim to create or distribute realistic content for human consumption.

We acknowledge that synthetic data generation may inherit biases, limitations, or undesirable behaviors from the underlying pre-trained LVLMs and generative models. Although PromptQ improves calibration efficiency under data-inaccessible settings, the resulting quantized models may still reflect biases present in their original training data. Therefore, additional evaluation and safety considerations are required before deploying quantized LVLMs in sensitive applications.

Our experiments focus on algorithmic improvements for efficient model compression and do not claim to address broader issues of fairness, privacy, or model safety. We follow responsible research practices by clearly documenting the methodology, evaluation protocols, and limitations of the proposed approach to facilitate reproducibility and further investigation.

REPRODUCIBILITY STATEMENT

We make efforts to ensure the reproducibility of our work through detailed descriptions of the proposed framework, experimental protocols, and evaluation settings. The complete algorithmic pipeline of PromptQ, including sensitivity-driven prompt construction and quantization-sensitive sample synthesis, is described in Section 2, while additional implementation details are provided in the appendix.

For theoretical analysis, we provide formal definitions, assumptions, and complete proofs in the appendix. For empirical evaluation, we document the evaluated LVLM architectures, quantization configurations, benchmark datasets, calibration procedures, and comparison settings in the main paper and appendix.

An anonymous implementation will be provided as supplementary materials to facilitate independent verification and future research.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13166–13175. IEEE, 2020.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Kanghyun Choi, Deokki Hong, Noseong Park, Youngsok Kim, and Jinho Lee. Qimera: Data-free quantization with synthetic boundary supporting samples. *Advances in Neural Information Processing Systems*, 34, 2021.

Kanghyun Choi, Hyeyoon Lee, Dain Kwon, SunJong Park, Kyuyeun Kim, Noseong Park, Jonghyun Choi, and Jinho Lee. Mimiq: Low-bit data-free quantization of vision transformers with encouraging inter-head attention similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16037–16045, 2025.

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

Juncan Deng and Kejie Huang. Vlm-ptq: Efficient post-training quantization for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24696–24705, 2026.

Hoang Anh Dung, Cuong Pham, Trung Le, Jianfei Cai, and Thanh-Toan Do. Sharpness-aware data generation for zero-shot quantization. In *Forty-first International Conference on Machine Learning*, 2024.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained 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. *Advances in Neural Information Processing Systems*, 38, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Minjun Kim, Jongjin Kim, and U Kang. Synq: Accurate zero-shot quantization by synthesis-aware fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023a.

Shile Li, Markus Karmann, and Onay Urfalioglu. Joint post-training quantization of vision transformers with learned prompt-guided data generation. *arXiv preprint arXiv:2602.18861*, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4167–4177. IEEE, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Yuhang Li, Youngeun Kim, Donghyun Lee, Souvik Kundu, and Priyadarshini Panda. Genq: Quantization in low data regimes with generative synthetic data. In *European conference on computer vision*, pp. 216–235. Springer, 2024.

Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *European conference on computer vision*, pp. 154–170. Springer, 2022.

Zhikai Li, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Psaq-vit v2: Toward accurate and general data-free quantization for vision transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023c.

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 on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100, 2024.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1325–1334, 2019.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, pp. 7197–7206. PMLR, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Adaptive data-free quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7960–7968, 2023.

Biao Qian, Yang Wang, Yong Wu, and Jungong Han. Selective coupling of decoupled informative regions: Masked attention alignment for data-free quantization of vision transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JeWtUdUk7l.

Akshat Ramachandran, Souvik Kundu, and Tushar Krishna. Clamp-vit: Contrastive data-free learning for adaptive post-training quantization of vits. In *European Conference on Computer Vision*, pp. 307–325. Springer, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Zhenhao Shang, Haizhao Jing, Guoting Wei, Haokui Zhang, Rong Xiao, Jianqing Gao, and Peng Wang. Rethinking practical and efficient quantization calibration for vision-language models. *arXiv preprint arXiv:2602.07899*, 2026.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45472–45496, 2024.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Q-vlm: Post-training quantization for large vision-language models. *Advances in Neural Information Processing Systems*, 37:114553–114573, 2024.

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 *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2020.

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.

Wenlun Zhang, Yunshan Zhong, Zihao Ding, Xinyu Li, and Kentaro Yoshioka. D4c: Data-free quantization for contrastive language-image pre-training models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2978–2987, 2026.

Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for accurate data-free quantization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15653–15662. IEEE, 2021.

Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and Rongrong Ji. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12339–12348, 2022.

Yunshan Zhong, Yuyao Zhou, Yuxin Zhang, Wanchen Sui, Shen Li, Yong Li, Fei Chao, and Rongrong Ji. Semantic alignment and reinforcement for data-free quantization of vision transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12479–12490. IEEE, 2025.

A Appendix

A.1 Related Work

A.1.1 Large Vision-Language Model Quantization

Large vision-language models (LVLMs) have recently achieved remarkable performance on a broad range of multimodal tasks by integrating large language models with vision encoders Li et al. (2023a); Liu et al. (2024a); Bai et al. (2025); Chen et al. (2024b). However, the rapidly increasing model size introduces substantial computational and memory costs, which hinders efficient deployment on resource-constrained devices. Existing approaches for efficient LVLM deployment mainly include pruning, distillation, and quantization. Among them, post-training quantization (PTQ) is particularly attractive because it reduces memory footprint and inference latency without requiring expensive retraining. Recent LVLM quantization methods build upon PTQ techniques developed for large language models and vision models. Representative LLM quantization approaches, including GPTQ Frantar et al. (2022), SmoothQuant Xiao et al. (2023), AWQ Lin et al. (2024), and OmniQuant Shao et al. (2024), improve low-bit inference through weight reconstruction, activation smoothing, and learnable quantization parameter optimization. Extending these ideas to multimodal architectures, recent studies investigate component-aware quantization strategies that consider the heterogeneous behaviors of vision encoders, projection modules, and language models Wang et al. (2024); Deng & Huang (2026); Shang et al. (2026). These works demonstrate that different LVLM components exhibit distinct quantization sensitivity, motivating specialized calibration and quantization strategies for multimodal systems. Despite their effectiveness, existing LVLM quantization methods generally assume access to representative calibration data. Such dependence limits their applicability in scenarios where deployment data are unavailable, privacy-sensitive, or difficult to collect. This motivates the study of data-free calibration strategies for LVLM quantization.

A.1.2 Data-Free Quantization

Data-Free Quantization (DFQ) aims to perform model quantization without access to original training or calibration datasets. Early DFQ methods typically generate synthetic calibration samples by exploiting internal statistics of full-precision models. ZeroQ Cai et al. (2020) introduces zero-shot calibration by synthesizing images that match batch-normalization statistics, while GDFQ Xu et al. (2020) further improves synthetic data generation through generative modeling and knowledge transfer. Subsequent works explore statistical matching, diversity enhancement, and optimization-based sample synthesis to improve calibration effectiveness Nagel et al. (2019); Zhang et al. (2021). With the emergence of Vision Transformers (ViTs), recent DFQ approaches begin to exploit transformer-specific structures. Unlike convolutional networks, ViTs rely heavily on attention interactions and token-level representations, motivating the incorporation of transformer-aware constraints during synthetic calibration. PSAQ-ViT Li et al. (2022) and PSAQ-ViT V2 Li et al. (2023c) leverage patch similarity to separate foreground regions from background regions. CLAMP-ViT Ramachandran et al. (2024) further introduces patch-level contrastive learning to capture meaningful inter-patch relationships. More recently, MimiQ Choi et al. (2025) exploits inter-head attention similarity to improve global structural consistency, while MaskAQ Qian et al. (2026) improve synthetic data generation via informative region coupling. Although these methods substantially improve DFQ performance for vision models, they are mainly designed for image classification tasks and usually construct a unified synthetic calibration distribution. They do not explicitly consider the diverse instruction-following capabilities and modality-specific sensitivity of large vision-language models.

A.1.3 Calibration Data Synthesis

Calibration data synthesis has become an important direction for data-free model compression. Early approaches mainly rely on optimization-based image synthesis, while recent methods increasingly leverage powerful generative models to construct semantically meaningful calibration samples. Diffusion models Ho et al. (2020); Rombach et al. (2022), particularly latent diffusion models, provide controllable image generation capabilities and have been widely adopted for constructing multimodal synthetic data. GenQ Li et al. (2024) demonstrates that text-guided synthetic images can improve calibration quality for Vision Transformers by introducing semantic priors into data-free calibration. Li et al. (2026) learns multi-mode prompts to improve semantic fidelity and sample

diversity. Recently, D4C Zhang et al. (2026) extends synthetic calibration to vision-language models by exploring instruction-aware sample generation and semantic constraints for LVLM quantization.

These studies highlight the importance of synthetic data quality for data-free calibration. However, existing approaches primarily focus on semantic fidelity, diversity, or distribution matching, while the problem of allocating calibration resources according to quantization-sensitive capabilities remains insufficiently explored. PromptQ follows this line of research and further investigates sensitivity-aware capability allocation and conditional synthetic calibration for LVLM quantization, aiming to construct calibration distributions that better preserve multimodal abilities under limited synthetic data budgets.

A.2 Additional Theoretical Analysis

A.2.1 Proof of Capability-Conditioned Calibration Bound

Theorem 3.1 (Capability-Conditioned Calibration Bound). *Condition on the pilot stage and the frozen synthesis procedure. Assume that samples within each capability are independently drawn from $\mathcal{D}_c^{\mathrm{syn}}$, that $\mathcal{Q}_b$ is fixed before sampling, and that $0 \le d_{\theta^Q}(\xi) \le B_c$, where B_c is the maximum discrepancy range for capability c. Let $t_\delta = \log \frac{2K|\mathcal{Q}_b|}{\delta}$. Then, with probability at least $1-\delta$, for all $\theta^Q \in \mathcal{Q}_b$, we can obtain*

$$R_{\mathrm{real}}(\theta^Q) \le \underbrace{\sum_c \mu_c \widehat{R}_c(\theta^Q)}_{\text{Synthetic empirical discrepancy}} + \underbrace{\sum_c \mu_c \,\mathrm{disc}_c}_{\text{Capability-wise distribution gap}} + \underbrace{\sum_c \mu_c B_c \sqrt{\frac{t_\delta}{2n_c}}}_{\text{Finite-budget estimation error}} . \tag{18}$$

Proof. For each capability c, define the population discrepancy on real and synthetic distributions as

$$R_c(\theta^Q) = \mathbb{E}_{\xi \sim \mathcal{D}_c}[d_{\theta^Q}(\xi)], \tag{19}$$

and

$$R_c^{\mathrm{syn}}(\theta^Q) = \mathbb{E}_{\xi \sim \mathcal{D}_c^{\mathrm{syn}}}[d_{\theta^Q}(\xi)]. \tag{20}$$

According to the definition of capability-wise discrepancy,

$$\mathrm{disc}_c = \sup_{\theta^Q \in \mathcal{Q}_b} |R_c(\theta^Q) - R_c^{\mathrm{syn}}(\theta^Q)|, \tag{21}$$

we have

$$R_c(\theta^Q) \le R_c^{\mathrm{syn}}(\theta^Q) + \mathrm{disc}_c \,. \tag{22}$$

Therefore, the deployment-level quantization risk satisfies

$$R_{\mathrm{real}}(\theta^Q) \le \sum_c \mu_c R_c^{\mathrm{syn}}(\theta^Q) + \sum_c \mu_c \,\mathrm{disc}_c \,. \tag{23}$$

Next, we bound the difference between the synthetic population discrepancy and its empirical estimation. For each capability c, since

$$0 \le d_{\theta^Q}(\xi) \le B_c, \tag{24}$$

Hoeffding's inequality gives

$$\Pr\left(\left| R_c^{\mathrm{syn}}(\theta^Q) - \widehat{R}_c(\theta^Q) \right| > B_c \sqrt{\frac{t}{2n_c}} \right) \le 2e^{-t}. \tag{25}$$

Applying the union bound over all capabilities and all candidate quantized models in $\mathcal{Q}_b$ yields

$$t = \log \frac{2K|\mathcal{Q}_b|}{\delta}. \tag{26}$$

Therefore, with probability at least $1 - \delta$,

$$R_c^{\mathrm{syn}}(\theta^Q) \le \widehat{R}_c(\theta^Q) + B_c\sqrt{\frac{t_\delta}{2n_c}} \tag{27}$$

holds simultaneously for all c and $\theta^Q \in \mathcal{Q}_b$.

Substituting this inequality into Eq. equation 23 gives

$$R_{\mathrm{real}}(\theta^Q) \le \sum_c \mu_c \widehat{R}_c(\theta^Q) + \sum_c \mu_c \,\mathrm{disc}_c + \sum_c \mu_c B_c \sqrt{\frac{t_\delta}{2n_c}}, \tag{28}$$

which completes the proof. □

A.2.2 Sensitivity-Aware Capability Allocation

Theorem 3.1 shows that the finite-budget uncertainty decreases as the number of samples n_c increases. Therefore, under a fixed calibration budget, an effective allocation strategy should prioritize capabilities that are more sensitive to quantization.

Let

$$\widehat{S}_c = \sum_m w_m \widehat{S}_{c,m} \tag{29}$$

denote the aggregated sensitivity score estimated in the pilot stage. SDPC determines the capability sampling distribution by solving the entropy-regularized optimization problem:

$$\max_{\pi \in \Delta_K} \sum_c \pi(c)\widehat{S}_c + \tau H(\pi), \tag{30}$$

where

$$H(\pi) = -\sum_c \pi(c) \log \pi(c) \tag{31}$$

is the Shannon entropy and

$$\Delta_K = \{\pi : \pi(c) \ge 0, \sum_c \pi(c) = 1\} \tag{32}$$

is the capability probability simplex.

The Lagrangian of Eq. equation 30 is

$$\mathcal{L} = \sum_c \pi(c)\widehat{S}_c - \tau \sum_c \pi(c) \log \pi(c) + \lambda(\sum_c \pi(c) - 1). \tag{33}$$

Taking derivative with respect to $\pi(c)$ gives

$$\widehat{S}_c - \tau(1 + \log \pi(c)) + \lambda = 0. \tag{34}$$

Therefore,

$$\pi(c) \propto \exp(\widehat{S}_c/\tau), \tag{35}$$

which leads to

$$\pi_{\text{sens}}(c) = \frac{\exp(\widehat{S}_c/\tau)}{\sum_{c'} \exp(\widehat{S}_{c'}/\tau)}. \tag{36}$$

This derivation shows that SDPC performs a sensitivity–diversity trade-off: high-sensitivity capabilities receive more calibration samples, while the entropy regularization prevents the allocation from collapsing to a small subset of capabilities.

A.2.3 Coverage Preservation of SDPC

To avoid insufficient calibration coverage for low-frequency capabilities, PromptQ further combines the sensitivity-aware distribution with a base prior:

$$\pi_{\text{SDPC}}(c) = (1-\rho)\pi_{\text{sens}}(c) + \rho\pi_{\text{base}}(c). \tag{37}$$

For any capability c,

$$\pi_{\text{SDPC}}(c) \geq \rho\pi_{\text{base}}(c). \tag{38}$$

Hence, after N independent capability selections, the probability that capability c receives no calibration sample is

$$\Pr[n_c = 0] = (1 - \pi_{\text{SDPC}}(c))^N. \tag{39}$$

Using $1 - x \leq e^{-x}$ gives

$$\Pr[n_c = 0] \leq e^{-N\rho\pi_{\text{base}}(c)}. \tag{40}$$

Therefore, prior mixing provides an explicit coverage guarantee while preserving sensitivity-aware allocation.

A.2.4 QSIS as Semantically Constrained Sensitivity Optimization

QSIS aims to identify synthetic samples that expose quantization-sensitive behaviors while maintaining semantic consistency.

For a prompt q, define the feasible semantic region:

$$\mathcal{X}_a(q) = \{x : \mathcal{L}_{\text{sem}}(x, q) \leq a\}. \tag{41}$$

The QSIS objective

$$\mathcal{L}_{\text{QSIS}} = \mathcal{L}_{\text{sem}}(x, q) - \lambda_q G_q(x; c, m) \tag{42}$$

can be interpreted as the Lagrangian relaxation of

$$\max_{x \in \mathcal{X}_a(q)} G_q(x; c, m). \tag{43}$$

Therefore, QSIS searches for samples inside a semantically valid region that maximize the sensitivity of the selected quantized module.

To connect G_q with the calibration discrepancy, assume that

$$0 \leq \Delta_{c,m}(x, q) \leq \kappa_{c,m} G_q(x; c, m) + \zeta_{c,m}, \tag{44}$$

where $\Delta_{c,m}$ denotes the quantization discrepancy of module m under capability c.

Then for any synthetic distribution supported on $\mathcal{X}_a(q)$,

$$\mathbb{E}[\Delta_{c,m}(x, q)] \leq \kappa_{c,m} \sup_{x \in \mathcal{X}_a(q)} G_q(x; c, m) + \zeta_{c,m}. \tag{45}$$

This result indicates that maximizing the QSIS objective increases the probability of generating samples with large quantization sensitivity, which improves the calibration signal while preserving semantic validity.

The guarantee is conditional on the proxy relationship above; it does not claim that maximizing G_q alone always improves downstream task performance.

A.3 More Experimental Details

A.3.1 Evaluation Metrics

We follow the official metric of each benchmark. For POPE, we report accuracy and F1 score over the adversarial, popular, and random subsets. For MME, we report the total score together with the perception and cognition subtotals. TextVQA, VQAv2, and VizWiz use VQA-style soft accuracy; GQA uses normalized exact match; MMBench and ScienceQA use multiple-choice accuracy; and MMMU uses its official aggregate accuracy. Full-precision and quantized models are evaluated with identical task prompts and deterministic generation. The extended evaluator validates labelled splits, rejects empty or malformed runs, and restores quantized checkpoints only after verifying that fake quantizers are enabled and observers are disabled.

A.3.2 Optimization Details

The overall optimization process of PromptQ follows three sequential stages, as summarized in Algorithm 1: sensitivity-driven prompt construction, quantization-sensitive image synthesis, and post-training quantization.

Stage 1: Sensitivity-driven prompt construction. For each capability $c \in \mathcal{C}$, we construct a small synthetic pilot set $\mathcal{D}_c^{\text{pilot}}$, where each pilot set is synthesized from random pixels using capability-specific prompt templates and semantic objectives, without real images or quantization-discrepancy optimization. We compare the full-precision model F_θ with $F_{\theta^{Q_m}}$, in which only module m is quantized, and aggregate behavioral and representation discrepancies to estimate $\widehat{S}_{c,m}$. The module-weighted scores yield $\widehat{S}_c$ through Eq.(8), which determines π_{SDPC} through sensitivity-based allocation and semantic-prior mixing in Eq.(10). For each calibration instance, we sample $c_i \sim \pi_{\text{SDPC}}$, select $m_i^* = \arg\max_{m \in \mathcal{M}} \widehat{S}_{c_i,m}$, and construct $q_i \sim P_{\text{SDPC}}(q \mid c_i, m_i^*, b)$. The resulting prompts and sensitivity estimates are fixed during synthesis.

Stage 2: Quantization-sensitive image synthesis. Given (c_i, m_i^*, q_i), we initialize a synthetic image and minimize $\mathcal{L}_{QSIS}$ in Eq.(13). The semantic term preserves correspondence with q_i, while the sensitivity term increases the normalized representation discrepancy at module m_i^*. Throughout synthesis, the full-precision and partially quantized models remain fixed; only the synthetic image is optimized. Upon reaching the synthesis stopping criterion, we retain x_i^* and construct $\mathcal{D}_{\text{syn}} = \{(x_i^*, q_i)\}_{i=1}^N$.

Stage 3: Post-training quantization. We fix $\mathcal{D}_{\text{syn}}$ and calibrate the target modules at bit-width b using Eq.(4), obtaining F_{θ^Q}. The synthetic pairs support quantization-parameter calibration and, where applicable, reconstruction under the selected PTQ procedure. Unlike the partially quantized probing models, the final model includes all target modules. Thus, QSIS exposes selected-module errors through image synthesis, whereas PTQ reduces model discrepancy on the resulting calibration set. No further prompt or image optimization is performed during PTQ.

Algorithm 1 PromptQ: Prompt-Guided Data-Free Quantization of LVLMs

Require: Full-precision LVLM F_θ; capabilities $\mathcal{C}$; target modules $\mathcal{M}$; bit-width b; calibration budget N; semantic prior π_{base}.
Ensure: Calibrated quantized LVLM $F_{\theta Q}$.

Stage 1: Sensitivity-Driven Prompt Construction (SDPC)

Construct a small synthetic pilot set $\mathcal{D}_c^{\text{pilot}}$ for each $c \in \mathcal{C}$.
Initialize $F_{\theta Q_m}$ for each $m \in \mathcal{M}$ by quantizing only module m at bit-width b.
for each capability $c \in \mathcal{C}$ **do**
  for each module $m \in \mathcal{M}$ **do**
    Evaluate $D_{\text{beh}}(\xi; m)$ and $D_{\text{rep}}(\xi; m)$ on $\xi \in \mathcal{D}_c^{\text{pilot}}$.
    Estimate $\widehat{S}_{c,m}$ by aggregating $\lambda_{\text{beh}} D_{\text{beh}} + \lambda_{\text{rep}} D_{\text{rep}}$ over the pilot set.
  end for
  Aggregate $\widehat{S}_c$ using module weights w_m according to Eq.(8).
end for
Compute π_{sens} by applying softmax to $\{\widehat{S}_c/\tau\}_{c\in\mathcal{C}}$.
Combine π_{sens} and π_{base} to obtain π_{SDPC} using Eq.(10).
for $i = 1, \ldots, N$ **do**
  Sample $c_i \sim \pi_{\text{SDPC}}$.
  Select $m_i^* \leftarrow \arg\max_{m\in\mathcal{M}} \widehat{S}_{c_i,m}$.
  Construct $q_i \sim P_{\text{SDPC}}(q \mid c_i, m_i^*, b)$.
end for

Stage 2: Quantization-Sensitive Image Synthesis (QSIS)

Initialize $\mathcal{D}_{\text{syn}} \leftarrow \varnothing$; freeze F_θ and the probing models $F_{\theta Q_m}$.
for $i = 1, \ldots, N$ **do**
  Initialize a synthetic image x_i conditioned on q_i.
  while the synthesis stopping criterion is not met **do**
    Evaluate the normalized representation discrepancy $G_q(x_i; c_i, m_i^*)$ at module m_i^*.
    Evaluate the semantic constraint loss $\mathcal{L}_{sem}(x_i, q_i)$.
    Update x_i to minimize $\mathcal{L}_{QSIS}(x_i, q_i, c_i, m_i^*)$ in Eq.(13).
  end while
  Set $x_i^* \leftarrow x_i$ and append (x_i^*, q_i) to $\mathcal{D}_{\text{syn}}$.
end for

Stage 3: Post-Training Quantization

Calibrate F_θ at bit-width b using $\mathcal{D}_{\text{syn}}$ and the PTQ objective in Eq.(4), obtaining $F_{\theta Q}$.
return $F_{\theta Q}$.

A.4 Sensitivity Analysis Across Models and Bit Widths

A.4.1 Model-Specific Sensitivity

To investigate whether synthetic calibration samples generated for one LVLM can be directly transferred to another model, we evaluate the model-specific sensitivity of PromptQ. Specifically, we exchange the source model (*e.g.*, Qwen2.5-VL-3B-Instruct, LLaVA-1.5-7B) used for calibration sample generation and the target quantized model (*e.g.*, Qwen2.5-VL-3B-Instruct, LLaVA-1.5-7B) , while keeping the quantization setting unchanged, denoted as Qwen $\rightarrow$ LLaVA and LLaVA $\rightarrow$ Qwen.

The results form Table 3 reveal that calibration samples exhibit strong model specificity. For LLaVA under W4A8, using LLaVA-specific samples achieves an average performance of 58.04%, while transferring Qwen-generated samples reduces the performance to 56.17%. A similar trend is observed under W6A6, where cross-model transfer decreases the average score from 53.32% to 51.46%. The degradation is particularly evident on capability-sensitive tasks such as TextVQA, indicating that *different LVLMs contain distinct vulnerable multimodal behaviors that require targeted data-free calibration* (Sec.1).

In contrast, Qwen exhibits relatively stable cross-model transfer, with minor degradation when using LLaVA-generated samples (70.20% vs. 70.02% under W4A8), suggesting that larger pre-trained LVLMs may share broader multimodal representations, but model-specific calibration still provides the optimal performance.

The above results confirm that quantization sensitivity is not solely determined by the quantization configuration, but also depends on the underlying model architecture and capability distribution.

Table 3: Model-specific sensitivity analysis across different source-target LVLM pairs. Calibration samples are most effective when generated for the target model, indicating architecture-dependent quantization sensitivity. The best results under each bit-width setting are reported in **boldface**.

Settings	Bit-width	MMBench	TextVQA	GQA	ScienceQA	MMMU	POPE	VizWiz	Avg.
LLaVA → LLaVA	W4A8	**56.44**	**54.10**	58.65	**63.96**	**34.00**	83.32	**55.81**	**58.04**
Qwen → LLaVA	W4A8	57.04	40.50	**59.49**	63.66	33.33	**83.65**	55.54	56.17
Qwen → Qwen	W4A8	77.32	**73.72**	59.53	**79.72**	46.77	**86.54**	**67.80**	**70.20**
LLaVA → Qwen	W4A8	**77.49**	73.23	**59.68**	79.62	**47.11**	86.20	66.78	70.02
LLaVA → LLaVA	W6A6	**46.90**	**47.19**	52.70	64.99	32.77	74.59	54.12	**53.32**
Qwen → LLaVA	W6A6	50.77	23.14	**54.38**	**65.10**	**33.67**	**78.42**	**54.73**	51.46
Qwen → Qwen	W6A6	73.88	**69.68**	58.07	**78.18**	**46.11**	**85.62**	**65.76**	**68.19**
LLaVA → Qwen	W6A6	**74.31**	68.81	**58.09**	77.24	45.89	84.61	64.38	67.62

Therefore, PromptQ performs calibration sample construction conditioned on the target model to better capture model-specific sensitivity patterns.

A.4.2 Bit-Width-Specific Sensitivity

We further investigate whether calibration samples generated under different quantization configurations can be transferred across varied bit-width settings. Specifically, we obtain synthetic calibration samples optimized for one bit-width configuration (*e.g.*, W4A8) and evaluate their effectiveness on other quantization settings (*e.g.*, W6A6), denoted as W4A8 → W6A6.

In Table 4, we observe an asymmetric transferability pattern across different bit-widths. Notably, calibration samples generated under more aggressive quantization settings generally provide stronger transferability to higher bit-width models. For example, W5A5-generated samples achieve 70.31% retention on W4A8 (W5A5 → W4A8) and 68.35% on W6A6 (W5A5 → W6A6), slightly outperforming the corresponding bit-width-specific calibration results, *i.e.*, 70.20% from W4A8 → W4A8 and 68.19% from W6A6 → W6A6. The results indicate that lower-bit quantization exposes stronger quantization perturbations during sample synthesis, enabling the generated samples to capture more comprehensive sensitivity patterns, which may remain effective under relaxed quantization constraints.

In contrast, samples generated from higher bit-width settings provide limited benefits when transferred to lower bit-width quantization. For instance, W4A8 and W6A6 samples (W4A8 → W5A5 and W6A6 → W5A5) achieve only 64.40% and 65.01% retention under W5A5 quantization, respectively, which are lower than the calibration results (65.25%) from W5A5 samples (W5A5 → W5A5). Such fact suggests that higher-bit calibration does not sufficiently expose the severe quantization-sensitive regions required by more aggressive low-bit deployment.

The above results reveal that quantization sensitivity is closely coupled with bit-width constraints, and calibration distributions optimized for stronger quantization perturbations can provide broader coverage across bit-width settings. Therefore, PromptQ explicitly incorporates the target quantization configuration during sample construction (Sec.2.2), allowing the synthesized calibration distribution to capture the sensitivity patterns required by different deployment scenarios.

A.5 Parameter Studies

To examine the prior mixing coefficient ρ in Eq.(10) and the balancing weight λ_q in Eq.(13), we perform the parameter studies with Qwen2.5-VL-3B-Instruct under w4a8 setting. The evaluation performance (average score) is measured by the unweighted average of the reported scores on seven benchmarks: MMBench, TextVQA, GQA, ScienceQA, MMMU, POPE, and VizWiz. We vary each parameter individually while keeping the quantization configuration, calibration budget, reconstruction schedule, and remaining settings fixed.

Prior mixing coefficient ρ. In Eq.(10), ρ is utilized to balance sensitivity-driven capability allocation and the base semantic prior. Fig. 4 shows relatively stable performance over the tested range, with the highest average score of 70.39 at $\rho = 0.3$. This suggests that retaining prior coverage complements

Table 4: Bit-width-specific sensitivity analysis. Settings indicate that calibration samples synthesized under one bit-width setting are applied to another quantization configuration. The asymmetric transferability reveals that stronger quantization constraints provide richer sensitivity signals. The best results under each target bit-width are reported in **boldface**.

Settings	MMBench	TextVQA	GQA	ScienceQA	MMMU	POPE	VizWiz	Avg.
W4A8 → W4A8	77.32	73.72	59.53	79.72	46.77	86.54	67.80	70.20
W6A6 → W4A8	76.71	**73.96**	**59.83**	79.47	**47.11**	86.51	**68.06**	70.24
W5A5 → W4A8	**77.57**	73.69	59.43	**79.82**	46.88	**86.76**	68.03	**70.31**
W4A8 → W6A6	73.54	69.34	58.08	77.24	46.11	**85.63**	65.58	67.93
W6A6 → W6A6	73.88	69.68	58.07	**78.18**	46.11	85.62	65.76	68.19
W5A5 → W6A6	**74.57**	**69.83**	**58.34**	77.98	**46.22**	85.60	**65.92**	**68.35**
W4A8 → W5A5	68.04	62.90	55.04	**75.75**	44.22	83.46	61.37	64.40
W6A6 → W5A5	68.55	**63.68**	55.62	75.55	45.55	83.73	62.36	65.01
W5A5 → W5A5	**68.81**	63.40	**55.91**	75.41	**45.77**	**84.88**	**62.56**	**65.25**

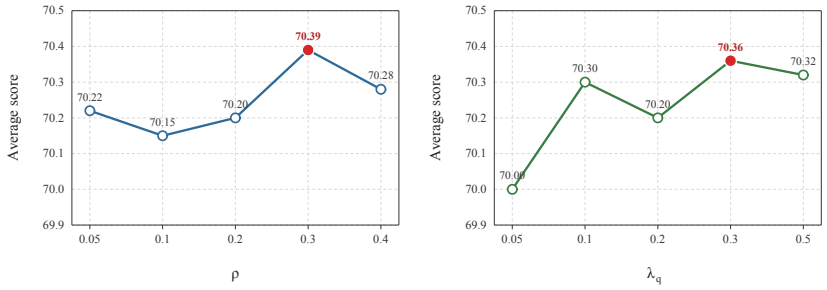

Figure 4: Parameter studies about hyper-parameters ρ and λ_q. The evaluation performance (average score) is measured by the unweighted average of the reported scores on seven benchmarks: MMBench, TextVQA, GQA, ScienceQA, MMMU, POPE, and VizWiz.

the emphasis on quantization-vulnerable capabilities, where smaller values place greater emphasis on sensitivity-driven allocation, potentially narrowing semantic coverage, whereas larger values weaken this targeted allocation. Such fact further confirms the necessity of balancing sensitivity awareness with broader capability coverage.

Balancing weight λ_q. In Eq.(13), λ_q controls the strength of quantization-error exposure relative to semantic constraints. Fig. 4 indicates that the highest average score of 70.36 is obtained at $\lambda_q = 0.3$. Lower performance at smaller values suggests that stronger quantization-sensitive guidance can improve calibration informativeness. However, further increasing λ_q provides no additional benefit, indicating that greater emphasis on representation discrepancy does not necessarily improve downstream performance, which supports balancing error exposure with semantic preservation.

A.6 Calibration Budget and Efficiency Analysis

We further investigate the calibration efficiency of PromptQ by varying the number of synthesized samples N and measuring the wall-clock time. The experiments are conducted on Qwen2.5-VL-3B-Instruct and LLaVA-1.5-7B. All experiments are conducted under identical quantization settings and hardware configurations. The evaluation performance (Avg.) is measured by the unweighted average of the reported scores on seven benchmarks: MMBench, TextVQA, GQA, ScienceQA, MMMU, POPE, and VizWiz. Table 5 and Fig.5 indicate that:

1) First, PromptQ achieves substantially higher calibration efficiency under limited sample budgets. With only $N = 32$ synthetic samples, PromptQ obtains an average score of 68.13, already surpassing D4C-adapted and Generic Prompt using four times more samples ($N = 128$), which achieve 67.24 and 67.66, respectively. This confirms that sensitivity-aware prompt allocation and quantization-

Table 5: Calibration budget and efficiency analysis of different synthetic calibration methods.

Method	Synthesis Time (h)	Calibration Time (h)	Total Time (h)
D4C-adapted	5.10	0.42	5.52
Generic Prompt	6.51	0.42	6.93
PromptQ ($N = 32$)	1.84	0.39	2.23
PromptQ ($N = 64$)	3.67	0.40	4.07
PromptQ ($N = 128$)	7.23	0.42	7.65
PromptQ ($N = 256$)	13.94	0.46	14.40

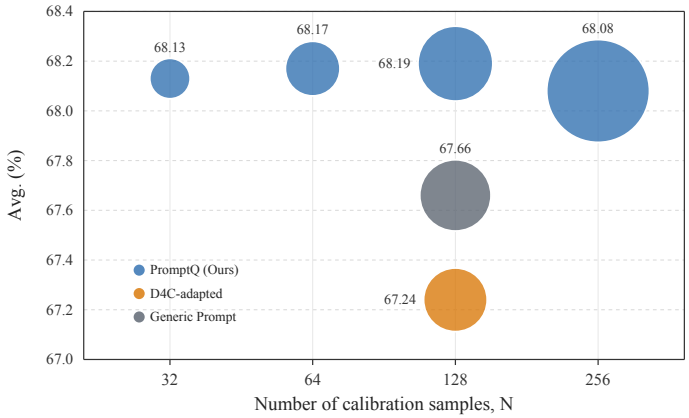

Figure 5: Visual analysis on calibration budget and efficiency. The smaller circle indicates the less total time (higher calibration efficiency).

sensitive image synthesis effectively improve the information density of each calibration sample, reducing the reliance on large synthetic calibration sets.

2) Second, PromptQ exhibits stable performance as the calibration budget increases. Increasing N from 32 to 256 only changes the average performance from 68.13 to 68.08, indicating that PromptQ can already capture the most quantization-critical capability distributions with a small calibration budget. In contrast, conventional synthetic calibration methods require more samples to compensate for the lack of capability-aware guidance.

3) Finally, PromptQ provides a *favorable efficiency-performance trade-off*. Although larger calibration sets naturally introduce additional synthesis cost, PromptQ ($N = 32$) reduces the total calibration pipeline time to 2.23 hours, achieving competitive performance with only a fraction of the synthesis cost required by existing methods. These results verify that PromptQ is particularly suitable for data-inaccessible LVLM quantization scenarios, where calibration data generation cost and storage constraints are critical considerations.

A.7 Visualization of Sensitivity-Aware Allocation

To further verify the merits of PromptQ, we further perform visual analysis on sensitivity-aware allocation with Qwen2.5-VL-3B-Instruct. Fig. 6 visualizes how SDPC translates capability-module sensitivity into prompt allocation. Notably, the heatmap reports normalized sensitivity utilities, where negative values indicate below-average sensitivity after normalization rather than negative quantization degradation. Instead, SDPC aggregates module-wise sensitivity with module-specific weights and further interpolates the resulting capability sensitivity with a base prior to preserve coverage. Fig. 6(a) shows that the sensitivity varies substantially across both capabilities and vision submodules. For instance, OCR is more sensitive to the early vision blocks, whereas reasoning shows higher sensitivity to the late vision/merger stage, confirming that quantization vulnerability is capability- and module-dependent. Accordingly, Fig. 6(b) shows a clear departure from uniform allocation: Spatial understanding and Reasoning receive larger budgets, while less critical capabilities

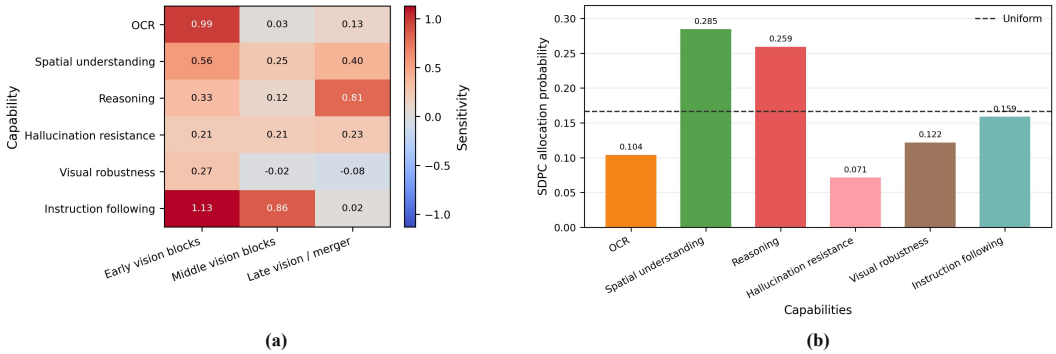

Figure 6: Visualization of sensitivity-aware prompt allocation. (a) Normalized capability-module sensitivity utilities estimated during the pilot stage, showing heterogeneous quantization vulnerability across capabilities and vision submodules. Negative values indicate below-average normalized sensitivity rather than negative quantization degradation. (b) SDPC prompt allocation under the same total calibration budget, with the dashed line denoting uniform allocation. The final allocation is derived from module-weighted capability sensitivity and base-prior interpolation, rather than a simple average of the heatmap values.

receive fewer samples, which demonstrates that SDPC converts heterogeneous quantization sensitivity into a coverage-aware allocation policy, providing direct evidence for the proposed sensitivity-driven prompt construction.