QuARI: Query Adaptive Retrieval Improvement

Eric Xing[1] **Abby Stylianou**[2] **Robert Pless**[3] **Nathan Jacobs**[1]
[1]Washington University in St. Louis [2]Saint Louis University [3]George Washington University

Abstract

Massive-scale pretraining has made vision-language models increasingly popular for image-to-image and text-to-image retrieval across a broad collection of domains. However, these models do not perform well when used for challenging retrieval tasks, such as instance retrieval in very large-scale image collections. Recent work has shown that linear transformations of VLM features trained for instance retrieval can improve performance by emphasizing subspaces that relate to the domain of interest. In this paper, we explore a more extreme version of this specialization by learning to map a given query to a query-specific feature space transformation. Because this transformation is linear, it can be applied with minimal computational cost to millions of image embeddings, making it effective for large-scale retrieval or re-ranking. Results show that this method consistently outperforms state-of-the-art alternatives, including those that require many orders of magnitude more computation at query time. Code and pre-trained models are available at `https://github.com/mvrl/QuARI`.

1 Introduction

Recent advances in language-image pretraining have significantly improved performance in various vision-language tasks, including image-to-image and text-to-image retrieval. These models learn to align images and their corresponding textual descriptions in a shared embedding space over large-scale datasets. Yet, despite their success, they show notable limitations in retrieval scenarios.

Pretrained models often rely on global image features that encapsulate the overall content of an image. While effective for general classification tasks, these global representations may not capture fine-grained details essential for distinguishing between visually similar images based on specific textual queries [51]. This limits retrieval performance when nuanced differences are critical.

To address the shortcomings of global feature representations, re-ranking methods have been proposed, which involve a secondary, more expensive analysis of the top retrieval candidates. Techniques such as two-stage retrieval pipelines [29] and re-ranking transformers [38] aim to refine initial retrieval results. However, these approaches often entail substantial computational overhead, making them impractical for real-time applications or large-scale deployments.

In this work, we propose a novel approach that integrates *query-specific* embedding projections into the retrieval process. By dynamically adjusting the embedding space based on the input, our method prioritizes the most relevant fine-grained semantic alignments between text and image features for each specific query. Figure 1 contrasts our query-specific retrieval approach with non-specific retrieval using a foundational vision-language model like CLIP [30] and domain-specific retrieval that has been adapted for a given dataset or task. Our framework, Query Adaptive Retrieval Improvement (QuARI), enhances retrieval accuracy without incurring the high computational costs associated with traditional re-ranking methods. Our specific contributions are as follows:

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

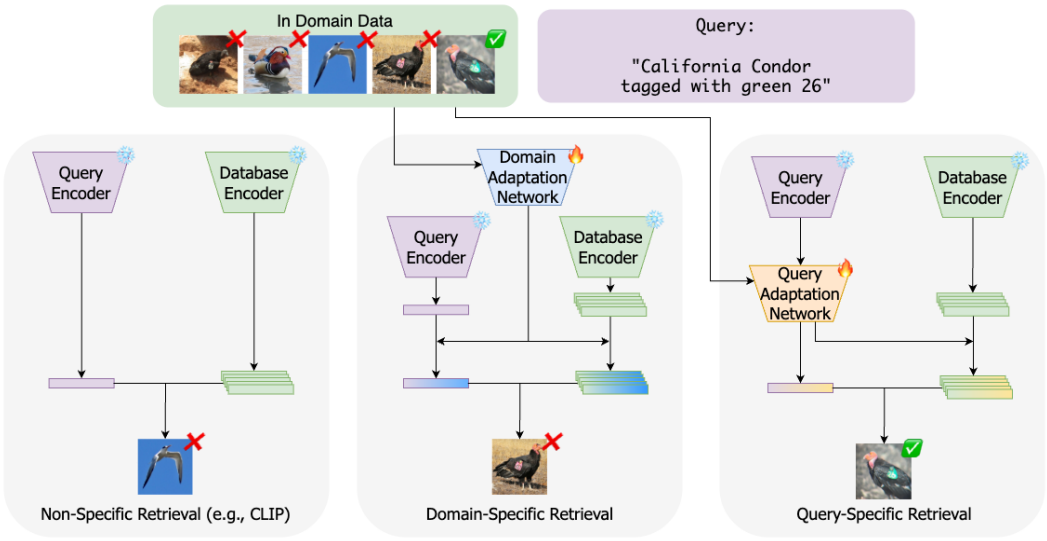

Figure 1: We propose a new query-specific approach to retrieval, QuARI. QuARI dynamically adapts embeddings *per-query* to significantly improve retrieval performance compared to non-specific retrieval with general-purpose embedding features like CLIP, and domain-specific retrieval with transformations learned for a specific domain, with little computational overhead. Figure 2 shows the details of the Query Adaptation Network.

- We identify and articulate the limitations of contrastively pretrained models in capturing fine-grained details necessary for accurate retrieval, and analyze the inefficiencies of existing re-ranking methods and their impact on retrieval performance.
- We introduce QuARI, an embedding projection framework that adapts global representations to individual queries, improving accuracy while maintaining computational efficiency.
- We demonstrate that QuARI yields large improvements in retrieval accuracy on multiple extremely challenging retrieval tasks.

2 Related Work

Text-Image Pretraining and Global Embeddings. Foundation models based on language-image pretraining, such as CLIP [30] and the SigLIP family of models [41, 50], have achieved remarkable success in aligning visual and textual modalities. These models learn global representations by maximizing the similarity between paired image-text inputs. However, the reliance on global features can be detrimental in retrieval tasks that require fine-grained discrimination. Methods like RegionCLIP [51] attempt to address this by incorporating region-level features, but challenges remain in capturing nuanced details essential for accurate retrieval. FILIP introduces token-wise late interaction to capture patch–word similarities [47], while SPARC sparsifies such interactions for efficiency in large-scale pre-training [5]. Region-centric pre-training further bridges image-level and region-level semantics for open-vocabulary detection [15].

These methods highlight a growing recognition of the limitations of global embeddings and attempt to address them through increasingly sophisticated alignment mechanisms. However, they still largely rely on a single static representation per image. In contrast, our method is dynamic, modifying the representation space per query to prioritize the learned aspects that are relevant for the query while de-emphasizing irrelevant aspects.

Efficient Late-Interaction and Re-ranking Paradigms. Late-interaction retrievers (e.g., ColBERT [14]) decouple encoder computation from expensive cross-attentions, enabling scalable yet fine-grained passage search. In vision, there have been a variety of approaches to re-ranking, including local feature based geometric re-ranking approaches [28, 33, 34, 37], and the recent Re-ranking Transformers [38], which refines top-k candidates with lightweight self-attention.

Methods like two-stage retrieval pipelines [29] and re-ranking transformers [38] refine initial retrieval results by re-evaluating top candidates with more sophisticated models. While effective, these approaches introduce significant computational overhead, making them unsuitable for real-time applications or large-scale systems. Our approach instead shifts the complexity into a lightweight projection computed once per query. The approach is computationally lightweight enough that it can be used to re-rank very large numbers of candidates, or even entire datasets.

Hypernetworks. Hypernetworks, or networks that predict the weights of other networks, were first introduced by Ha *et al.* to generate recurrent neural network parameters on the fly [12]. Recent advancements have explored the use of hypernetworks for personalization in generative and task-specific models. HyperDreamBooth [32] introduces a hypernetwork that generates customized weights for text-to-image diffusion models, enabling the synthesis of subject-specific images with minimal data. HyperCLIP [1] employs a hypernetwork to generate the weights of a task-specific image encoder. These approaches demonstrate the potential of hypernetworks in capturing customized semantics efficiently. However, reliance on task-level customization of entire encoders, as in HyperCLIP, yields only modest performance gains on challenging retrieval tasks. Generating complete sets of encoder weights is both computationally expensive and difficult to optimize. In contrast, we show that using hypernetworks to adapt off-the-shelf features with lightweight, query-specific transformations can achieve strong performance without significant computational overhead.

Transformers as Hypernetworks. The concept of using transformers as hypernetworks has been explored in various domains. For example, Transformers as Meta-Learners [7] leverage transformers to predict the weights of implicit neural representations, showcasing their capability in dynamic weight generation. We draw inspiration from these works, employing a transformer to predict query-specific projection matrices. In the following section, we discuss how we tailor this idea to the retrieval setting.

3 Methodology

Our approach introduces a *transformer-based hypernetwork* for customized text-to-image retrieval. Unlike traditional methods relying on static embeddings, QuARI predicts customized linear projections conditioned on each query embedding to adapt database features for each query.

3.1 Hypernetwork-Augmented Retrieval

We begin by extracting global embeddings for queries and candidate images using a pretrained vision-language model, such as CLIP [30] or SigLIP [50]. Formally, for a query q and a set of "gallery" images $\mathcal{G} = \{I_n | n = 1, 2, ..., N\}$, we obtain embeddings for a query $\mathbf{q}_i = \text{Enc}(q_i)$ and a database of gallery embeddings $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_N\} = \{\text{Enc}(I_1), \text{Enc}(I_2), ...\text{Enc}(I_N)\}$. We learn a hypernetwork H_θ as follows:

$$(\mathbf{q}'_i, T_i) = H_\theta(\mathbf{q}_i), \tag{1}$$

Which outputs a transformed query representation $\mathbf{q}'_i$ and a transformation function $\mathbf{T}_i$. Retrieval is then performed by transforming the database:

$$\mathcal{D}'_i = \{\mathbf{d}'_1, \mathbf{d}'_2, ..., \mathbf{d}'_N\} \tag{2}$$

$$= \{T(\mathbf{d}_1), T(\mathbf{d}_2), ..., T(\mathbf{d}_N)\} \tag{3}$$

The transformed database of embeddings $\mathcal{D}'_i$ may then be used to perform retrieval by selecting database indices maximizing some embedding similarity between the query $\text{sim}(\mathbf{q}', \mathbf{d}'_i)$.

3.2 Hypernetwork Architecture

Our query adaptation network, shown in Figure 2, is a hypernetwork which we will denote as H_θ. It is a transformer-based module that conditions on the query embedding $\mathbf{q} \in \mathbb{R}^E$ and predicts both a customized query embedding $\mathbf{q}' \in \mathbb{R}^E$ and a transformation matrix $T \in \mathbb{R}^{E \times E}$ to adapt image embeddings of dimension E. To constrain computation and promote generalization, we parameterize T as a low-rank matrix of rank $r = 64$.

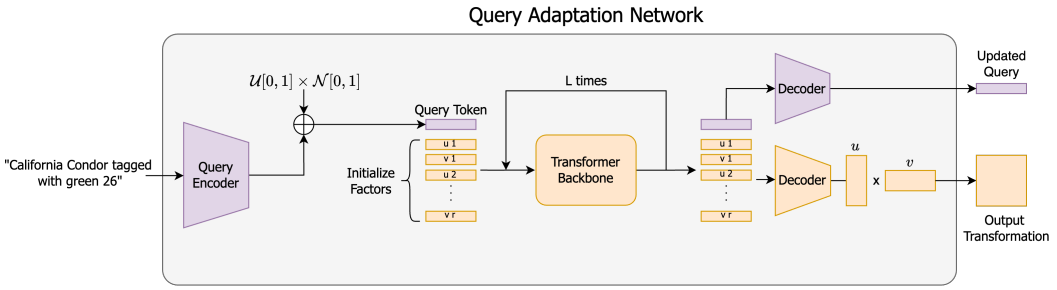

Figure 2: An overview of our query adaptation network. A zero initialization of the transformation matrix is tokenized by columns and passed to a transformer backbone with a conditioning token to obtain refined columns. This process is repeated L times, refining the transformation.

Tokenization of the Projection Matrix. We define the transformation matrix T via a learned low-rank decomposition:

$$T = \sum_{j=1}^{r} \mathbf{u}_j \mathbf{v}_j^\top, \tag{4}$$

where each pair $(\mathbf{u}_j, \mathbf{v}_j) \in \mathbb{R}^E \times \mathbb{R}^E$ represents a learned rank-one component of the projection matrix. These pairs are generated from a shared set of column-wise tokens, which are iteratively refined via a transformer encoder. Intuitively, each token pair $(\mathbf{u}_j, \mathbf{v}_j)$ defines a single direction in the transformed feature space, contributing one semantic feature to the customized representation.

Initialization and Conditioning. We initialize a bank of $2r$ tokens per sample in the batch: r *U-tokens* for projecting into the output space and r *V-tokens* for selecting features from the input space. The initial token representations are zero-initialized and refined over L denoising steps. To condition the generation process on the query embedding, we first encode $\mathbf{q}$ using a two-layer MLP and add a learned timestep embedding to capture iterative update dynamics.

Iterative Refinement with Transformers. At each step $t \in \{1, ..., L\}$, we concatenate the query-conditioned control token with the current U/V token sequence and apply a shared transformer encoder. We also apply sinusoidal positional encodings [42] to the token sequence. The output of the transformer updates the tokens via residual addition:

$$\mathbf{u}_j^{(t+1)} = \mathbf{u}_j^{(t)} + \Delta\mathbf{u}_j^{(t)}, \tag{5}$$

$$\mathbf{v}_j^{(t+1)} = \mathbf{v}_j^{(t)} + \Delta\mathbf{v}_j^{(t)}. \tag{6}$$

Decoding to Projection Matrices. Once the tokens are fully refined, we decode each U-token and V-token using separate MLP decoders to produce vectors in $\mathbb{R}^E$. The final transformation matrix is then constructed as:

$$T = \sum_{j=1}^{r} \text{MLP}_\text{u}(\mathbf{u}_j) \cdot \text{MLP}_\text{v}(\mathbf{v}_j)^\top. \tag{7}$$

The customized query representation $\mathbf{q}'$ is decoded from the control token by another MLP.

3.3 Training

We train the model using a symmetric contrastive loss over the transformed query and image embeddings. Given a minibatch of B text-image embeddings $\{(\mathbf{q}_i, \mathbf{d}_i)\}_{i=1}^B$, we first add noise to every query embedding to help bridge the text-image modality gap. We borrow this formulation from LinCIR [11]. This acts as a regularizer that also allows high performance with image queries, training only on text-image datasets.

$$\mathbf{q}_i \leftarrow \mathbf{q}_i + \mathcal{U}[0,1] \times \mathcal{N}[0,1] \tag{8}$$

Then, we compute the customized query text embedding $\mathbf{q'}_i$ and the personalization matrix T_i using the hypernetwork H_θ.

Semi-Positive Sample Mining. In the standard contrastive formulation, there is only one positive sample for each negative sample. This leads the learned transformations to overfit to the query. This results in samples that are completely different from the target and samples which share some, but not all, attributes with the target having near-equivalent similarities to the transformed query.

In order to address this, we also compute a set of "semi-positive" samples $\mathcal{P}_i$ for each target image. For every target image we compute a set of 100 nearest-neighbors using precomputed backbone embeddings and select the top 2 as "semi-positives." We apply a softmax over the distribution of these 100 cosine similarities and use their logits as target similarity values.

To train QuARI, we optimize a symmetric contrastive objective, as is standard for retrieval with semi-positive samples:

$$p_{i,j} = \frac{\exp(S_{i,j})}{\sum_{k=1}^{B} \exp(S_{i,k})}, \qquad q_{j,i} = \frac{\exp(S_{j,i})}{\sum_{k=1}^{B} \exp(S_{k,i})}, \qquad S_{i,j} = \frac{\mathbf{q'}_i \cdot \mathbf{d}_j^i}{\tau}, \tag{9}$$

$$\mathcal{L} = \frac{1}{2B} \sum_{i=1}^{B} \left[-\sum_{j=1}^{B} \alpha_{i,j} \log p_{i,j} - \sum_{j=1}^{B} \alpha_{j,i} \log q_{j,i} \right], \quad \alpha_{i,j} = \begin{cases} 1 & \text{if } j = i \text{ (positive)}, \\ w_{i,j} & \text{if } j \in \mathcal{P}_i \text{ (semi-positive)}, \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

Where $\mathbf{d}_i^j$ is the i-th target image embedding transformed by the j-th personalization transformation T_j, and τ is a temperature parameter. We only consider similarities where the embedding similarity of a transformed query $\mathbf{q'}_i$ is only computed with target image embeddings transformed by T_i.

The incorporation of semi-positive samples discourages the behavior of overfitting to training query-target pairs, and encourages lower ranked images that seem visually similar to the target to be returned with higher similarity than images that lack visual similarity.

Implementation Details We use the AdamW optimizer [23] with a cosine annealed learning rate cycling between `1e-5` and `2e-7` and a weight decay of `1e-2`. Experiments were conducted on an NVIDIA H100 with 80GB of VRAM. At inference, retrieval is performed by computing cosine similarity scores between the L2-normalized adapted embeddings of a query and database images.

4 Evaluation

Evaluation Datasets. We focus our evaluation on two challenging benchmarks: ILIAS and INQUIRE. ILIAS (Instance-Level Image retrieval At Scale) is a large-scale dataset designed to assess instance-level image retrieval capabilities [16]. It has 1,000 object instances, each represented by query and positive images, totaling 5,947 manually collected images. To evaluate retrieval performance under large-scale settings, ILIAS includes 100 million distractor images from YFCC100M [40]. It also includes a retrieval task over a curated set of 5M distractor images, and an image-to-image retrieval re-ranking task.

INQUIRE [43] is a text-to-image retrieval benchmark tailored for expert-level ecological queries. It is built on the iNaturalist 2024 dataset, containing five million natural world images across 10,000 species. The benchmark features 250 expert-crafted queries spanning categories such as species identification, context, behavior, and appearance. INQUIRE evaluates two retrieval tasks: INQUIRE-Fullrank, requiring models to perform retrieval over the entire dataset, and INQUIRE-Rerank, focusing on refining initial retrieval results.

We also provide results on popular text-to-image and image-to-image retrieval benchmarks COCO [19], Flickr30k [48], FORB [44], and TextCaps [35].

Training Datasets. Our framework can be trained on any paired text-image dataset. We utilize Microsoft Common Objects in Context (MS COCO) [19], Conceptual Captions 12M, and synthetically augmented BioTrove [46] to train QuARI. MS COCO includes over 330,000 images annotated with

model	backbone	resolution	I2I @ 100M	T2I @ 5M	T2I @ 100M
OAI CLIP	ViT-B	224	4.2	2.7	1.6
OAI CLIP + QuARI	ViT-B	224	8.4 (+4.2)	11.5 (+8.8)	9.2 (+7.6)
SigLIP	ViT-B	512	16.6	14.6	11.1
SigLIP + QuARI	ViT-B	512	24.6 (+8.0)	28.9 (+14.3)	25.6 (+14.5)
SigLIP2	ViT-B	512	15.4	14.6	10.4
SigLIP2 + QuARI	ViT-B	512	25.6 (+10.2)	30.7 (+16.1)	27.2 (+16.8)
OpenCLIP	ViT-L	384	9.4	9.4	7.0
OpenCLIP + QuARI	ViT-L	384	15.6 (+6.2)	20.9 (+11.5)	18.7 (+11.7)
OAI CLIP	ViT-L	336	9.4	8.4	5.8
OAI CLIP + QuARI	ViT-L	336	15.9 (+6.5)	20.6 (+12.2)	18.9 (+13.1)
SigLIP	ViT-L	384	19.6	22.2	18.1
SigLIP + QuARI	ViT-L	384	30.9 (+11.3)	36.5 (+14.3)	34.2 (+16.1)
SigLIP2	ViT-L	512	20.8	24.7	19.8
SigLIP2 + QuARI	ViT-L	512	36.2 (+15.4)	41.2 (+16.5)	38.7 (+18.9)

(a) Comparison to baselines.

model	I2I @ 100M
OAI CLIP + TA	7.9
OAI CLIP + QuARI	21.0 (+13.1)
SigLIP + TA	23.0
SigLIP + QuARI	42.3 (+19.3)
SigLIP2 + TA	23.5
SigLIP2 + QuARI	43.4 (+19.9)
OpenCLIP + TA	13.7
OpenCLIP + QuARI	28.4 (+14.7)
OAI CLIP + TA	15.2
OAI CLIP + QuARI	31.1 (+15.9)
SigLIP + TA	28.9
SigLIP + QuARI	45.8 (+16.9)
SigLIP2 + TA	31.3
SigLIP2 + QuARI	50.4 (+19.1)

(b) Comparison to static task adaptation (TA).

Table 1: We show mAP@1k for image-to-image (I2I) and text-to-image (T2I) retrieval on ILIAS.

model	COCO T2I R@1	Flickr30k T2I R@1	FORB mAP@5	FORB t-mAP@5	TextCaps T2I R@1
SigLIP2	55.2	85.3	93.74	69.24	44.6
FT SigLIP2	55.0	85.5	92.89	70.03	45.2
SigLIP2+QuARI	**77.4**	**92.9**	**95.67**	**78.53**	**55.8**

Table 2: Comparison on other benchmarks with SigLIP2 ViT-L @ 512.

five human-written captions per image. Conceptual Captions 12M [6] is a collection of approximately 12 million image-text pairs harvested from the web. BioTrove [46] contains 161.9 million images across approximately 366,600 species, each annotated with taxonomic data. We extract a random subset of 5M samples from BioTrove for training. Since the BioTrove dataset only includes taxonomic and common-name level annotations, we use `Qwen2.5-VL-7B-Instruct` to caption a 500K subset of BioTrove, providing natural-language captions that include visual information beyond species identity (the prompt for constructing these captions is in the supplemental materials).

Metrics. On ILIAS, we measure mean Average Precision @1k (mAP@1k) across both the full 100M distractor set (@100M) and the 5M mini distractor set (@5M). On INQUIRE, we measure mean Average Precision @50 (mAP@50), Normalized Discounted Cumulative Gain @50 (nDCG@50), and Mean Recall Rank (MRR).

Baseline Models. We use popular contrastively pretrained backbone models CLIP [30], SigLIP [50], OpenCLIP [13], and SigLIP2 [41] as backbone feature encoders. We replicate the baselines published with ILIAS for image-to-image re-ranking built on local feature descriptors from DINOv2 [27]. These include query expansion-based methods like αQE [8], local feature and geometric-based matching-based methods like Chamfer Similarity (CS) and Spatial Verification (SP) [45], and transformer-based methods like AMES [37]. αQE-k refers to query expansion with k nearest-neighbors. We also replicate baselines from INQUIRE using vision-language models (VLMs) as re-rankers, including open-source VLMs LLaVA [20–22], VILA [18], PaliGemma [4], InstructBLIP [9], and BLIP2 [17].

5 Results

5.1 Embedding-based Retrieval

ILIAS. In Table 1a, we report the retrieval performance using mAP@1k on the image-to-image and text-to-image tasks on ILIAS. For each baseline row, there is a corresponding row building QuARI

model	backbone	resolution	mAP@50	nDCG@50	MRR
OAI CLIP	ViT-B	224	10.4	20.9	0.40
OAI CLIP + QuARI	ViT-B	224	15.7 (+5.3)	25.4 (+4.5)	0.44 (+0.04)
OAI CLIP	ViT-L	336	23.4	37.7	0.59
OAI CLIP + QuARI	ViT-L	336	32.5 (+9.1)	43.6 (+5.9)	0.64 (+0.05)
SigLIP	ViT-L	384	31.1	46.6	0.68
SigLIP + QuARI	ViT-L	384	41.3 (+10.2)	54.7 (+8.1)	0.72 (+0.04)
SigLIP	SoViT-400m	384	34.2	49.1	0.69
SigLIP + QuARI	SoViT-400m	384	45.4 (+11.2)	56.8 (+7.7)	0.74 (+0.05)

Table 3: Comparison of QuARI and baselines on INQUIRE

Re-ranking Method	mAP@1k
Initial Ranking	19.6
αQE1	22.1
αQE2	20.4
αQE5	14.3
CS (Chamfer Similarity)	22.9
SP (Spatial Verification)	21.8
AMES	26.4
QuARI	**29.1**

(a) ILIAS Top-1k Re-ranking

Re-ranking Method	mAP@50	nDCG@50	MRR
Initial ranking	33.3	48.8	0.69
Best possible re-rank	65.6	72.7	0.96
Open-source VLMs			
BLIP-2 FLAN-T5-XXL	31.2	46.5	0.58
InstructBLIP-T5-XXL	33.0	48.3	0.64
PaliGemma-3B-mix-448	35.6	50.6	0.68
LLaVA-1.5-13B	32.2	47.9	0.64
LLaVA-v1.6-7B	32.3	47.9	0.62
LLaVA-v1.6-34B	35.7	51.2	0.69
VILA-13B	35.7	50.8	0.65
VILA-40B	40.2	54.6	0.72
Closed Source VLMs			
GPT-4V	36.5	51.9	0.72
GPT-4o	43.7	**57.9**	**0.78**
SigLIP2 + QuARI	**45.7**	55.2	0.76

(b) INQUIRE Top-100 Re-ranking

Table 4: Comparison of Re-ranking Methods on ILIAS and INQUIRE

on the same frozen baseline as a backbone feature encoder. Across all backbones, QuARI provides a strong performance boost from off-the-shelf encoders commonly used for retrieval.

Insufficiency of Static Task Adaptation. A common method to improve the performance of large pretrained encoders is to adapt their features with a simple projection operation learned using a dataset that is relevant to the specific task [24, 31]. While this allows models to be adapted for that general task, we show that learning a task adaptation performs significantly worse than QuARI's query-specific adaptations. The authors of ILIAS show that a task adaptation trained on a sample of 1M images from the Universal Embeddings dataset [49] improves image-to-image instance retrieval in ILIAS 100M by between 3.7 and 10.5 mAP@1k. In Table 1b, we compare this task adaptation approach (TA) with a version of QuARI that was also trained on a 1M image sample of the Universal Embeddings dataset. QuARI shows significant improvements, between 13.1 and 19.7 mAP@1k, on top of task adaptation improvements.

INQUIRE. Table 3 shows retrieval performance across baseline model and backbone encoder sizes, along with corresponding QuARI models built upon their frozen features. In all cases, QuARI provides a significant performance improvement over general-purpose global features.

Other benchmarks. Table 2 provides results on other popular text-to-image and image-to-image retrieval benchmarks. QuARI demonstrates strong performance improvements over both the off-the-shelf and fine-tuned SigLIP2 backbone model.

	ILIAS		INQUIRE	
method	**I2I @ 100M**	**T2I @ 5M**	**mAP@50**	**nDCG@50**
SigLIP2	20.8	24.7	37.2	52.3
Fine-tuned SigLIP2	21.1	25.1	38.9	53.1
QuARI w/o Iterative Generation	28.5	32.1	43.8	54.0
QuARI w/o Semi-Positives	30.3	35.6	45.9	53.8
QuARI w/o Noise	20.0	32.4	44.1	56.7
QuARI (query transformation only)	23.2	26.8	40.1	54.9
QuARI (database transformation only)	33.4	38.2	48.4	57.4
QuARI (query & database transformations)	**36.2**	**41.2**	**50.9**	**58.4**

Table 5: Algorithmic ablation studies on QuARI.

5.2 Re-ranking

In Table 4, we compare QuARI to other approaches for image-to-image re-ranking. We replicate the baselines used by ILIAS [16], and pre-compute 100 local feature descriptors from DINOv2 [27] for each gallery image, and 600 local feature descriptors for each query image. We evaluate these methods on the image-to-image re-ranking task over a set of top-1000 initial retrievals.

We demonstrate higher performance than baseline methods, most of which require computing and storing many local feature descriptors per query, while we only use global features that would typically already be stored in a database index.

We also evaluate QuARI on text-to-image re-ranking in INQUIRE against large VLMs. Table 4b shows that QuARI outperforms all open-source VLMs on text-to-image re-ranking on this task while using precomputed global features and much lower computational overhead (explored further in Section 5.5). QuARI is also competitive with closed-source VLMs GPT-4V and GPT-4o [25, 26].

5.3 Ablation Studies

We present ablations on both algorithmic design choices and hyperparameter selection.

5.3.1 Algorithmic Ablations

We use the SigLIP2 backbone as the frozen feature extractor and evaluate both image-to-image and text-to-image tasks on ILIAS and INQUIRE. First, we compare the performance of the pretrained SigLIP2 model with models that are fine-tuned on the same datasets on which we train QuARI. This shows that fine-tuning alone, even on data from relevant domains, only provides a small improvement. We also present ablations demonstrating that it is not sufficient to learn a model learning only query transformations, and that it is relatively much more important to learn query-specific transformations of the database embeddings.

We then consider the performance of QuARI when we remove different components of the algorithm. First, we remove the iterative generation process and use a one-step generation process instead, resulting in a 6.8 decrease in mAP@1k on image-to-image retrieval @ 100M, and an 8.5 decrease in mAP@1k in text-to-image retrieval @ 5M on ILIAS, and a 6.9 decrease in mAP@50 on INQUIRE. Next, we remove the consideration of semi-positive samples used during training. This results in a degradation of 5.0 mAP@1k on both retrieval tasks on ILIAS and 4.8 mAP@50 on INQUIRE. Finally, we consider the case where noise is not added to the query representation during training to bridge the modality gap. This has the most significant impact on performance, with a drop of 15.3 image-to-image mAP@1k on ILIAS 100M, 8.2 text-to-image mAP@1k on ILIAS 5M, and 6.6 mAP@50 on INQUIRE. Notably, on the image-to-image task, QuARI without adding noise to the query representation during training does worse than the baseline SigLIP2 model, indicating that with only text-image data, this method could be prone to over-fitting without the additional regularization.

rank	I2I @ 100M	T2I @ 5M
16	23.5	40.6
32	30.2	32.8
64	**35.3**	**40.6**
128*	33.6	38.9
256*	29.4	32.6

(a)

semi-positives	I2I @ 100M	T2I @ 5M
0	30.3	35.6
1	33.1	37.2
2	**35.3**	**40.6**
3*	35.1	39.8

(b)

Table 6: Hyperparameter ablation studies on the rank of the QuARI projection (left) and the number of semi-positive examples considered during training (right). For some experiments, changed hyperparameters necessitated a decrease in batch size—we note those cases with an asterisk. We report all results using a SigLIP2 ViT-L backbone.

5.3.2 Hyperparameter Ablations

Table 6 provides ablation studies on the key parameters r, the rank of the learned transformation, and the number of semi-positive samples used during training. We show that both a increasing rank and increasing the number of semi-positive samples is beneficial until the batch size has to be reduced to train on a single NVIDIA H100 GPU.

5.4 Embedding Visualizations

In Figure 3, we explore feature transformations for two queries—an image query from ILIAS and a text query from INQUIRE. To visualize the original embedding space, the middle panel of each row shows the SigLIP2 embeddings for a collection composed of the two queries, their corresponding ground-truth images, and 5,000 distractor images (2,500 sampled at random from each of the ILIAS and INQUIRE datasets). In each row, the query embedding is highlighted in blue and the ground-truth responses in red. The right panel shows the t-SNE embedding of the same set after query-specific adaptation; here, the ground-truth responses are mapped much closer to their query embedding.

5.5 Computational Efficiency

One of the primary strengths of QuARI is that it adapts features from a precomputed database of off-the-shelf features. Figure 4 shows the time to run a fixed-size query versus the performance of the method. QuARI not only achieves state-of-the-art performance, but is also very lightweight. On the ILIAS image-to-image re-ranking task, QuARI achieves around 3% improvement over the highest accuracy re-ranking approach in over two orders of magnitude less time, and over 6% better than approaches that have similar speeds *without* the use of auxiliary local feature descriptors. On the INQUIRE text-to-image re-ranking task, QuARI is almost 10% better than the best-performing vision-language model, and is orders of magnitude faster.

6 Limitations

An inherent limitation of our method is that it applies a linear transformation to the retrieved results. While this design choice enables fast inference, it restricts the expressiveness of the adaptation. If the original representation space lacks relevant features for the retrieval, a linear transformation of it will be insufficient to improve results. Additionally, like most re-ranking approaches, our method is constrained by the set of top-k results initially retrieved: it cannot recover relevant items that are excluded from this initial set. However, because QuARI is extremely efficient, this top-k restriction is not as limiting in practice: we can afford to deploy QuARI over a very large initial set. Exploring non-linear transformation strategies to overcome representational limitations while maintaining computational tractability is a promising avenue for future research.

7 Conclusions

In this work, we introduced a query-specific retrieval framework, QuARI, that significantly outperforms strong baselines, including large vision-language models and models with learned domain-

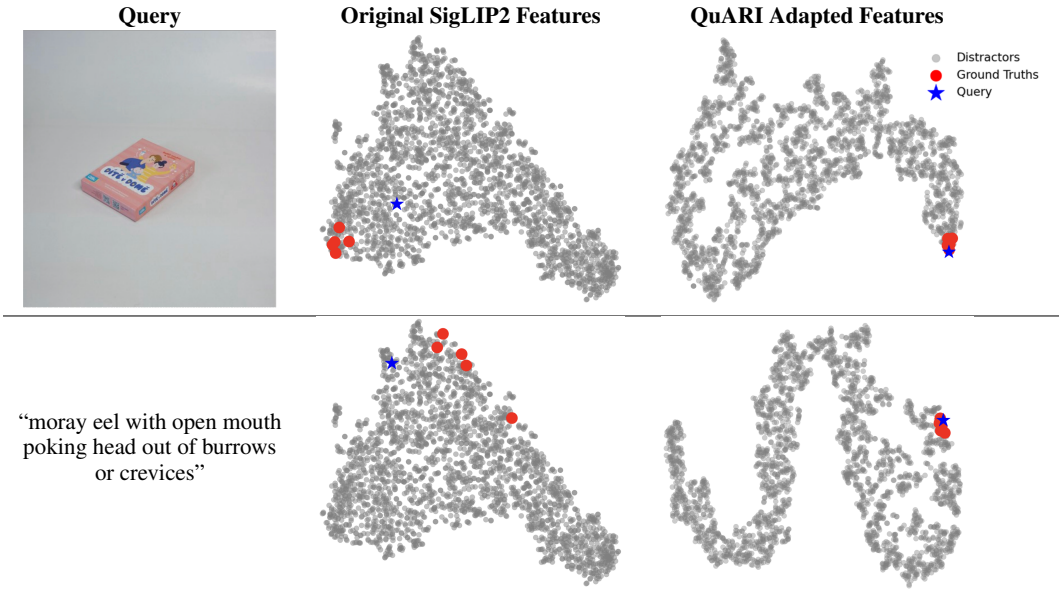

Figure 3: t-SNE visualizations comparing original features and QuARI features.

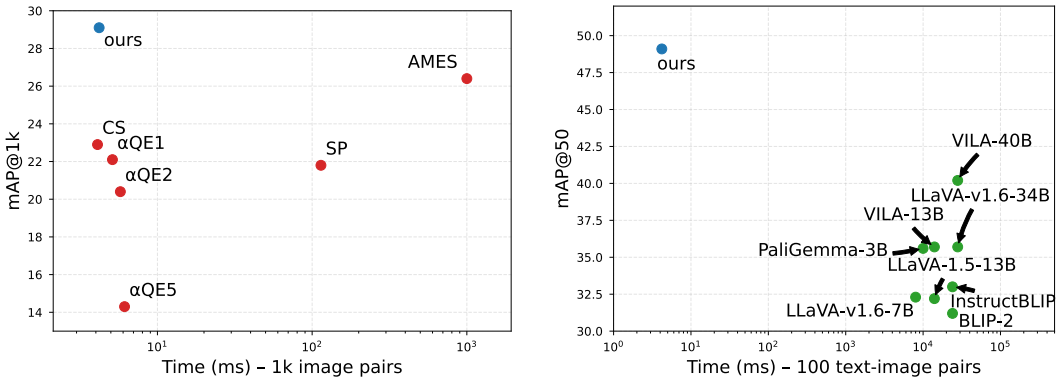

Figure 4: Comparison of re-ranking performance and inference cost for image-to-image retrieval on the ILIAS dataset (left) and text-to-image retrieval on the INQUIRE dataset (right).

specific adaptations, on challenging retrieval benchmarks. By learning to predict transformations tailored to each query, our method significantly improves image-to-image and text-to-image retrieval performance without incurring substantial computational overhead. Unlike traditional re-ranking pipelines that rely on expensive local descriptors or multi-stage processing, our approach operates directly on global embeddings and scales efficiently for searching large databases of images. Our results demonstrate that retrieval performance can be meaningfully improved not by making the underlying encoders larger or more specialized, but by learning lightweight, query-conditioned adaptations of their features.

8 Acknowledgments

This material is based upon work supported in whole or in part with funding from the National Science Foundation (DGE-2125677, IIS-2441774) and the Department of Defense (DoD). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DOD/NSF and/or any agency or entity of the United States Government.

References

[1] Victor Akinwande, Mohammad Sadegh Norouzzadeh, Devin Willmott, Anna Bair, Madan Ravi Ganesh, and J. Zico Kolter. Hyperclip: Adapting vision-language models with hypernetworks, 2024. URL https://arxiv.org/abs/2412.16777.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] Lucas Beyer, Andreas Steiner, André S. Pinto, Alexander Kolesnikov, Xiao Wang, et al. Paligemma: A versatile 3b vision–language model for transfer. Technical report, Google Research, 2024. URL https://ai.google.dev/gemma/docs/paligemma.

[5] Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrović. Improving fine-grained understanding in image-text pre-training, 2024. URL https://arxiv.org/abs/2401.09865.

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[7] Yinbo Chen and Xiaolong Wang. Transformers as meta-learners for implicit neural representations. In *European Conference on Computer Vision*, 2022.

[8] Ondřej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. URL https://api.semanticscholar.org/CorpusID:570516.

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony M. H. Tiong, Junqi Zhao, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS*, pages 17246–17262, 2023.

[10] Vinh Dang, Thanh-Son Nguyen, Minh-Triet Tran, and Duc-Tien Dang-Nguyen. Detecting misinformation in photos utilizing reverse image search. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1321–1323, 2024.

[11] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[12] David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[13] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. https://doi.org/10.5281/zenodo.5143773, July 2021.

[14] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR*, 2020.

[15] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-centric image–language pretraining for open-vocabulary detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[16] Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Šuma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiří Matas, Ondřej Chum, and Giorgos Tolias. Ilias: Instance-level image retrieval at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[17] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pages 10439–10460, 2023.

[18] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23178–23189, 2024.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16112–16128, 2023.

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22926–22936, 2024.

[22] Haotian Liu, Chunyuan Li, and the LLaVA Team. Llava-1.6: Release notes and model card. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, 2025. Accessed 14 May 2025.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[24] Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5474–5490, 2023.

[25] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023. URL `https://arxiv.org/abs/2303.08774`. Describes the multimodal GPT-4V variant.

[26] OpenAI. Introducing gpt-4o. `https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/`, 2025. Official launch post.

[27] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

[28] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[29] Pinecone. Rerankers and two-stage retrieval. `https://www.pinecone.io/learn/series/rag/rerankers/`, 2023.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, and et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[31] Elan Rosenfeld, Preetum Nakkiran, Hadi Pouransari, Oncel Tuzel, and Fartash Faghri. Ape: Aligning pretrained encoders to quickly learn aligned multimodal representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL `https://arxiv.org/abs/2210.03927`.

[32] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, and et al. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[33] Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Bingyi Cao. Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11036–11046, 2023.

[34] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[35] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioningwith reading comprehension. 2020.

[36] Abby Stylianou, Hong Xuan, Maya Shende, Jonathan Brandt, Richard Souvenir, and Robert Pless. Hotels-50k: A global hotel recognition dataset. In *The AAAI Conference on Artificial Intelligence (AAAI)*, January 2019.

[37] Pavel Suma, Giorgos Kordopatis-Zilos, Ahmet Iscen, and Giorgos Tolias. Ames: Asymmetric and memory-efficient similarity estimation for instance-level retrieval. In *European Conference on Computer Vision (ECCV)*, 2024.

[38] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[39] Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

[40] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016. ISSN 0001-0782. doi: 10.1145/2812802. URL `https://doi.org/10.1145/2812802`.

[41] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL `https://arxiv.org/abs/2502.14786`.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[43] Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. *Advances in Neural Information Processing Systems*, 37:126500–126514, 2024.

[44] Pengxiang Wu, Siman Wang, Kevin Dela Rosa, and Derek Hu. Forb: a flat object retrieval benchmark for universal image embedding. *Advances in Neural Information Processing Systems*, 36:25448–25460, 2023.

[45] Tong Wu, Liang Pan, Junzhe Zhang, Tai WANG, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. In *In Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[46] Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37:102101–102120, 2024.

[47] Lewei Yao, Runhui Huang, Lu Hou, and et al. Filip: Fine-grained interactive language–image pre-training. In *ICLR*, 2022.

[48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[49] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Bingyi Cao, Mário Lipovský, Pelin Dogan-Schönberger, Grzegorz Makosa, Boris Bluntschli, Mojtaba Seyedhosseini, Ondřej Chum, and André Araujo. Towards universal image embeddings: A large-scale dataset and challenge for generic image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11290–11301, 2023.

[50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, October 2023.

[51] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the main contributions of the paper: the introduction of QuARI, a query-specific hypernetwork for retrieval; its lightweight computational design; and its strong empirical performance on large-scale datasets like ILIAS and INQUIRE. These contributions are all supported by experimental results presented in Sections 4 and 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 ("Limitations") explicitly discusses the primary limitations of the method, including the reliance on linear transformations which can restrict expressive power and the dependence on an initial shortlist from the base retriever. These reflections are honest and appropriately scoped given the claims made.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper includes comprehensive implementation details in Section 3.3 ("Training"), including model architecture, loss functions, optimization parameters, datasets used, and evaluation metrics. Additional configuration information such as token initialization, noise injection, and ablation variants are described in detail. Additionally, code to reproduce all experimental results will be released with the camera ready submission.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: All code and datasets will be released upon acceptance of this paper, but is not included with this submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: Section 3.3 and Section 4 provide details about training and evaluation setups, including hyperparameters (optimizer, learning rate, weight decay), backbone models used, dataset splits, and evaluation metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Experiments conducted are too computationally expensive to run multiple trials. Following conventions for large-scale datasets, we report experimental results based on a single trial.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided the most important details in the main paper, with a more detailed section on hyperparameters in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We include a section on Broader Impacts in the Supplemental Materials.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The models and datasets used in this work are not high-risk for misuse and involve no sensitive or restricted assets. Therefore, no special safeguards were necessary or implemented.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and models used (e.g., MS COCO, Conceptual Captions, BioTrove, CLIP, SigLIP) are credited and cited with references in the bibliography. Their licenses (e.g., open access or research use) are respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We primarily leverage existing assets (datasets and models) that are publicly available. The methods for generating the synthetic BioTrove data are included in the Supplemental Materials. We will release trained model weights upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or human subject studies.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve any crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were only used for minor editing purposes and were not involved in the core methodology or formulating ideas.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Implementation Details

QuARI's transformer backbone is randomly initialized with 4-8 transformer layers depending on the size of the backbone encoder. The query encoder and both the query and column decoders are two-layer MLPs with `GeLU` activation functions and layer normalization [1]. We train with a batch size of 320 and a contrastive temperature of 0.07. All backbone model embeddings are precomputed before training.

B Data Generation Prompt

For all datasets other than BioTrove [6], we use the provided natural language annotations as the text label. However, BioTrove does not provide natural language annotations outside of taxonomic and common-name identities. Therefore, we provide the species annotation along with the image to `Qwen2.5-VL-7B-Instruct` [5] with the following instruction:

"For the image shown, write one plain, human-sounding sentence that someone might type into an image search system to find this exact picture of a {species_name}. Mention the main objects, their key attributes, and any distinctive action or setting. Keep it brief and objective, avoiding flowery descriptors unless they are essential to identify the scene. Output only this sentence."

We collect these annotations on 500K images sampled from BioTrove to augment our training dataset with natural language descriptions of biodiversity-domain imagery.

C Broader Impacts

Improving retrieval systems to be both more accurate and more computationally efficient has broad positive implications, especially in domains where real-time or large-scale search is critical – such as recognizing where victims of human trafficking are photographed [4], monitoring biodiversity using camera trap images in ecological surveys [2], or identifying the spread of disinformation through manipulated visual media [3]. QuARI enables high-quality retrieval even with limited resources, making advanced search capabilities more accessible in a wider range of applications. We do not foresee unique negative societal impacts associated with QuARI beyond those that already exist with general-purpose image retrieval systems. Nevertheless, the broader implications of visual search technologies—including potential misuse in surveillance or disinformation—remain important areas for ongoing community oversight and ethical consideration.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL `https://arxiv.org/abs/1607.06450`.

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[3] Vinh Dang, Thanh-Son Nguyen, Minh-Triet Tran, and Duc-Tien Dang-Nguyen. Detecting misinformation in photos utilizing reverse image search. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1321–1323, 2024.

[4] Abby Stylianou, Hong Xuan, Maya Shende, Jonathan Brandt, Richard Souvenir, and Robert Pless. Hotels-50k: A global hotel recognition dataset. In *The AAAI Conference on Artificial Intelligence (AAAI)*, January 2019.

[5] Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

[6] Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37:102101–102120, 2024.