VISTA: ACCURATE CLIP MODEL INVERSION FRAMEWORK WITH SEMANTIC TARGETS AND APPEARANCE COVERAGE FOR DATA-FREE QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

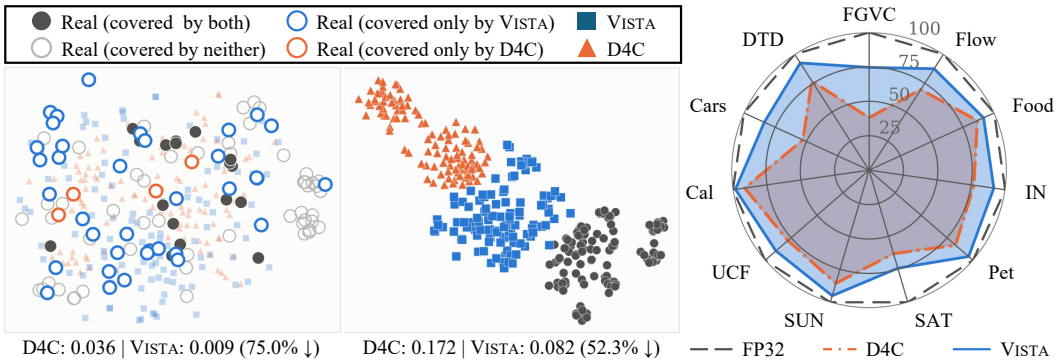
How can we quantize a CLIP model without access to real data while preserving its generalization capability? Quantization reduces the memory and computational cost of large models, but most methods require real data for calibration. Data-free quantization (DFQ) eliminates the need for real data, yet most existing methods synthesize images conditioned on predefined class labels as inversion targets. In contrast, CLIP models support open-vocabulary recognition without a fixed label space. Therefore, DFQ for CLIP models requires semantically diverse text prompts to extract broad semantic knowledge encoded in the model. A recent method manually constructs these targets, covering a narrow semantic range and degrading the generalization of the quantized model. In this paper, we propose VISTA, an accurate CLIP inversion framework for DFQ. VISTA broadens the semantic coverage of inversion targets by systematically sampling visual prompts from the lexical hierarchy of WordNet. VISTA captures semantic relationships among prompts by replacing one-hot supervision with identity-preserving soft targets derived from pairwise prompt similarities. VISTA expands the visual coverage of synthetic images by decorrelating the appearance components that the target prompts do not explain, while enforcing view consistency to preserve semantic alignment. Experimental results show that VISTA improves the average accuracy of quantized CLIP models by up to 9.45%p over existing DFQ methods and better preserves the generalization capability of CLIP models across various CLIP backbones, bit-widths, and downstream datasets.

1 INTRODUCTION

How can we quantize a CLIP model without access to real data while preserving its generalization capability? Recent deep learning models achieve remarkable success across visual recognition tasks, but their growing size and computational cost hinder deployment in resource-constrained environments. Model compression reduces this burden (Cheng et al., 2017; Deng et al., 2020) through pruning (Han et al., 2015; He & Xiao, 2024), quantization (Wu et al., 2016; Rokh et al., 2023), knowledge distillation (Hinton et al., 2015; Gou et al., 2021), and low-rank approximation (Zhang et al., 2015). Among them, quantization represents weights and activations at low precision, reducing the memory footprint and enabling hardware-friendly computation. Most quantization methods require real data for calibration to mitigate accuracy degradation (Nagel et al., 2020; Li et al., 2021). However, privacy and security concerns often restrict access to such data in real-world deployment settings (Zhang et al., 2021). Data-free quantization (DFQ) removes this dependency and enables efficient deployment in data-constrained settings (Cai et al., 2020; Kim et al., 2025a).

However, most existing DFQ methods do not directly apply to CLIP models because they synthesize images conditioned on the predefined class labels of a closed-set classifier as inversion targets (Li et al., 2022; Kim et al., 2025b). In contrast, CLIP models align images and text prompts in a shared embedding space through contrastive learning, enabling open-vocabulary recognition without a fixed label space (Radford et al., 2021). Therefore, DFQ for CLIP models requires semantically diverse text prompts to serve as inversion targets. As in DFQ for closed-set classifiers, the synthetic images for each target should also be visually varied to yield the range of activation patterns needed for calibration. A recent DFQ method for CLIP models (Zhang et al., 2025) curates its prompt

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



(a) Prompt coverage (b) Real-image alignment (c) DFQ accuracy across various datasets

Figure 1: Comparison of D4C and VISTA (proposed). (a) VISTA covers a broader portion of the reference prompt embeddings. (b) Its synthetic image embeddings are more closely aligned with the real image embeddings. The lower MMD scores in (a) and (b) quantify these improvements. (c) VISTA better preserves full-precision accuracy than D4C under W8/A8 quantization of ViT-B/16.

set manually, but manual curation lacks a systematic mechanism to scale to the open-vocabulary semantic space of CLIP models. This curated set therefore covers only a limited semantic range, and the resulting synthetic calibration data capture limited semantic knowledge, degrading the accuracy of quantized models (see Figure 1).

In this paper, we propose VISTA (in**V**ers**I**on framework with **S**emantic **T**argets and **A**ppearance coverage), an accurate CLIP inversion framework for DFQ. VISTA broadens the semantic coverage of inversion targets by systematically sampling visual prompts from the lexical hierarchy of WordNet. VISTA captures semantic relationships among prompts by replacing one-hot supervision with identity-preserving soft targets derived from pairwise prompt similarities. VISTA increases visual diversity of the synthetic images by decorrelating the components that the target prompts leave unexplained, while maintaining consistency across different views to preserve semantic alignment. Figure 1 summarizes the representation- and quantization-level improvements achieved by VISTA. Figure 1(a) shows that prompt embeddings of VISTA cover a larger portion of the reference prompt embeddings. Figure 1(b) shows that synthetic image embeddings of VISTA are more closely aligned with the real image embeddings. Experimental results show that VISTA improves the accuracy of quantized CLIP models by up to 9.45%p over existing DFQ methods across various CLIP backbones, bit-widths, and downstream datasets, and better preserves their zero-shot generalization capability (see Figure 1(c)).

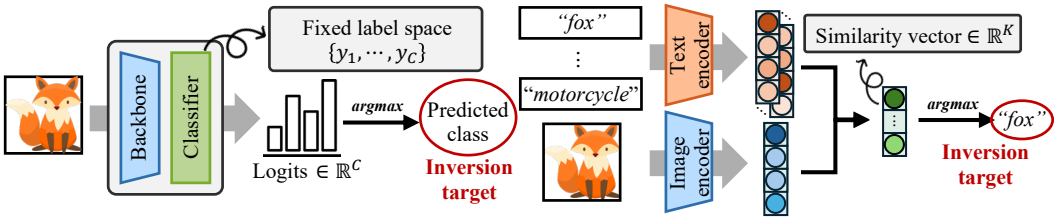
Our contributions are summarized as follows:

- **Problem analysis.** We identify three key challenges in existing DFQ for CLIP models: (1) manually constructed inversion targets with limited semantic coverage, (2) one-hot supervision that ignores semantic relationships among prompts, and (3) limited visual diversity within each target.
- **Framework.** We propose VISTA, an accurate CLIP inversion framework that addresses these limitations. VISTA broadens the semantic coverage of inversion targets, captures semantic relationships among prompts, and promotes visual diversity while maintaining semantic alignment.
- **Experiments.** We experimentally show that VISTA improves the average accuracy of quantized CLIP models over competitors by up to 9.45%p across various backbones, bit-widths, and datasets.

In the rest of the paper, we provide the preliminaries and related works in Section 2, describe the proposed method in Section 3, evaluate the proposed method and competitors on various datasets in Section 4, and conclude this work in Section 5. All of our implementations are available at `https://anonymous.4open.science/r/Vista-F19B`.

2 PRELIMINARIES AND RELATED WORKS

We provide background on data-free quantization (DFQ) and CLIP models. We then formulate the DFQ problem for CLIP models. We summarize the notation used in this paper in Appendix A.



(a) Closed-set classifier (b) CLIP model

Figure 2: Comparison of inversion targets. (a) A closed-set classifier directly provides fixed class labels as inversion targets. (b) CLIP requires externally supplied candidate prompts, motivating explicit construction of the inversion targets.

2.1 Data-Free Quantization

Quantization reduces memory footprint and computational cost by representing model weights and activations at low precision (Wu et al., 2016). We focus on B-bit asymmetric uniform quantization, which maps a full-precision matrix $\mathbf{M}$ to an integer matrix $\mathbf{M}_q$ as formulated in Equation 1.

$$\mathbf{M}_q = \left\lfloor \frac{\mathbf{M}}{s} + z + \frac{1}{2} \right\rfloor, \tag{1}$$

where $s = (r_{\max} - r_{\min})/(2^B - 1)$ and $z = -\lfloor r_{\min}/s \rceil$ indicate the scale factor and zero-point, respectively. $(r_{\min}, r_{\max})$ denote the minimum and maximum values of $\mathbf{M}$, respectively. Despite its efficiency, quantization incurs numerical errors that perturb model outputs. Therefore, accurate quantization typically requires representative real data for calibration to mitigate these errors.

DFQ aims to quantize pre-trained models without access to their original training data. Existing DFQ methods are categorized according to their data-generation strategies (Kim et al., 2025a): synthesis-free (Nagel et al., 2019; Guo et al., 2022), generator-based (Xu et al., 2020; Li et al., 2024), and noise-optimization-based (Li et al., 2023a; Ramachandran et al., 2024) methods. We focus on noise-optimization-based methods, which commonly utilize model inversion to optimize synthetic images toward predefined model outputs, enabling direct control over their semantic content.

For closed-set classifiers, model inversion combines a target loss $\mathcal{L}_{\text{CE}}$ with an architecture-specific loss $\mathcal{L}_{\text{arch}}$ and image regularization $\mathcal{R}$, as formulated in Equation 2.

$$\mathcal{L}_{\text{inv}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{arch}}\mathcal{L}_{\text{arch}} + \mathcal{R}, \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ aligns predictions with fixed labels, $\mathcal{L}_{\text{arch}}$ imposes architecture-specific priors, and $\mathcal{R}$ promotes spatial smoothness and discourages excessive pixel magnitudes. The weight λ_{arch} controls the architecture-specific prior; Appendix C details both $\mathcal{L}_{\text{arch}}$ and $\mathcal{R}$. However, $\mathcal{L}_{\text{CE}}$ assumes a fixed label space and is therefore not directly applicable to CLIP models.

2.2 Contrastive Language-Image Pre-training Models

A contrastive language-image pre-training (CLIP) model aligns image representations from an image encoder $\mathcal{I}_\theta$ and text representations from a text encoder $\mathcal{T}_\phi$ in a shared embedding space through contrastive learning (Radford et al., 2021). Given N_b image-text pairs $\{(\mathbf{x}_i, t_i)\}_{i=1}^{N_b}$, a CLIP model minimizes the symmetric contrastive loss $\mathcal{L}_{\text{CLIP}}$ during pre-training as formulated in Equation 3.

$$\begin{aligned} \mathcal{L}_{\text{CLIP}} &= \frac{1}{2}\Big(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}\Big), \quad \mathcal{L}_{I\to T} = -\frac{1}{N_b}\sum_{i=1}^{N_b} \log \frac{\exp(\text{sim}(\mathbf{x}_i, t_i)/\tau)}{\sum_{j=1}^{N_b}\exp(\text{sim}(\mathbf{x}_i, t_j)/\tau)}, \\ \mathcal{L}_{T\to I} &= -\frac{1}{N_b}\sum_{i=1}^{N_b} \log \frac{\exp(\text{sim}(\mathbf{x}_i, t_i)/\tau)}{\sum_{j=1}^{N_b}\exp(\text{sim}(\mathbf{x}_j, t_i)/\tau)}, \quad \text{sim}(\mathbf{x}, t) = \frac{\mathcal{I}_\theta(\mathbf{x})^\top \mathcal{T}_\phi(t)}{\|\mathcal{I}_\theta(\mathbf{x})\|\|\mathcal{T}_\phi(t)\|}, \end{aligned} \tag{3}$$

where $\mathcal{L}_{I\to T}$ and $\mathcal{L}_{T\to I}$ denote the image-to-text and text-to-image contrastive losses, respectively. The temperature τ controls the sharpness of the softmax distribution over pairwise similarities.

In the inference phase, given candidate prompts $\{t_k\}_{k=1}^K$, a CLIP model selects the index k^* of the most similar prompt: $k^* = \arg\max_k \text{sim}(\mathbf{x}, t_k)$. The class associated with t_{k^*} becomes the predicted

label. The candidate prompt set therefore determines the output space for each task without a fixed classification head. As illustrated in Figure 2, a closed-set classifier directly provides fixed class labels as inversion targets, whereas CLIP requires candidate prompts to be supplied before optimizing synthetic images. Among prior DFQ methods, D4C (Zhang et al., 2025) specifically targets CLIP and adapts model inversion to CLIP models by replacing $\mathcal{L}_{\mathrm{CE}}$ used in closed-set classifiers with $\mathcal{L}_{\mathrm{CLIP}}$. However, D4C derives its inversion targets from a manually curated prompt set without a systematic mechanism for expanding their coverage across the open-vocabulary semantic space. In contrast, our proposed VISTA systematically constructs a broad inversion-target pool from WordNet (Miller, 1995) and samples text prompts from this pool.

2.3 PROBLEM DEFINITION

Given a pre-trained CLIP model and a target bit-width, DFQ for CLIP models aims to obtain a quantized CLIP model without access to real training data while preserving the zero-shot recognition performance of the full-precision model. Preserving image-text alignment is crucial to this goal because CLIP models compute task-specific predictions from image-text similarities rather than logits from a fixed classification head. We formalize this problem in Problem 1.

Problem 1 (DFQ for CLIP)**.** *Given a pre-trained CLIP model $\mathcal{M} = (\mathcal{I}_\theta, \mathcal{T}_\phi)$ and a target bit-width B, obtain a B-bit quantized CLIP model $\mathcal{M}^q = (\mathcal{I}^q_{\theta'}, \mathcal{T}^q_{\phi'})$ that minimizes the degradation in zero-shot recognition performance across diverse downstream tasks without access to real training data.*

3 PROPOSED METHOD

We propose VISTA (in**V**ers**I**on framework with **S**emantic **T**argets and **A**ppearance coverage), an accurate CLIP inversion framework for data-free quantization (DFQ). Existing DFQ methods for CLIP models suffer from the following three challenges.

- **C1. Limited semantic coverage.** CLIP models have no fixed classification head and therefore provide no predefined inversion targets. D4C (Zhang et al., 2025) constructs these targets from a manually curated prompt set with limited semantic coverage. How can we systematically construct inversion targets with broad semantic coverage?
- **C2. Ignoring semantic relationships.** Standard one-hot prompt supervision treats all non-target prompts as negatives. However, some non-target prompts are semantically related to the target prompt. How can we synthesize images that reflect semantic relationships among prompts?
- **C3. Limited visual diversity.** Semantic alignment alone does not explicitly encourage synthetic images to cover diverse visual characteristics. How can we expand visual diversity across synthetic images while preserving alignment with their target prompts?

We address these challenges with the following main ideas:

- **I1. WordNet-guided prompt sampling (Section 3.1).** VISTA systematically constructs a broad pool of inversion targets by sampling visual concepts from WordNet.
- **I2. Identity-preserving semantic relaxation (Section 3.2).** VISTA captures semantic relationships among prompts through identity-preserving soft targets derived from WordNet similarities.
- **I3. Visual diversity with semantic preservation (Section 3.3).** VISTA enforces representation consistency across augmented views and expands visual diversity by reducing correlations among prompt-orthogonal visual components.

Figure 3 illustrates how the three ideas are integrated into the inversion pipeline of VISTA. VISTA generates synthetic images by minimizing the overall inversion objective in Equation 4 while keeping the CLIP model frozen.

$$\mathcal{L}_{\text{VISTA}} = \mathcal{L}_{\text{relax}} + \lambda_{\text{view}}\mathcal{L}_{\text{view}} + \lambda_{\text{app}}\mathcal{L}_{\text{app}} + \lambda_{\text{arch}}\mathcal{L}_{\text{arch}} + \mathcal{R}, \tag{4}$$

where $\mathcal{L}_{\text{relax}}$ denotes the semantic relaxation loss defined in Section 3.2. $\mathcal{L}_{\text{view}}$ and $\mathcal{L}_{\text{app}}$ denote the view consistency and appearance diversity losses defined in Section 3.3, respectively. $\mathcal{L}_{\text{arch}}$ imposes architecture-specific priors from the pre-trained image encoder, while $\mathcal{R}$ encourages spatial smoothness and constrains pixel magnitudes. The hyperparameters λ_{view}, λ_{app}, and λ_{arch} control the contributions of the corresponding terms. Appendix B presents the algorithm of VISTA.

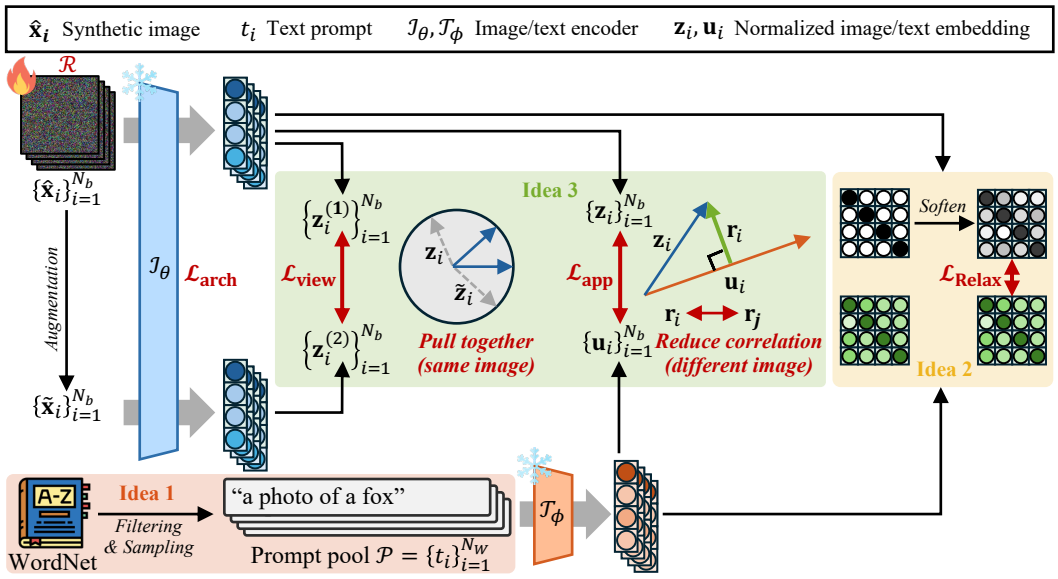


Figure 3: Overall architecture of VISTA. Our main ideas are 1) WordNet-guided prompt sampling, 2) identity-preserving semantic relaxation, and 3) visual diversity with semantic preservation.

3.1 WORDNET-GUIDED PROMPT SAMPLING

CLIP inversion requires text prompts that cover diverse visual concepts because CLIP has no fixed label space. D4C (Zhang et al., 2025) manually selects these prompts, which may limit their semantic coverage. Figure 1(a) compares the CLIP text embeddings derived from the reference, D4C, and VISTA prompt sets.We derive the reference prompts from downstream dataset labels and project all embeddings onto a two-dimensional space using t-SNE. We evaluate the coverage of each prompt set by computing IRP recall (Kynkäänniemi et al., 2019) in the original CLIP text-embedding space, while the two-dimensional projection serves only for visualization. For each inversion-prompt embedding, IRP recall defines a local neighborhood whose radius equals the distance to its k-th nearest neighbor within the same prompt set and computes recall as the proportion of reference prompt embeddings that fall within at least one of these neighborhoods. Each reference prompt is accordingly marked as covered only by D4C, only by VISTA, by both, or by neither method. D4C covers just 20.0% of the reference prompts, indicating limited semantic coverage. We therefore aim to construct inversion targets with broad semantic coverage.

VISTA systematically constructs inversion targets with broad semantic coverage by sampling visual concepts from WordNet, an image-free lexical database. WordNet (Miller, 1995) organizes English words into synonym sets (synsets) and connects them through semantic relations such as hypernymy and hyponymy. However, not every WordNet synset is suitable for inversion because some lexical categories, such as cognition and feeling, primarily contain abstract concepts. We therefore restrict the candidate set to four categories: animal, artifact, food, and plant. Appendix F.2 empirically evaluates the effect of this category filtering against all WordNet noun categories. VISTA randomly samples concepts from the filtered categories and converts each sampled concept into a text prompt using the template '`a photo of a {concept}`' to construct a prompt pool $\mathcal{P} = \{t_i\}_{i=1}^{N_W}$, where N_W denotes the pool size. This construction uses neither real images nor downstream task labels. We then form $\mathbb{D}_{\text{syn}} = \{(\hat{\mathbf{x}}_i, t_i)\}_{i=1}^{N_W}$ for $\mathcal{L}_{\text{CLIP}}$ by pairing each prompt t_i with a synthetic image $\hat{\mathbf{x}}_i$ initialized from Gaussian noise.

As shown in Figure 1(a), the WordNet-guided prompt set increases the IRP recall from 20.0% to 49.1%, a 29.1%p increase over D4C. While IRP recall measures local coverage of the reference prompts, maximum mean discrepancy (MMD) measures the global discrepancy between two distributions, with a lower value indicating closer distributional alignment. VISTA reduces the MMD to the reference prompt distribution from 0.036 to 0.009, a 75% reduction relative to D4C. Together, these

results show that the WordNet-guided prompt embeddings cover a broader portion of the reference embeddings and more closely match the reference prompt distribution.

3.2 IDENTITY-PRESERVING SEMANTIC RELAXATION

One-hot supervision in CLIP inversion ignores semantic relationships among sampled prompts. It assigns a probability of one to the corresponding prompt and zero to all other prompts, thereby treating all non-target prompts as equally negative. However, prompts in a mini-batch sampled from WordNet may represent semantically related concepts. One-hot supervision therefore penalizes alignment between a synthetic image and related non-target prompts. Consequently, inversion fails to preserve semantic relationships encoded by the CLIP model. Overall, we aim to incorporate semantic relationships among prompts into inversion supervision.

VISTA incorporates semantic relationships among prompts through an identity-preserving soft target matrix $\mathbf{Q} = [q_{ij}] \in [0,1]^{N_b \times N_b}$, where N_b denotes the mini-batch size. For each prompt t_i in a mini-batch, let s_i denote its corresponding WordNet synset. We define the semantic affinity between prompts t_i and t_j as $a_{ij} = \mathrm{sim}_{\mathrm{WN}}(s_i, s_j)$, where $\mathrm{sim}_{\mathrm{WN}}$ measures the similarity between two synsets along the hypernym-hyponym hierarchy of WordNet. Let $\mathbf{A} = [a_{ij}]$ denote the resulting affinity matrix. A naive approach constructs a soft target distribution by normalizing each row of $\mathbf{A}$. However, as the total affinity with non-target prompts increases, this normalization assigns less probability to the corresponding prompt and thereby weakens its target identity. VISTA prevents this dilution by placing probability mass δ on the corresponding prompt and distributing $1-\delta$ among non-target prompts according to their affinities, as defined in Equation 5.

$$q_{ij} = \begin{cases} \delta, & i = j, \\ (1-\delta)\frac{a_{ij}}{\sum_{k \neq i} a_{ik}}, & i \neq j, \end{cases} \tag{5}$$

where $\delta \in (0,1]$ controls the probability reserved for the corresponding prompt. Each row of $\mathbf{Q}$ forms a valid probability distribution, satisfying $\sum_{j=1}^{N_b} q_{ij} = 1$. VISTA replaces the one-hot target in Equation 3 with $\mathbf{Q}$ and defines the semantic relaxation loss in Equation 6.

$$\mathcal{L}_{\mathrm{relax}} = \frac{1}{2}\left(\mathcal{L}_{I\to T}^{\mathrm{relax}} + \mathcal{L}_{T\to I}^{\mathrm{relax}}\right) \quad \mathcal{L}_{I\to T}^{\mathrm{relax}} = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{j=1}^{N_b} q_{ij} \log \frac{\exp\left(\mathrm{sim}(\hat{\mathbf{x}}_i, t_j)/\tau\right)}{\sum_{k=1}^{N_b}\exp\left(\mathrm{sim}(\hat{\mathbf{x}}_i, t_k)/\tau\right)}, \tag{6}$$

where $\mathcal{L}_{T\to I}^{\mathrm{relax}}$ is defined analogously by exchanging the roles of images and prompts. When $\delta = 1$, $\mathbf{Q}$ becomes the identity matrix, and $\mathcal{L}_{\mathrm{relax}}$ reduces to the original CLIP loss in Equation 3. For $\delta < 1$, the diagonal entries reserve probability for the corresponding image-prompt pair, while the off-diagonal entries reflect semantic affinities among prompts.

3.3 VISUAL DIVERSITY WITH SEMANTIC PRESERVATION

Accurate quantization requires synthetic calibration images that cover diverse visual characteristics (Zhang et al., 2021; Zhong et al., 2022; Shang et al., 2023; Ramachandran et al., 2024). Highly similar synthetic images yield a narrow range of representations and activation patterns, limiting the calibration of quantization parameters. Figure 1(b) compares the CLIP image embeddings of real images with those of synthetic images generated by D4C and VISTA. We project the image embeddings onto a two-dimensional space using t-SNE for visualization. The synthetic image embeddings produced by D4C remain separated from the real image embeddings, indicating limited alignment with real images. This discrepancy suggests that the synthetic images capture only a limited range of the visual characteristics needed for quantization. A straightforward approach to expanding visual coverage is to reduce pairwise similarities among the entire image representations. However, this approach can also penalize components aligned with the target prompts, thereby weakening semantic alignment. Overall, we aim to expand the visual coverage of synthetic images while preserving their alignment with the target prompts.

VISTA addresses this trade-off with two objectives: view consistency and appearance diversity. The view consistency loss $\mathcal{L}_{\mathrm{view}}$ stabilizes the representation of each synthetic image across augmentations. The appearance diversity loss $\mathcal{L}_{\mathrm{app}}$ expands visual coverage by reducing correlations among prompt-orthogonal components without directly decorrelating components aligned with the target prompts.

Table 1: Top-1 accuracy [%] of quantized CLIP ViT-B/16 across different bit-widths and datasets. W/A denotes the weight and activation bit-widths, respectively. Avg reports the average accuracy across the datasets. Note that VISTA achieves the highest performance.

Method	W/A	Cal*	Cars*	DTD	FGVC*	Flow*	Food*	IN*	Pet*	SAT*	SUN*	UCF*	Avg
Full Precision	32/32	93.31	65.64	44.15	24.72	70.77	85.88	68.79	89.10	48.27	62.57	67.46	65.51
Gaussian	8/8	80.41	20.46	33.16	5.73	32.24	41.91	39.60	60.48	19.77	43.34	30.95	37.10
PSE		87.26	40.50	36.88	9.06	51.40	72.86	54.86	73.81	31.05	54.33	53.93	51.45
D4C (Zhang et al., 2025)		86.82	40.80	38.83	8.25	57.65	75.17	55.08	78.52	33.65	55.76	54.56	53.19
VISTA (proposed)		**92.41**	**58.94**	**43.38**	**20.18**	**68.87**	**82.90**	**66.07**	**88.17**	**34.02**	**61.43**	**63.91**	**61.84**
Gaussian	6/6	6.53	0.46	8.04	1.14	0.85	1.40	0.66	4.91	9.32	1.46	1.35	3.28
PSE		49.29	2.04	15.13	1.86	7.96	9.29	12.37	29.68	9.95	11.97	12.13	14.70
D4C (Zhang et al., 2025)		75.66	18.51	28.78	4.11	23.87	35.87	32.33	53.56	**27.84**	34.57	30.48	33.23
VISTA (proposed)		**80.11**	**31.48**	**31.82**	**10.13**	**38.56**	**43.78**	**45.47**	**78.44**	20.48	**45.12**	**44.09**	**42.68**
Gaussian	4/8	52.66	4.12	22.58	1.86	17.21	12.29	15.97	30.53	16.17	18.04	10.60	18.37
PSE		87.38	31.82	36.88	7.20	43.89	64.09	48.00	67.95	28.48	52.55	47.69	46.90
D4C (Zhang et al., 2025)		85.15	32.00	33.92	5.43	43.32	65.21	49.49	67.62	33.48	51.72	49.27	46.96
VISTA (proposed)		**89.98**	**45.44**	**40.23**	**15.63**	**60.02**	**72.69**	**59.59**	**81.94**	**38.11**	**59.15**	**57.48**	**56.39**
Gaussian	4/4	0.77	**0.53**	1.65	**1.11**	**1.26**	0.99	0.09	3.11	8.80	0.29	1.08	1.79
PSE		2.84	0.36	1.77	0.96	0.61	0.99	0.10	2.75	9.75	0.25	1.00	1.94
D4C (Zhang et al., 2025)		1.66	0.45	**2.07**	0.99	0.85	**1.57**	**0.30**	3.03	**11.19**	**0.30**	**1.45**	2.17
VISTA (proposed)		**6.21**	**0.53**	1.91	1.08	0.89	1.05	0.13	**3.82**	9.71	**0.30**	1.06	**2.43**

* Cal: Caltech101, Cars: StanfordCars, FGVC: FGVCAircraft, Flow: Flowers102, Food: Food101, IN: ImageNet, Pet: Oxford-IIIT Pet, SAT: EuroSAT, SUN: SUN397, UCF: UCF101.

First, VISTA enforces representation consistency between two views of each synthetic image. For each $\hat{\mathbf{x}}_i$, we apply two view transformations $g^{(1)}, g^{(2)} \in \mathcal{G}$ and generate $\tilde{\mathbf{x}}_i^{(v)} = g^{(v)}(\hat{\mathbf{x}}_i)$ for $v \in \{1, 2\}$. VISTA then aligns their normalized image representations using the view consistency loss $\mathcal{L}_{\text{view}}$ defined in Equation 7. Minimizing $\mathcal{L}_{\text{view}}$ increases the cosine agreement between the two views and stabilizes the synthetic image representations under augmentations.

$$\mathcal{L}_{\text{view}} = \frac{1}{N_b} \sum_{i=1}^{N_b} \left(1 - \mathbf{z}_i^{(1)\top} \mathbf{z}_i^{(2)}\right), \qquad \mathbf{z}_i^{(v)} = \frac{\mathcal{I}_\theta(\tilde{\mathbf{x}}_i^{(v)})}{\left\|\mathcal{I}_\theta(\tilde{\mathbf{x}}_i^{(v)})\right\|}, \quad v \in \{1, 2\}. \tag{7}$$

Second, VISTA expands visual coverage by reducing redundancy only among prompt-orthogonal components of the synthetic image representations. Let $\mathbf{u}_i = \mathcal{T}_\phi(t_i)/\|\mathcal{T}_\phi(t_i)\|$ denote the normalized representation of the target prompt, and let $\mathbf{z}_i = \mathbf{z}_i^{(2)}$ denote the normalized image representation used for appearance regularization. VISTA removes the component of $\mathbf{z}_i$ aligned with $\mathbf{u}_i$ and obtains the residual $\mathbf{r}_i = \mathbf{z}_i - (\mathbf{z}_i^\top \mathbf{u}_i)\mathbf{u}_i$. VISTA minimizes the appearance diversity loss defined in Equation 8 to reduce pairwise correlations among the normalized residuals.

$$\mathcal{L}_{\text{app}} = \frac{1}{N_b(N_b - 1)} \sum_{i=1}^{N_b} \sum_{\substack{j=1 \\ j \neq i}}^{N_b} \left(\bar{\mathbf{r}}_i^\top \bar{\mathbf{r}}_j\right)^2, \qquad \bar{\mathbf{r}}_i = \frac{\mathbf{r}_i}{\|\mathbf{r}_i\|}, \quad \mathbf{r}_i = \mathbf{z}_i - \left(\mathbf{z}_i^\top \mathbf{u}_i\right) \mathbf{u}_i, \tag{8}$$

Minimizing $\mathcal{L}_{\text{app}}$ encourages the prompt-orthogonal components to represent distinct visual variations while avoiding direct decorrelation of the prompt-aligned components. The squared term penalizes both positive and negative correlations between residuals.

We quantify the discrepancy between the synthetic and real image distributions in the CLIP image embedding space using MMD, where a lower value indicates closer distributions. Figure 1(b) shows that the synthetic image distribution produced by VISTA is more closely aligned with the real image distribution than that produced by D4C. Consistent with this observation, VISTA reduces the MMD from 0.172 to 0.082, a 52.3% reduction relative to D4C.

4 EXPERIMENTS

We perform experiments to answer the following questions about VISTA.

Q1. Accuracy in Data-Free Quantization (DFQ) (Section 4.2). How does VISTA compare with existing DFQ methods in quantization accuracy across bit-widths and CLIP backbones?

Q2. Ablation study (Section 4.3). How does each component of VISTA improve quantization accuracy?

Table 2: Average top-1 accuracy [%] over 11 datasets for quantized CLIP models across different backbones and bit-widths. Avg reports the mean across the six backbones. Note that VISTA achieves the highest average accuracy at all evaluated bit-widths.

Method	W/A	B/32	L/14	RN50	RN101	RN50x4	RN50x16	Avg
Full Precision	32/32	62.17	72.77	58.88	59.66	63.01	66.34	63.81
Gaussian	8/8	27.09	49.62	45.73	47.97	52.11	56.25	46.46
PSE & BNS		46.12	60.29	45.75	47.78	53.41	56.54	51.65
D4C (Zhang et al., 2025)		47.80	60.17	47.02	48.65	52.49	55.79	51.99
VISTA (proposed)		**56.26**	**69.88**	**56.63**	**58.27**	**61.77**	**64.86**	**61.28**
Gaussian	6/6	19.51	2.63	35.03	40.79	46.15	48.41	32.09
PSE & BNS		40.39	9.67	40.64	46.50	49.42	51.88	39.75
D4C (Zhang et al., 2025)		43.12	8.84	42.26	44.29	48.15	51.39	39.67
VISTA (proposed)		**52.68**	**45.48**	**51.87**	**55.23**	**58.86**	**62.83**	**54.49**
Gaussian	4/8	15.50	32.26	21.55	34.61	39.88	47.99	31.97
PSE & BNS		39.27	59.05	41.10	44.34	48.47	52.10	47.39
D4C (Zhang et al., 2025)		41.77	59.21	40.29	42.10	46.40	53.08	47.14
VISTA (proposed)		**49.22**	**66.88**	**48.28**	**52.90**	**56.44**	**61.96**	**55.95**
Gaussian	4/4	2.09	2.08	2.65	5.27	3.20	2.43	2.95
PSE & BNS		13.04	1.80	14.77	21.78	24.42	17.25	15.51
D4C (Zhang et al., 2025)		15.43	**2.74**	13.10	19.14	20.32	15.02	14.29
VISTA (proposed)		**19.39**	2.67	**17.35**	**26.02**	**29.31**	**17.57**	**18.72**

Q3. Hyperparameter analysis (Section 4.4). How robust is the quantization accuracy of VISTA to changes in its key hyperparameters?

4.1 EXPERIMENTAL SETUP

We briefly introduce the experimental setup. Further experimental details are provided in Appendix E.

Setup. We evaluate the zero-shot top-1 accuracy of quantized CLIP models on 11 image classification datasets: Caltech101 (Fei-Fei et al., 2007), StanfordCars (Krause et al., 2013), DTD (Cimpoi et al., 2014), FGVCAircraft (Maji et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), ImageNet (Deng et al., 2009), Oxford-IIIT Pet (Parkhi et al., 2012), EuroSAT (Helber et al., 2019), SUN397 (Xiao et al., 2010), and UCF101 (Soomro et al., 2012). We further evaluate seven CLIP backbones (Radford et al., 2021) with different architectures and model sizes: ViT-B/16, ViT-B/32, ViT-L/14, RN50, RN101, RN50x4, and RN50x16.

Implementation details. We base our experimental setup on that of D4C (Zhang et al., 2025) for a fair comparison. As in D4C, we construct a text prompt pool from 180 concepts. We then sample prompts from this pool to synthesize 128 images with a batch size of 16, a learning rate of 0.01, and 3,000 iterations. For quantization, we calibrate the image and text encoders with 128 generated images and 512 text prompts, respectively, and perform layer-wise reconstruction for 20,000 iterations. All experiments are conducted on a workstation with an Intel Xeon Gold 6534 and NVIDIA L40S.

4.2 ACCURACY IN DATA-FREE QUANTIZATION

We compare VISTA against existing DFQ methods to assess the effectiveness of its synthetic calibration data. Table 1 reports the top-1 accuracy of quantized ViT-B/16. First, VISTA achieves the highest average accuracy at all evaluated bit-widths, outperforming the strongest competitor by up to 9.45%p. Second, activation precision has a greater influence on CLIP quantization than weight precision in these settings, as W4/A8 consistently outperforms W6/A6 across all methods despite its lower weight precision. Third, W4/A4 remains the most challenging setting, with all methods showing substantial accuracy degradation.

Table 2 extends the comparison to six additional CLIP backbones. VISTA achieves the highest average accuracy across backbones at all evaluated bit-widths. This result supports its applicability across diverse architectures and model scales. Detailed per-dataset results are provided in Appendix F.

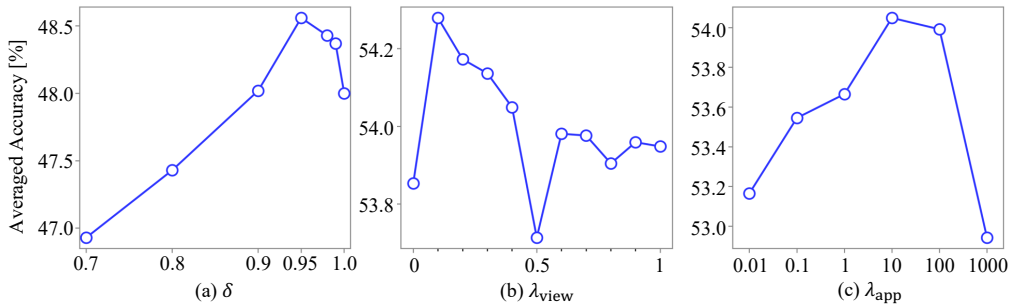


Figure 4: Hyperparameter analysis of (a) identity confidence δ, (b) view consistency weight λ_{view}, and (c) appearance diversity weight λ_{app}. See Section 4.4 for details.

4.3 Ablation Study

We perform an ablation study to show that each main idea of VISTA, such as WordNet-guided prompt sampling (I1) in Section 3.1, identity-preserving semantic relaxation (I2) in Section 3.2, and visual diversity with semantic preservation (I3) in Section 3.3, improves the classification accuracy of the quantized model. We evaluate I2 and I3 only in combination with I1 because they require a prompt-based inversion loss that cannot be computed without the candidate prompts supplied by I1. We summarize the W4/A8 quantization accuracy of ViT-B/32 averaged over 11 datasets in Table 3. Our analysis shows that all proposed ideas improve accuracy, with I3 having the strongest impact of 1.21%p.

Table 3: Ablation study on the main ideas of VISTA. All ideas contribute to the improved performance

I1	I2	I3	Avg. [%]	Gain [%p]
			46.89	0.00
✓			47.76	+0.87
✓	✓		47.93	+1.04
✓		✓	48.97	+2.08
✓	✓	✓	49.39	+2.50

4.4 Hyperparameter Analysis

We analyze the sensitivity of VISTA to three newly introduced hyperparameters: δ, λ_{view}, and λ_{app} in Figure 4. We report the W8/A8 quantization accuracy of ViT-B/32, averaged across the 11 datasets. We have three observations from the results. First, Figure 4(a) shows that the accuracy increases as δ increases from 0.7 to 0.95 but decreases beyond 0.95. Notably, values below 0.9 underperform one-hot supervision ($\delta = 1$), indicating that excessive semantic relaxation can be more harmful than applying no relaxation. This pattern highlights the importance of balancing semantic relaxation with target identity: a small δ assigns excessive probability mass to related prompts, whereas a value close to 1 approaches one-hot supervision and limits the benefit of semantic relaxation. Second, Figure 4(b) shows that setting λ_{view} to 0.1 yields the highest accuracy. Performance generally declines as λ_{view} moves away from 0.1, with a particularly pronounced drop at $\lambda_{\text{view}} = 0.5$. These results suggest that a small λ_{view} is effective, while the isolated drop at 0.5 indicates sensitivity at that setting rather than a monotonic degradation with increasing weight. Third, Figure 4(c) shows that accuracy increases as λ_{app} increases from 0.01 to 10 but decreases as λ_{app} increases beyond 10. This pattern suggests that appearance diversity regularization is effective across a relatively broad range of λ_{app}, whereas an excessively large weight can degrade quantization accuracy.

5 Conclusion

Data-free quantization (DFQ) for CLIP models is challenging because the absence of a fixed classification head leaves inversion targets undefined, while existing CLIP inversion is limited by manually curated prompts, one-hot supervision, and insufficient within-prompt visual diversity. In this paper, we propose VISTA, an accurate CLIP inversion framework for DFQ. VISTA samples visual prompts from WordNet, constructs identity-preserving soft targets, and promotes appearance diversity in prompt-orthogonal components while enforcing view consistency. Experiments on various CLIP backbones, bit-widths, and downstream datasets show that VISTA improves the average accuracy by up to 9.45%p over existing DFQ methods.

AI USE STATEMENT

We employ ChatGPT[1] (GPT-5.6) and Claude[2] (Opus 5.5) for language editing, including grammatical correction and improvements to clarity at the sentence level. We do not use them for research-related tasks, including theoretical derivation and result analysis. The authors take all the responsibility for the final content of this work.

REFERENCES

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, 2014.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *CVPR*, pp. 13166–13175, 2020.

Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Weiming Hu, Xiaohu Qie, and Jianping Wu. Vilem: Visual-language error modeling for image-text retrieval. In *CVPR*, 2023.

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *ICCV Workshops*, 2019.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proc. IEEE*, 108(4):485–532, 2020.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *ICLR*, 2020.

Li Fei-Fei, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 2007.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789–1819, 2021.

Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, and Minyi Guo. Squant: On-the-fly data-free quantization via diagonal hessian approximation. In *ICLR*, 2022.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural network. In *NIPS*, pp. 1135–1143, 2015.

Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5):2900–2919, 2024.

[1] https://chatgpt.com/

[2] https://claude.ai/

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 2019.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Minjun Kim, Jaehyeon Choi, Jongkeun Lee, Wonjin Cho, and U Kang. Zero-shot quantization: A comprehensive survey. In *IJCAI*, pp. 10508–10516, 2025a.

Minjun Kim, Jongjin Kim, and U Kang. Synq: Accurate zero-shot quantization by synthesis-aware fine-tuning. In *ICLR*, 2025b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019.

Huantong Li, Xiangmiao Wu, Fanbing Lv, Daihai Liao, Thomas H. Li, Yonggang Zhang, Bo Han, and Mingkui Tan. Hard sample matters a lot in zero-shot quantization. In *CVPR*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023b.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: pushing the limit of post-training quantization by block reconstruction. In *ICLR*, 2021.

Yuhang Li, Youngeun Kim, Donghyun Lee, Souvik Kundu, and Priyadarshini Panda. Genq: Quantization in low data regimes with generative synthetic data. In *ECCV*, 2024.

Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *ECCV*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

He Liu, Yikai Wang, Huaping Liu, Fuchun Sun, and Anbang Yao. Small scale data-free knowledge distillation. In *CVPR*, 2024.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *ICCV*, 2019.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *ICML*, 2020.

Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *ICML*, 2019.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Akshat Ramachandran, Souvik Kundu, and Tushar Krishna. Clamp-vit: Contrastive data-free learning for adaptive post-training quantization of vits. In *ECCV*, 2024.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022.

Babak Rokh, Ali Azarpeyvand, and Alireza Khanteymoori. A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Trans. Intell. Syst. Technol.*, 14(6):97:1–97:50, 2023.

Yuzhang Shang, Bingxin Xu, Gaowen Liu, Ramana Rao Kompella, and Yan Yan. Causal-dfq: Causality guided data-free network quantization. In *ICCV*, 2023.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Jialiang Tang, Mingjin Liu, Ning Jiang, Huan Cai, Wenxin Yu, and Jinjia Zhou. Data-free network pruning for model compression. In *ISCAS*, pp. 1–5, 2021.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: clip-driven referring image segmentation. In *CVPR*, 2022.

Zi Wang. Data-free knowledge distillation with soft targeted transfer set synthesis. In *AAAI*, 2021.

Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *CVPR*, pp. 4820–4828, 2016.

Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, 2023.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In *ECCV*, 2020.

Hongxu Yin, Pavlo Molchanov, José M. Álvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, 2020.

Shikang Yu, Jiachen Chen, Hu Han, and Shuqiang Jiang. Data-free knowledge distillation via feature exchange and activation region constraint. In *CVPR*, 2023.

Boqiang Zhang, Lei Ke, Ruihan Yang, Qi Gao, Tianyuan Qu, Rossell Chen, Dong Yu, and Leoweiliang. Penguin-vl: Exploring the efficiency limits of VLM with llm-based vision encoders. *arXiv preprint arXiv:2603.06569*, 2026.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *ECCV*, 2022.

Wenlun Zhang, Yunshan Zhong, Zihao Ding, Xinyu Li, and Kentaro Yoshioka. D4C: data-free quantization for contrastive language-image pre-training models. In *CVPR Findings*, 2025.

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 *CVPR*, 2021.

Xiangyu Zhang, Jianhua Zou, Xiang Ming, Kaiming He, and Jian Sun. Efficient and accurate approximations of nonlinear convolutional networks. In *CVPR*, pp. 1984–1992, 2015.

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 *CVPR*, 2022.

A NOTATION

We summarize the frequently used symbols in Table 4.

Table 4: Frequently used symbols.

Symbol	Description
$\mathcal{M} = (\mathcal{I}_\theta, \mathcal{T}_\phi)$	Pre-trained CLIP model with an image encoder $\mathcal{I}_\theta$ and a text encoder $\mathcal{T}_\phi$
$\mathcal{M}^q = (\mathcal{I}^q_{\theta'}, \mathcal{T}^q_{\phi'})$	Quantized CLIP model with an image encoder $\mathcal{I}^q_{\theta'}$ and a text encoder $\mathcal{T}^q_{\phi'}$
$\mathbb{D}_{\text{syn}} = \{(\hat{\mathbf{x}}_i, t_i)\}_{i=1}^N$	Synthetic image-text dataset
$\hat{\mathbf{x}}_i, t_i$	i-th synthetic image and its corresponding text prompt, respectively
N, N_b	Total number of synthetic samples and the batch size, respectively
B	Target quantization bit-width
$\mathcal{P} = \{t_i\}_{i=1}^{N_W}$	WordNet-derived text prompt pool
N_W	Number of prompts in the WordNet-derived prompt pool
s_i	WordNet synset associated with the i-th text prompt
a_{ij}	WordNet-based semantic affinity between the i-th and j-th text prompts
$\mathbf{A} = [a_{ij}]$	WordNet-based semantic affinity matrix
$\mathbf{Q} = [q_{ij}]$	Identity-preserving soft target matrix
δ	Probability reserved for the corresponding prompt in the soft target distribution
$\mathcal{G}$	Set of stochastic view transformations
$g^{(v)}$	Stochastic transformation used to generate the v-th augmented view
$\widetilde{\mathbf{x}}_i^{(v)}$	v-th augmented view of the i-th synthetic image
$\mathbf{z}_i^{(v)}$	Normalized image representation of $\widetilde{\mathbf{x}}_i^{(v)}$
$\mathbf{u}_i$	Normalized text representation of the i-th text prompt
$\mathbf{r}_i, \bar{\mathbf{r}}_i$	Prompt-orthogonal residual and its normalized representation, respectively
$\mathcal{L}_{\text{VISTA}}$	Overall inversion objective of VISTA
$\mathcal{L}_{\text{relax}}$	Identity-preserving semantic relaxation loss
$\mathcal{L}_{\text{view}}, \lambda_{\text{view}}$	View consistency loss and its weight, respectively
$\mathcal{L}_{\text{app}}, \lambda_{\text{app}}$	Prompt-orthogonal appearance diversity loss and its weight, respectively
$\mathcal{L}_{\text{arch}}, \lambda_{\text{arch}}$	Architecture-specific inversion loss and its weight, respectively
$\mathcal{R}$	Image regularization

B ALGORITHM

We describe the inversion procedure of VISTA in Algorithm 1. VISTA first constructs a WordNet-derived text prompt pool $\mathcal{P}$ and samples text prompts together with corresponding synthetic images initialized from Gaussian noise (Lines 1-5). For each mini-batch, VISTA constructs the identity-preserving soft target matrix $\mathbf{Q}$ from the WordNet-based semantic affinities matrix $\mathbf{A}$ and the identity confidence δ (Lines 6-7). During inversion, VISTA computes normalized text and image representations, then calculates semantic relaxation, view consistency, and appearance diversity to optimize the synthetic images (Lines 8-15). The semantic relaxation loss incorporates semantic relationships among sampled prompts into the inversion target, while the view consistency loss preserves stable representations across augmented views. The appearance diversity loss further encourages distinct prompt-orthogonal visual components across synthetic images without directly disturbing their prompt-aligned components. The architecture-specific inversion loss and image regularization are then incorporated into the overall objective $\mathcal{L}_{\text{VISTA}}$, which is minimized to update the synthetic images (Lines 16-18). After T iterations, the optimized image-text pairs are added to the synthetic dataset, and this procedure is repeated until the target number of samples is generated (Lines 19-21). As a result, the generated samples jointly capture semantic relationships among prompts and diverse visual characteristics while preserving alignment with the pre-trained CLIP model. This synthetic dataset can serve as calibration or training data for downstream data-free or zero-shot applications, including data-free quantization.

Algorithm 1 Overall process of VISTA.

Input: pre-trained CLIP model $\mathcal{M} = (\mathcal{I}_\theta, \mathcal{T}_\phi)$, number N of synthetic samples, prompt-pool size N_W, batch size N_b, number T of inversion iterations, identity confidence δ, and hyperparameters λ_{view}, λ_{app}, and λ_{arch}

Output: Synthetic image-text dataset $\mathbb{D}_{\text{syn}} = \{(\hat{\mathbf{x}}_i, t_i)\}_{i=1}^{N}$

Construct a WordNet-derived text prompt pool $\mathcal{P}$. ▷ *Idea 1: WordNet-guided prompt sampling*
Initialize $\mathbb{D}_{\text{syn}} \leftarrow \emptyset$.
while $|\mathbb{D}_{\text{syn}}| < N$ **do**
  Sample N_b text prompts from $\mathcal{P}$.
  Initialize the corresponding synthetic images $\{\hat{\mathbf{x}}_i\}_{i=1}^{N_b}$ with Gaussian noise.
  Compute the WordNet-based semantic affinity matrix $\mathbf{A} = [a_{ij}]$.
  Construct the identity-preserving soft target matrix $\mathbf{Q} = [q_{ij}]$ using $\mathbf{A}$ and δ.
    ▷ *Idea 2: Identity-preserving semantic relaxation*
  Normalize text representations $\mathbf{u}_i \leftarrow \mathcal{T}_\phi(t_i)/\|\mathcal{T}_\phi(t_i)\|$.
  for $k = 0$ to $T - 1$ **do**
    Generate two augmented views $\tilde{\mathbf{x}}_i^{(v)} = g^{(v)}(\hat{\mathbf{x}}_i)$, $v \in \{1, 2\}$.
    Normalize image representations $\mathbf{z}_i^{(v)} \leftarrow \mathcal{I}_\theta(\tilde{\mathbf{x}}_i^{(v)})/\|\mathcal{I}_\theta(\tilde{\mathbf{x}}_i^{(v)})\|$.
    Compute the semantic relaxation loss $\mathcal{L}_{\text{relax}}$ using $\mathbf{Q}$.
    Compute the view consistency loss $\mathcal{L}_{\text{view}}$ using $\mathbf{z}_i^{(1)}$ and $\mathbf{z}_i^{(2)}$.
      ▷ *Idea 3-1: View consistency*
    Compute prompt-orthogonal residuals $\mathbf{r}_i \leftarrow \mathbf{z}_i^{(2)} - \left((\mathbf{z}_i^{(2)})^\top \mathbf{u}_i\right) \mathbf{u}_i$.
    Normalize $\bar{\mathbf{r}}_i = \mathbf{r}_i/\|\mathbf{r}_i\|$ and compute appearance diversity loss $\mathcal{L}_{\text{app}}$.
      ▷ *Idea 3-2: Appearance diversity*
    Compute $\mathcal{L}_{\text{arch}}$ and $\mathcal{R}$.
    Compute overall loss $\mathcal{L}_{\text{VISTA}} \leftarrow \mathcal{L}_{\text{relax}} + \lambda_{\text{view}}\mathcal{L}_{\text{view}} + \lambda_{\text{app}}\mathcal{L}_{\text{app}} + \lambda_{\text{arch}}\mathcal{L}_{\text{arch}} + \mathcal{R}$.
    Update $\{\hat{\mathbf{x}}_i\}_{i=1}^{N_b}$ by minimizing $\mathcal{L}_{\text{VISTA}}$.
  end for
  Add the optimized image-text pairs to $\mathbb{D}_{\text{syn}}$.
end while
  return $\mathbb{D}_{\text{syn}}$

C ARCHITECTURE-SPECIFIC LOSSES AND IMAGE REGULARIZATION

We provide detailed formulations of the architecture-specific loss $\mathcal{L}_{\text{arch}}$ and image regularization $\mathcal{R}$ used in Equation 2. $\mathcal{L}_{\text{arch}}$ constrains internal representations according to the architecture of the image encoder, while $\mathcal{R}$ imposes spatial and magnitude priors directly on the synthetic images.

C.1 ARCHITECTURE-SPECIFIC LOSSES

Following existing DFQ methods, we implement $\mathcal{L}_{\text{arch}}$ as batch-normalization statistics (BNS) for CNN image encoders and patch similarity entropy (PSE) for ViT image encoders.

Batch-Normalization Statistics. Batch-normalization (BN) layers in a pre-trained CNN retain channel-wise running statistics accumulated during pre-training. BNS matches these statistics with those induced by a synthetic mini-batch, constraining inversion to produce internal activations consistent with the retained statistics and thereby providing a more representative signal for quantization calibration than unconstrained noise. For the l-th BN layer, let $\hat{\boldsymbol{\mu}}_l$ and $\hat{\boldsymbol{\sigma}}_l$ denote the channel-wise mean and standard deviation of its input activations over the mini-batch and spatial dimensions. Let $\boldsymbol{\mu}_l$ and $\boldsymbol{\sigma}_l$ denote the corresponding running statistics stored in the pre-trained model. The BNS loss $\mathcal{L}_{\text{BNS}}$ is formulated as in Equation 9, where N_{BN} is the number of BN layers.

$$\mathcal{L}_{\text{BNS}} = \frac{1}{N_{\text{BN}}} \sum_{l=1}^{N_{\text{BN}}} \left(\|\hat{\boldsymbol{\mu}}_l - \boldsymbol{\mu}_l\|_2^2 + \|\hat{\boldsymbol{\sigma}}_l - \boldsymbol{\sigma}_l\|_2^2 \right). \tag{9}$$

Patch Similarity Entropy. Unlike CNN image encoders, ViT image encoders do not retain BN running statistics that can serve as an inversion prior. PSE instead utilizes the relationships among patch representations produced by self-attention, encouraging diverse patch-wise responses and thereby providing informative internal activations for quantization calibration. For the l-th self-attention layer, let $\mathbf{A}_b^{(l,r)}$ denote the attention matrix of head r for the b-th synthetic image, and let $\mathbf{V}_b^{(l)}$ denote its value representations. After averaging the attention-weighted contexts over the H_l heads and removing the class token, the context representation of patch i is formulated as in Equation 10.

$$\mathbf{c}_{b,i}^{(l)} = \frac{1}{H_l} \sum_{r=1}^{H_l} \left(\mathbf{A}_b^{(l,r)} \mathbf{V}_b^{(l)} \right)_i . \tag{10}$$

We measure the relationship between patches i and j using their cosine similarity as in Equation 11.

$$s_{b,ij}^{(l)} = \frac{\mathbf{c}_{b,i}^{(l)\top} \mathbf{c}_{b,j}^{(l)}}{\left\| \mathbf{c}_{b,i}^{(l)} \right\|_2 \left\| \mathbf{c}_{b,j}^{(l)} \right\|_2}. \tag{11}$$

For each image, we estimate the density of the N_l^2 pairwise similarities using a Gaussian kernel K_h with bandwidth h as in Equation 12.

$$\hat{f}_b^{(l)}(x) = \frac{1}{N_l^2} \sum_{i=1}^{N_l} \sum_{j=1}^{N_l} K_h \left(x - s_{b,ij}^{(l)} \right), \tag{12}$$

where N_l is the number of patch tokens at layer l. The PSE loss $\mathcal{L}_{\mathrm{PSE}}$ is defined as in Equation 13.

$$\begin{aligned} \mathcal{L}_{\mathrm{PSE}} &= -\sum_{l=1}^{N_{\mathrm{SA}}} \mathcal{H}_l \\ \mathcal{H}_l &= -\frac{1}{N_b} \sum_{b=1}^{N_b} \int \left(\hat{f}_b^{(l)}(x) + \epsilon_{\mathrm{KDE}} \right) \log \left(\hat{f}_b^{(l)}(x) + \epsilon_{\mathrm{KDE}} \right) dx, \end{aligned} \tag{13}$$

where N_{SA} is the number of self-attention layers. Minimizing $\mathcal{L}_{\mathrm{PSE}}$ maximizes the entropy of patch similarities, discouraging homogeneous patch relationships across the synthetic images.

C.2 Image Regularization

Image regularization $\mathcal{R}$ combines total variation (TV) $\mathcal{R}_{\mathrm{tv}}$ and L2 penalties $\mathcal{R}_{\ell_2}$ to promote spatial smoothness and discourage excessive pixel magnitudes, as formulated in Equation 14.

$$\mathcal{R} = \lambda_{\mathrm{tv}} \mathcal{R}_{\mathrm{tv}} + \lambda_{\ell_2} \mathcal{R}_{\ell_2}, \tag{14}$$

where λ_{tv} and λ_{ℓ_2} control the contributions of the corresponding penalties.

Total Variation Penalty. The total variation penalty $\mathcal{R}_{\mathrm{tv}}$ promotes spatial smoothness by penalizing squared differences between horizontally and vertically adjacent pixels. Given a mini-batch of synthetic images $\{\hat{\mathbf{x}}_i\}_{i=1}^{N_b}$, where each image has C channels, height H, and width W, we formulate $\mathcal{R}_{\mathrm{tv}}$ as in Equation 15.

$$\begin{aligned} \mathcal{R}_{\mathrm{tv}} = \frac{1}{N_b} \sum_{i=1}^{N_b} \Bigg[& \frac{1}{CH(W-1)} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W-1} \left(\hat{\mathbf{x}}_{i,c,h,w+1} - \hat{\mathbf{x}}_{i,c,h,w} \right)^2 \\ & + \frac{1}{C(H-1)W} \sum_{c=1}^{C} \sum_{h=1}^{H-1} \sum_{w=1}^{W} \left(\hat{\mathbf{x}}_{i,c,h+1,w} - \hat{\mathbf{x}}_{i,c,h,w} \right)^2 \Bigg], \end{aligned} \tag{15}$$

where the first and second terms penalize horizontal and vertical pixel differences, respectively.

L2 Penalty. The L2 penalty $\mathcal{R}_{\ell_2}$ discourages excessive pixel magnitudes, as defined in Equantion 16.

$$\mathcal{R}_{\ell_2} = \frac{1}{N_b CHW} \sum_{i=1}^{N_b} \left\| \hat{\mathbf{x}}_i \right\|_2^2 . \tag{16}$$

D TRAINING AND EVALUATION

D.1 SYNTHETIC IMAGE OPTIMIZATION

We provide details for the synthetic image optimization procedure summarized in Algorithm 1. Given prompts sampled from WordNet, VISTA optimizes synthetic images initialized from a Gaussian distribution while keeping both CLIP encoders frozen. The following paragraphs describe the prompt pool construction, image optimization, and view augmentation.

Prompt pool construction. WordNet (Miller, 1995) organizes concepts into synsets according to their parts of speech and lexical categories. Among its noun synsets, we select four lexical categories that primarily describe visually depictable entities: `noun.animal`, `noun.artifact`, `noun.food`, and `noun.plant`. The prompt pool can be extended to additional noun lexical categories without modifying the subsequent inversion procedure. We sample 180 concepts from these categories to match the prompt-pool size used by D4C (Zhang et al., 2025). We then sample 128 concepts from this pool and convert each into a text prompt using the template "`a photo of a {concept}`" to obtain the inversion targets.

Image initialization and optimization. We describe the synthetic image optimization procedure, including image initialization, inversion target construction, and iterative image optimization. Each sampled prompt is paired with a synthetic image initialized from a standard Gaussian distribution at the native input resolution of the corresponding CLIP image encoder. For each mini-batch, we construct an identity-preserving soft target matrix based on WordNet affinities among the sampled prompts, as defined in Equation 5. During optimization, only the synthetic images are updated, while both the CLIP image and text encoders of the evaluation model remain frozen. We use the Adam (Kingma & Ba, 2015) optimizer with $(\beta_1, \beta_2) = (0.5, 0.99)$ and an initial learning rate of 0.01. We optimize 128 synthetic images for 3,000 optimization iterations using a batch size of 16. After each optimization step, the synthetic images are clipped to the valid RGB range according to the CLIP normalization statistics.

View augmentation. We describe the view augmentation procedure used to compute the view consistency loss $\mathcal{L}_{\text{view}}$ in Section 3.3. Given a synthetic image $\hat{\mathbf{x}}_i$, we apply two view transformations $g^{(1)}$ and $g^{(2)}$ to generate $\tilde{\mathbf{x}}_i^{(1)} = g^{(1)}(\hat{\mathbf{x}}_i)$ and $\tilde{\mathbf{x}}_i^{(2)} = g^{(2)}(\hat{\mathbf{x}}_i)$, respectively. We utilize the augmentation pipeline from D4C (Zhang et al., 2025); we adopt it to construct the transformed views used by $\mathcal{L}_{\text{view}}$. We generate each augmented view by combining multiple transformations, including horizontal flipping, affine transformation, color jittering, Gaussian blurring, and random erasing. We briefly describe each transformation and its supported parameter range below.

- **Horizontal Flipping.** Each image is horizontally flipped with probability $p_{\text{flip}} \in [0, 1]$, where $p_{\text{flip}} = 0$ disables flipping and $p_{\text{flip}} = 1$ flips every image.
- **Affine Transformation.** Each image undergoes a random affine transformation consisting of rotation, scaling, and translation. The rotation angle is uniformly sampled from $[-d_{\max}, d_{\max}]$, where $d_{\max} \geq 0$ denotes the maximum rotation angle in degrees. The scale coefficient is sampled from $[0.95, 1.05]$, while horizontal and vertical translation offsets are independently sampled from $[-0.05, 0.05]$ in normalized coordinates.
- **Color Jittering.** When enabled, brightness, contrast, and saturation are independently perturbed using multiplicative factors sampled from $[1 - s, 1 + s]$, where s denotes the jitter strength. We restrict s to $[0, 1]$ to keep these factors non-negative, with $s = 0$ disabling color jittering.
- **Gaussian Blurring.** Each image is blurred with probability $p_{\text{blur}} \in [0, 1]$ using a fixed 3×3 Gaussian kernel. Setting $p_{\text{blur}} = 0$ disables blurring, whereas $p_{\text{blur}} = 1$ applies it to every image.
- **Random Erasing.** Each image undergoes random erasing with probability $p_{\text{erase}} \in [0, 1]$. A square region with a side length of $\lfloor 0.2H \rfloor$, where H is the image height, is centered at a randomly sampled spatial location and set to zero in normalized image space; regions extending beyond the image boundary are clipped accordingly.

In our experiments, we set $g^{(1)}$ to the identity transformation and configure $g^{(2)}$ using the augmentations described above. For $g^{(2)}$, we use $p_{\text{flip}} = 0.5$, $d_{\max} = 15$, $s = 0.4$, $p_{\text{blur}} = 0.3$, and $p_{\text{erase}} = 0.3$.

D.2 Quantization and Reconstruction

We describe how the generated synthetic data are used to calibrate and reconstruct quantized CLIP models. To isolate the effect of synthetic data generation, we select the same quantization configuration, calibration budget, and reconstruction procedure for all compared methods.

Quantization configuration. We evaluate four weight/activation bit-width configurations: W8/A8, W6/A6, W4/A8, and W4/A4. Weights and activations in all configurations are quantized according to the asymmetric uniform scheme described in Section 2.1. We follow the quantization pipeline of D4C (Zhang et al., 2025) and quantize weights per channel and activations per tensor by default. We use AdaRound (Nagel et al., 2020) to quantize weights in internal convolutional and linear layers. The input layers and embedding parameters remain in full precision, as do the final projection matrices of the vision and text Transformers. We adopt LSQ (Esser et al., 2020) to quantize activations at three locations: the input to each quantized layer, the output of each attention projection (query, key, or value), and the output of each softmax operation. We instead apply per-channel activation quantization at the inputs to these attention projections and to the first projection of each Transformer MLP because the corresponding quantization parameters can be folded into the preceding LayerNorm layers (Ba et al., 2016). Inputs to the second projection of each text-encoder MLP remain at 8-bit precision across all configurations. We retain normalization layers, residual additions, and the CLIP logit scale in full precision. Softmax operations also remain in full precision, whereas their outputs are quantized with LSQ.

Details of calibration data. We calibrate the image and text encoders using 128 synthetic images and 512 text prompts, respectively. The text prompts are sampled without replacement from 3,240 candidates formed by pairing 180 concepts with 18 CLIP prompt templates. We use OMSE (Choukroun et al., 2019) to initialize the quantization parameters from the first 32 images and 32 text prompts in their respective calibration pools. The full calibration pools are subsequently used for block- or layer-wise reconstruction.

Details of reconstruction. We perform block-wise reconstruction for CNN image encoders and layer-wise reconstruction for Transformer-based encoders. For each reconstruction unit, we minimize the L_2 discrepancy between its quantized output and the corresponding output of the frozen full-precision model. We reconstruct the units of each encoder in input-to-output order, completing the image encoder before proceeding to the text encoder. For each reconstruction unit, we jointly optimize the AdaRound rounding parameters and LSQ activation scales for 20,000 iterations with a batch size of 32. We optimize the rounding parameters and activation scales with separate Adam (Kingma & Ba, 2015) optimizers. The learning rates are set to $1e-3$ for the rounding parameters, $4e-5$ for the activation scales in the image encoder, and $4e-6$ for the activation scales in the text encoder. We apply cosine annealing only to the learning rates for activation scales, decaying them to zero over the 20,000 iterations for each reconstruction unit.

D.3 Evaluation Protocol

We follow the zero-shot evaluation protocol of Tip-Adapter (Zhang et al., 2022) and apply identical class names, fixed prompt templates, model-specific preprocessing, and evaluation splits to all compared methods. The evaluation prompts are distinct from the 512 prompts used for text encoder calibration. For each backbone, we report the mean top-1 accuracy across the 11 datasets.

E Experimental Setup

E.1 Datasets

We evaluate quantized CLIP models on the following 11 downstream classification datasets. Consistent with the data-free setting, real images from these datasets are used only for evaluation and are not

Table 5: Hyperparameter ranges for VISTA.

Hyperparameter	Range
δ	$[0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.97, 0.99, 1.00]$
λ_{view}	$[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$
λ_{app}	$[0.01, 0.1, 1, 10, 100, 1000]$

accessed during synthetic image generation or quantization calibration. The datasets are summarized as follows:

- **Caltech101** (Fei-Fei et al., 2007) is a generic object-recognition dataset containing 100 classes.
- **StanfordCars** (Krause et al., 2013) is a fine-grained recognition dataset containing 196 car classes.
- **DTD** (Cimpoi et al., 2014) evaluates texture recognition across 47 descriptive texture classes.
- **FGVCAircraft** (Maji et al., 2013) contains 100 fine-grained aircraft variants.
- **Flowers102** (Nilsback & Zisserman, 2008) contains 102 fine-grained flower categories.
- **Food101** (Bossard et al., 2014) contains 101 categories of food images.
- **ImageNet** (Deng et al., 2009) is a large-scale benchmark with 1,000 object classes.
- **Oxford-IIIT Pet** (Parkhi et al., 2012) contains 37 fine-grained cat and dog breed classes.
- **EuroSAT** (Helber et al., 2019) contains 10 satellite-image land-cover classes.
- **SUN397** (Xiao et al., 2010) contains 397 indoor and outdoor scene classes.
- **UCF101** (Soomro et al., 2012) contains 101 human-action classes represented by video frames.

E.2 COMPETITORS

We compare VISTA with four baselines for synthetic data generation:

- **Gaussian** directly samples synthetic images from a standard Gaussian distribution without optimizing them against the pre-trained CLIP model.
- **Patch-Similarity Entropy (PSE)** (Li et al., 2022) synthesizes images for ViTs by maximizing the entropy of pairwise similarities among self-attention patch representations.
- **Batch-Normalization Statistics (BNS)** (Cai et al., 2020) synthesizes images for CNNs by matching batch-wise activation statistics to the running statistics stored in the pre-trained model.
- **D4C** (Zhang et al., 2025) is a CLIP inversion method that constructs a predefined text prompt pool and synthesizes images by aligning them with sampled prompts.

E.3 HYPERPARAMETERS

We conduct a hyperparameter search for the identity confidence δ, the view-consistency weight λ_{view}, and the appearance diversity weight λ_{app}. Table 5 summarizes the evaluated values and the settings adopted in all other experiments. We set the remaining inversion-related hyperparameters, such as λ_{arch}, λ_{tv} and λ_{ℓ_2}, based on DeepInversion (Yin et al., 2020) and D4C (Zhang et al., 2025), while retaining the temperature τ of each pre-trained CLIP checkpoint. We configure each competitor using the hyperparameter settings recommended in its original paper.

E.4 IMPLEMENTATION AND HARDWARE

We implement VISTA in PyTorch (Paszke et al., 2019) and utilize the official pre-trained OpenAI CLIP checkpoints (Radford et al., 2021). Our quantization and reconstruction code builds on D4C (Zhang et al., 2025), while our zero-shot evaluation code is adapted from Tip-Adapter (Zhang et al., 2022). All experiments are conducted on a workstation equipped with an Intel Xeon Gold 6534 CPU and an NVIDIA L40S GPU.

Table 6: Top-1 accuracy [%] of quantized CLIP ViT-B/32 across different bit-widths and datasets.

Method	W/A	Cal*	Cars*	DTD	FGVC*	Flow*	Food*	IN*	Pet*	SAT*	SUN*	UCF*	Avg
Full Precision	32/32	91.44	60.17	44.09	19.14	66.71	80.40	63.79	87.41	45.20	62.08	63.49	62.17
Gaussian	8/8	74.12	11.90	23.64	4.23	24.16	17.43	26.85	50.04	9.99	34.49	21.12	27.09
PSE		82.31	26.66	33.75	7.62	45.07	64.42	47.68	72.04	29.51	51.79	46.42	46.12
D4C (Zhang et al., 2025)		83.53	31.81	33.92	7.29	46.45	69.58	49.12	72.74	28.56	53.16	49.59	47.80
VISTA (proposed)		**89.76**	**50.39**	**40.80**	**14.28**	**58.40**	**73.60**	**58.47**	**83.73**	**33.54**	**59.02**	**56.88**	**56.26**
Gaussian	6/6	61.26	4.32	19.39	1.71	16.32	7.71	16.07	35.32	16.60	22.53	13.38	19.51
PSE		83.08	22.20	28.13	5.94	35.61	50.49	40.98	62.01	27.22	46.80	41.85	40.39
D4C (Zhang et al., 2025)		80.32	22.71	32.39	6.33	35.97	58.63	43.84	64.92	**33.52**	50.52	45.12	43.12
VISTA (proposed)		**88.18**	**41.58**	**37.39**	**12.58**	**53.82**	**68.33**	**54.21**	**81.51**	31.98	**56.24**	**53.64**	**52.68**
Gaussian	4/8	54.20	3.72	17.97	1.98	12.79	6.11	12.41	22.02	12.09	18.00	9.20	15.50
PSE		82.64	20.25	27.25	5.07	34.55	48.36	40.01	66.37	21.85	46.20	39.39	39.27
D4C (Zhang et al., 2025)		81.01	23.77	28.90	6.99	31.02	58.26	43.35	67.02	**25.88**	49.58	43.72	41.77
VISTA (proposed)		**87.09**	**37.57**	**35.91**	**11.92**	**48.89**	**61.42**	**51.03**	**78.93**	24.10	**54.21**	**50.36**	**49.22**
Gaussian	4/4	4.26	0.50	4.08	1.02	1.66	1.50	0.59	2.94	3.31	1.14	1.96	2.09
PSE		49.61	1.93	11.58	2.22	6.50	6.08	9.57	19.76	13.15	11.84	11.23	13.04
D4C (Zhang et al., 2025)		57.85	3.23	13.65	1.86	6.41	10.12	12.83	23.58	9.48	16.73	13.98	15.43
VISTA (proposed)		**59.11**	**4.58**	**16.37**	**3.49**	**12.19**	**12.11**	**16.59**	**37.25**	**13.76**	**20.49**	**17.40**	**19.39**

* Cal: Caltech101, Cars: StanfordCars, FGVC: FGVCAircraft, Flow: Flowers102, Food: Food101, IN: ImageNet, Pet: Oxford-IIIT Pet, SAT: EuroSAT, SUN: SUN397, UCF: UCF101.

Table 7: Top-1 accuracy [%] of quantized CLIP ViT-L/14 across different bit-widths and datasets.

Method	W/A	Cal*	Cars*	DTD	FGVC*	Flow*	Food*	IN*	Pet*	SAT*	SUN*	UCF*	Avg
Full Precision	32/32	95.13	76.88	53.01	32.70	79.54	90.91	75.93	93.49	60.21	67.68	74.97	72.77
Gaussian	8/8	89.78	37.30	40.78	9.33	45.96	58.54	51.94	78.63	37.05	49.31	47.16	49.62
PSE		92.66	51.96	41.90	11.34	64.80	82.56	65.20	86.70	44.25	59.99	61.88	60.29
D4C (Zhang et al., 2025)		90.71	55.75	42.38	9.39	62.65	84.57	63.66	84.08	47.09	59.52	62.04	60.17
VISTA (proposed)		**94.81**	**71.77**	**53.19**	**26.13**	**77.06**	**88.90**	**74.16**	**92.83**	**51.31**	**67.05**	**71.50**	**69.88**
Gaussian	6/6	6.45	0.47	4.43	1.20	1.26	1.05	0.27	3.52	8.51	0.79	0.93	2.63
PSE		23.37	1.37	12.71	1.08	4.43	7.32	4.84	21.72	**14.23**	5.97	9.28	9.67
D4C (Zhang et al., 2025)		28.60	1.62	12.29	0.93	1.87	7.07	5.44	15.86	11.22	8.01	4.31	8.84
VISTA (proposed)		**85.52**	**22.99**	**40.96**	**10.47**	**45.55**	**52.66**	**51.12**	**82.72**	11.28	**48.61**	**48.45**	**45.48**
Gaussian	4/8	80.24	14.59	35.46	3.69	26.19	29.45	31.35	49.20	23.25	32.39	29.02	32.26
PSE		90.99	51.18	41.90	10.11	55.34	81.80	62.97	84.22	48.51	59.47	63.05	59.05
D4C (Zhang et al., 2025)		90.51	48.39	44.27	11.25	60.58	80.31	63.18	84.71	**49.96**	58.93	59.27	59.21
VISTA (proposed)		**93.79**	**63.80**	**51.60**	**22.53**	**72.43**	**86.36**	**71.19**	**90.54**	49.11	**65.80**	**68.52**	**66.88**
Gaussian	4/4	0.89	0.49	2.72	0.96	1.06	1.05	0.15	3.11	11.11	0.28	1.08	2.08
PSE		1.14	0.44	1.42	0.99	0.57	0.96	0.10	2.75	10.41	0.21	0.82	1.80
D4C (Zhang et al., 2025)		**5.19**	**0.62**	2.66	0.99	1.42	**1.49**	0.32	2.73	**12.90**	0.63	1.16	**2.74**
VISTA (proposed)		1.87	0.57	**3.37**	**1.14**	**1.71**	1.36	**0.46**	**3.84**	12.30	**0.87**	**1.85**	2.67

* Cal: Caltech101, Cars: StanfordCars, FGVC: FGVCAircraft, Flow: Flowers102, Food: Food101, IN: ImageNet, Pet: Oxford-IIIT Pet, SAT: EuroSAT, SUN: SUN397, UCF: UCF101.

F EXPERIMENTS

F.1 DFQ PERFORMANCE ACROSS DIFFERENT BACKBONES

We present the data-free quantization performance of VISTA across different backbones. Tables 6, 7, 8, 9, 10, 11 report the results on ViT-B/32, ViT-L/14, RN50, RN101, RN50x4, and RN50x16. VISTA outperforms D4C in average accuracy in 23 of the 24 (6 backbones and 4 quantization configurations) evaluated backbone-quantization configurations. We find that PSE and BNS outperform D4C in several settings, highlighting that architecture-specific inversion priors are important alongside text-image alignment for synthesizing effective calibration data.

F.2 EFFECT OF WORDNET CATEGORY FILTERING

We examine the effect of the WordNet category filtering in Section 3.1. We compare the proposed setting, which samples concepts from the four selected noun categories (animal, artifact, food, and plant), with unfiltered sampling from all 26 WordNet noun categories. As shown in Table 12, the filtered setting slightly improves the average accuracy of ViT-B/32 under W8/A8 quantization across 11 datasets from 53.31% to 53.91%.

Table 12: Effect of WordNet category filtering on CLIP ViT-B/32 under W8/A8.

Setting	Avg. [%]	Gain [%p]
Unfiltered	53.31	0.00
Filtered	**53.91**	+0.60

Table 8: Top-1 accuracy [%] of quantized CLIP ResNet50 across different bit-widths and datasets. Avg reports the average accuracy across the datasets.

Method	W/A	Cal*	Cars*	DTD	FGVC*	Flow*	Food*	IN*	Pet*	SAT*	SUN*	UCF*	Avg
Full Precision	32/32	85.76	55.69	42.91	16.98	66.06	77.39	60.34	85.72	36.07	58.89	61.83	58.88
Gaussian	8/8	79.76	23.79	32.74	8.10	46.29	67.05	48.43	67.16	30.63	51.64	47.48	45.73
BNS		81.01	31.45	33.33	6.27	45.31	66.55	49.00	58.33	32.78	51.46	47.71	45.75
D4C (Zhang et al., 2025)		79.55	33.48	35.28	8.76	48.23	69.80	49.56	57.56	**36.21**	52.24	46.50	47.02
VISTA (proposed)		**85.03**	**49.66**	**40.84**	**13.56**	**64.68**	**76.42**	**59.31**	**84.06**	33.20	**57.54**	**58.60**	**56.63**
Gaussian	6/6	70.43	11.04	27.78	2.61	35.81	51.45	38.08	49.01	17.73	42.41	39.02	35.03
BNS		77.20	26.08	31.91	4.56	38.98	59.28	44.67	47.56	24.42	48.32	44.09	40.64
D4C (Zhang et al., 2025)		78.13	25.51	30.50	7.26	38.94	59.84	45.40	64.35	20.22	48.69	46.05	42.26
VISTA (proposed)		**80.20**	**42.53**	**38.95**	**10.23**	**57.04**	**70.60**	**56.01**	**79.99**	**26.98**	**53.48**	**54.61**	**51.87**
Gaussian	4/8	48.36	11.13	22.28	2.67	14.82	19.11	21.80	33.52	23.90	25.67	13.83	21.55
BNS		78.70	26.71	30.67	5.79	38.77	59.51	43.41	54.92	21.88	48.48	43.25	41.10
D4C (Zhang et al., 2025)		77.16	25.76	29.26	5.25	36.58	55.86	42.37	64.05	**22.37**	45.96	38.59	40.29
VISTA (proposed)		**79.23**	**35.93**	**35.93**	**11.61**	**53.71**	**65.89**	**52.62**	**78.93**	15.27	**54.07**	**47.90**	**48.28**
Gaussian	4/4	5.19	0.70	2.84	1.02	1.66	1.07	1.43	4.77	6.99	1.03	2.41	2.65
BNS		42.11	6.12	**5.14**	2.40	4.14	**17.95**	15.98	28.73	8.65	11.87	19.40	14.77
D4C (Zhang et al., 2025)		43.61	4.74	2.84	**3.72**	4.51	14.14	13.80	21.37	11.12	8.99	15.23	13.10
VISTA (proposed)		**43.69**	**8.42**	4.67	2.37	**11.94**	15.32	**18.88**	**38.65**	**13.79**	**12.95**	**20.14**	**17.35**

* Cal: Caltech101, Cars: StanfordCars, FGVC: FGVCAircraft, Flow: Flowers102, Food: Food101, IN: ImageNet, Pet: Oxford-IIIT Pet, SAT: EuroSAT, SUN: SUN397, UCF: UCF101.

Table 9: Top-1 accuracy [%] of quantized CLIP ResNet101 across different bit-widths and datasets. Avg reports the average accuracy across the datasets.

Method	W/A	Cal*	Cars*	DTD	FGVC*	Flow*	Food*	IN*	Pet*	SAT*	SUN*	UCF*	Avg
Full Precision	32/32	89.98	63.24	38.00	18.09	64.23	80.69	62.50	86.89	32.81	59.00	60.88	59.66
Gaussian	8/8	83.49	33.53	33.69	7.56	47.42	69.34	50.81	70.67	29.23	51.62	50.28	47.97
BNS		84.14	39.15	35.28	6.18	46.12	71.00	51.03	71.79	21.21	50.14	49.59	47.78
D4C (Zhang et al., 2025)		84.18	39.30	33.27	6.33	49.41	70.63	51.10	76.53	23.48	50.49	50.41	48.65
VISTA (proposed)		**88.80**	**57.64**	**39.83**	**16.59**	**63.01**	**79.69**	**61.32**	**85.64**	**31.16**	**58.55**	**58.71**	**58.27**
Gaussian	6/6	77.81	23.99	30.02	6.54	35.57	57.48	42.95	60.51	24.56	46.16	43.14	40.79
BNS		82.43	33.07	32.45	6.45	42.67	65.17	48.92	72.66	**30.69**	49.37	47.63	46.50
D4C (Zhang et al., 2025)		83.81	34.95	31.86	6.72	40.11	58.16	47.81	68.82	22.10	47.40	45.44	44.29
VISTA (proposed)		**88.28**	**52.48**	**38.00**	**15.81**	**58.42**	**77.05**	**58.73**	**83.02**	22.69	**56.43**	**56.62**	**55.23**
Gaussian	4/8	71.52	15.73	26.65	5.07	33.62	43.61	36.85	62.33	9.28	39.97	36.11	34.61
BNS		81.30	35.47	31.86	6.54	40.03	60.32	45.38	64.35	**23.84**	49.72	48.96	44.34
D4C (Zhang et al., 2025)		82.23	29.57	30.56	6.33	36.54	59.84	44.58	61.08	22.93	45.97	43.46	42.10
VISTA (proposed)		**85.11**	**48.34**	**36.17**	**13.29**	**55.14**	**72.72**	**55.66**	**80.73**	23.01	**55.84**	**55.86**	**52.90**
Gaussian	4/4	16.96	1.42	4.26	1.62	2.64	1.47	2.88	8.91	11.86	2.35	3.62	5.27
BNS		58.26	11.13	7.15	2.85	5.68	30.91	23.46	35.05	**15.73**	20.97	28.39	21.78
D4C (Zhang et al., 2025)		54.32	11.98	6.97	2.67	7.19	24.70	20.33	34.40	8.60	17.39	21.99	19.14
VISTA (proposed)		**63.37**	**17.12**	**8.87**	**5.22**	**13.97**	**35.38**	**28.19**	**50.01**	12.64	**20.94**	**30.50**	**26.02**

* Cal: Caltech101, Cars: StanfordCars, FGVC: FGVCAircraft, Flow: Flowers102, Food: Food101, IN: ImageNet, Pet: Oxford-IIIT Pet, SAT: EuroSAT, SUN: SUN397, UCF: UCF101.

Table 10: Top-1 accuracy [%] of quantized CLIP ResNet50x4 across different bit-widths and datasets. Avg reports the average accuracy across the datasets.

Method	W/A	Cal*	Cars*	DTD	FGVC*	Flow*	Food*	IN*	Pet*	SAT*	SUN*	UCF*	Avg
Full Precision	32/32	88.44	68.15	50.89	21.45	70.28	83.61	66.40	88.88	30.85	62.24	61.91	63.01
Gaussian	8/8	86.00	40.14	39.30	9.96	50.79	72.88	56.66	80.19	29.47	54.84	52.97	52.11
BNS		86.37	40.34	39.60	8.31	53.76	76.87	55.96	79.86	**40.27**	54.40	51.76	53.41
D4C (Zhang et al., 2025)		85.15	43.88	37.06	7.77	49.61	76.99	54.85	76.26	36.86	54.45	54.53	52.49
VISTA (proposed)		**87.95**	**63.16**	**47.99**	**16.68**	**68.94**	**82.72**	**65.13**	**87.76**	37.07	**61.36**	**60.75**	**61.77**
Gaussian	6/6	81.46	25.59	34.04	7.38	43.08	69.16	48.25	71.19	30.91	48.55	48.08	46.15
BNS		84.30	37.67	35.52	8.04	41.21	72.19	51.36	73.32	**35.72**	51.85	52.52	49.43
D4C (Zhang et al., 2025)		84.10	34.72	35.22	7.62	40.76	67.32	51.88	73.97	30.90	50.41	52.76	48.15
VISTA (proposed)		**86.69**	**58.91**	**46.51**	**15.39**	**64.96**	**80.77**	**62.94**	**86.18**	26.35	**59.05**	**59.71**	**58.86**
Gaussian	4/8	78.54	21.55	33.10	4.68	34.15	53.55	43.77	66.86	18.54	45.54	38.44	39.88
BNS		82.27	35.53	36.05	6.57	42.06	71.11	49.86	74.13	**29.81**	53.02	52.74	48.47
D4C (Zhang et al., 2025)		82.39	28.74	33.33	6.00	42.96	69.37	48.78	73.43	24.52	51.76	49.09	46.40
VISTA (proposed)		**84.91**	**54.55**	**43.20**	**15.24**	**60.41**	**76.49**	**60.70**	**83.18**	26.14	**59.35**	**56.70**	**56.44**
Gaussian	4/4	7.55	0.76	2.13	0.93	1.91	2.40	1.13	6.68	9.28	1.16	1.30	3.20
BNS		61.83	15.46	11.58	3.87	11.00	38.96	24.93	40.20	**10.02**	19.68	31.14	24.42
D4C (Zhang et al., 2025)		54.56	10.15	6.62	2.91	9.18	25.50	20.89	40.28	9.35	15.41	28.71	20.32
VISTA (proposed)		**63.12**	**20.63**	**13.06**	**6.03**	**18.68**	**43.22**	**29.77**	**63.89**	9.51	**21.54**	**32.99**	**29.31**

* Cal: Caltech101, Cars: StanfordCars, FGVC: FGVCAircraft, Flow: Flowers102, Food: Food101, IN: ImageNet, Pet: Oxford-IIIT Pet, SAT: EuroSAT, SUN: SUN397, UCF: UCF101.

Table 11: Top-1 accuracy [%] of quantized CLIP ResNet50x16 across different bit-widths and datasets. Avg reports the average accuracy across the datasets.

Method	W/A	Cal*	Cars*	DTD	FGVC*	Flow*	Food*	IN*	Pet*	SAT*	SUN*	UCF*	Avg
Full Precision	32/32	89.82	72.55	50.18	27.24	72.19	88.01	70.95	90.05	38.62	63.79	66.32	66.34
Gaussian	8/8	83.29	49.57	40.48	9.81	59.16	83.30	62.35	81.30	**33.56**	58.12	57.78	56.25
BNS		85.76	51.21	40.07	9.84	59.76	82.72	62.22	82.28	31.36	57.53	59.19	56.54
D4C (Zhang et al., 2025)		85.60	51.77	38.12	8.91	58.95	81.17	62.24	82.53	28.83	56.72	58.90	55.79
VISTA (proposed)		**89.37**	**68.56**	**51.54**	**23.10**	**71.38**	**87.48**	**70.49**	**89.13**	31.12	**64.09**	**67.25**	**64.86**
Gaussian	6/6	80.89	38.19	36.29	6.93	41.78	74.37	54.10	69.17	27.06	51.69	52.08	48.41
BNS		83.65	47.56	35.70	7.38	48.56	79.23	58.27	74.30	28.31	50.91	56.78	51.88
D4C (Zhang et al., 2025)		81.58	43.96	36.82	8.94	49.74	77.02	56.18	75.12	28.57	51.86	55.49	51.39
VISTA (proposed)		**87.46**	**64.00**	**46.75**	**21.87**	**66.67**	**85.86**	**68.22**	**87.33**	**39.27**	**59.38**	**64.29**	**62.83**
Gaussian	4/8	78.66	33.70	37.94	8.55	45.23	70.26	52.76	66.86	31.02	51.63	51.31	47.99
BNS		83.45	44.40	39.01	9.60	54.41	75.79	58.16	73.43	27.22	54.35	53.26	52.10
D4C (Zhang et al., 2025)		85.19	51.73	38.48	9.33	51.16	75.73	58.14	75.25	29.14	55.43	54.27	53.08
VISTA (proposed)		**89.66**	**61.44**	**47.34**	**19.11**	**65.04**	**84.38**	**67.06**	**87.63**	**35.68**	**62.02**	**62.15**	**61.96**
Gaussian	4/4	1.54	0.60	2.66	0.93	1.38	1.14	0.62	5.07	11.57	0.28	0.90	2.43
BNS		34.77	**13.17**	3.96	**3.09**	2.76	21.50	**22.96**	29.14	**18.17**	**11.75**	**28.50**	17.25
D4C (Zhang et al., 2025)		23.53	9.25	2.54	2.88	4.59	21.32	17.67	**37.83**	10.42	9.32	25.88	15.02
VISTA (proposed)		**36.39**	9.55	**6.32**	3.06	**7.84**	**22.61**	22.18	31.34	16.27	11.68	26.04	**17.57**

* Cal: Caltech101, Cars: StanfordCars, FGVC: FGVCAircraft, Flow: Flowers102, Food: Food101, IN: ImageNet, Pet: Oxford-IIIT Pet, SAT: EuroSAT, SUN: SUN397, UCF: UCF101.

F.3 Cross-Combination Analysis of Prompts and Image Optimization

We evaluate the individual and joint effects of prompt selection and image optimization on ViT-B/32 through all four combinations of the prompt sets and image optimization of D4C and VISTA. We measure MMD^2 and embedding spread for each combination. A lower MMD^2 indicates a smaller discrepancy between the synthetic and real image embedding distributions, whereas embedding spread measures within-set dispersion and is compared with the corresponding real-image value. Table 13 shows that the full VISTA combination achieves the lowest MMD^2 and spread. Compared with the full D4C combination, replacing only the prompts with those of VISTA reduces MMD^2 by 0.005 and the spread by 0.02. Replacing only the image optimization yields larger corresponding reductions of 0.096 and 0.22. These results indicate that both components improve image distribution matching, with image optimization producing the larger improvement.

Table 13: Cross-combination analysis of prompt selection and image optimization. Details in Section F.3.

Prompt	Image	MMD^2 (↓)	Spread
Real		–	0.523
D4C	D4C	0.175	0.788
VISTA	D4C	0.170	0.768
D4C	VISTA	0.079	**0.568**
VISTA	VISTA	**0.076**	**0.568**

G Limitations and Future Directions

We discuss the limitations of VISTA and outline promising directions for future research. We first examine the current scope and assumptions of the framework and then consider how it can be extended to broader tasks, compression settings, and model architectures.

G.1 Limitations

Limited contribution to quantization methodology. We primarily contribute a CLIP inversion framework for generating synthetic calibration data rather than a new quantization algorithm. We adopt an existing quantization scheme and layer-wise reconstruction procedures to evaluate the effectiveness of the generated synthetic data for quantization. Accordingly, quantizer design and quantization-specific optimization are outside the scope of this work. Nevertheless, the independence of VISTA from the quantization scheme offers a practical advantage, enabling it to be readily combined with a wide range of quantization methods. Despite this flexibility, the substantial performance degradation under aggressive quantization settings such as W4/A4 suggests that synthetic calibration data quality is not the only bottleneck and that the quantization method itself may constitute another major bottleneck. Future work should therefore extend beyond model inversion to address the broader problem of DFQ for CLIP models.

G.2 Future Directions

We consider three directions for extending VISTA: broader vision-language tasks, data-free compression beyond quantization, and visual encoders in vision-language models (VLMs).

Broader vision-language tasks. We focus on zero-shot image classification and evaluate the effectiveness of VISTA across diverse image domains. CLIP models also support various vision-language tasks beyond image classification, such as image-text retrieval (Baldrati et al., 2022; Chen et al., 2023), object detection (Minderer et al., 2022; Wu et al., 2023), and segmentation (Rao et al., 2022; Wang et al., 2022). Whether the gains obtained by VISTA transfer to these tasks remains unexplored. Future work can evaluate VISTA on these tasks and examine whether its synthetic data preserve the cross-modal representations required after quantization.

Data-free compression beyond quantization. We currently utilize the synthetic image-text pairs generated by VISTA for quantization of CLIP models. These pairs can also provide surrogate data for other data-free compression methods, including pruning (Yin et al., 2020; Tang et al., 2021) and knowledge distillation (Nayak et al., 2019; Wang, 2021; Yu et al., 2023; Liu et al., 2024). Future work can examine whether they preserve the activation statistics and cross-modal supervision required by these methods, thereby establishing the applicability of VISTA beyond quantization.

Extension to VLM image encoders. Many generative VLMs employ CLIP-based visual encoders to extract image representations (Liu et al., 2023; Li et al., 2023b; Zhang et al., 2026). For models that retain a CLIP-derived visual encoder, future work could utilize its paired text encoder to guide synthetic-data generation and then utilize the resulting images to calibrate the visual encoder. Extending VISTA to visual encoders without a paired text encoder would require adapting its inversion objective to the multimodal projector and language model. These extensions would enable an evaluation of whether the generated data preserve visual representations and downstream multimodal performance.