

# MOCHA: Multi-modal Objects-aware Cross-arcHitecture Alignment

Elena Camuffo<sup>1,2</sup>, Francesco Barbato<sup>1,2</sup>, Mete Ozay<sup>1</sup>, Simone Milani<sup>2</sup>,  
and Umberto Michieli<sup>1</sup>

<sup>1</sup> Samsung R&D Institute UK, United Kingdom

<sup>2</sup> University of Padova, Italy

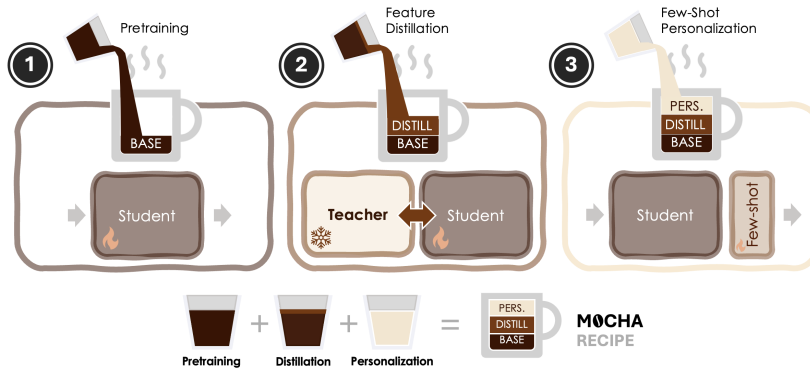
<https://github.com/SamsungLabs/MOCHA>

**Abstract.** Personalized object detection aims to adapt a general-purpose detector to recognize user-specific instances from only a few examples. Lightweight models often struggle in this setting due to their weak semantic priors, while large vision-language models (VLMs) offer strong object-level understanding but are too computationally demanding for real-time or on-device applications. We introduce MOCHA (Multi-modal Objects-aware Cross-arcHitecture Alignment), a distillation framework that transfers multimodal region-level knowledge from a frozen VLM teacher into a lightweight vision-only detector. MOCHA extracts fused visual and textual teacher’s embeddings and uses them to guide student training through a dual-objective loss that enforces accurate local alignment and global relational consistency across regions. This process enables efficient transfer of semantics without the need for teacher modifications or textual input at inference. MOCHA consistently outperforms prior baselines across four personalized detection benchmarks under strict few-shot regimes, yielding a +10.1 average improvement, with minimal inference cost.

**Keywords:** Knowledge Distillation · Object Detection · Personalization

## 1 Introduction

Recent advances in vision-language models (VLMs) such as CLIP [39], Flamingo [1], and LLaVa [24] have demonstrated remarkable zero-shot and open-vocabulary capabilities, owing to their rich representations and scale. These models, however, are typically large and computationally intensive, limiting their applicability in real-time or resource-constrained scenarios such as mobile devices (*e.g.*, smartphones or robots). In contrast, lightweight architectures such as YOLO [16, 17] offer fast and memory-efficient object detection, at the cost of degraded performance in low-data regimes and vulnerability to neural collapse [19, 36], where features become indistinguishably aligned and lose their discriminative power. In this work, we aim to bridge this gap by proposing MOCHA, a knowledge distillation approach that transfers object-centric multimodal embeddings from a vision-language teacher into a compact vision-only student detector. MOCHA



**Fig. 1: MOCHA recipe.** (1) Pretraining student model. (2) Knowledge distillation on rich joint visual and textual features from a frozen teacher. (3) Few-shot personalization with frozen student and prototypical learner.

builds on the observation that semantically related concepts tend to exhibit similar structures in the embedding space across modalities under a well-generalized backbone [4, 8, 13]. By aligning and regularizing the student’s feature space, it achieves improved generalization in few-shot detection settings. Our approach consists of three stages (Fig. 1): (1) pretraining a student detector, (2) distilling rich joint visual-textual features from a frozen teacher, and (3) performing few-shot personalization with a frozen student and a prototype-based classifier [41].

Our primary contribution lies in the distillation stage, where we introduce: (i) extraction of joint visual-language features from a frozen teacher, using both image regions and class labels to provide strong object-level supervision; (ii) a translation module that maps the student’s visual features into the teacher’s multimodal space; and (iii) a distillation objective that combines local feature alignment with a relational constraint, helping the student match the teacher’s semantics and get relationships between object regions.

Unlike prior cross-modal or cross-architecture distillation approaches (*e.g.*, AuXFT [3], OFA [11], GLIP [21]), MOCHA targets object-level few-shot personalized detection by aligning lightweight detector features with compact multimodal embeddings from a frozen VLM teacher. In particular, unlike AuXFT’s dense visual feature transfer between related visual architectures, MOCHA bridges both a modality gap (vision-only vs. vision-language) and a structural mismatch (multi-scale detector features vs. object-centric VLM representations) through multimodal fusion and relational alignment.

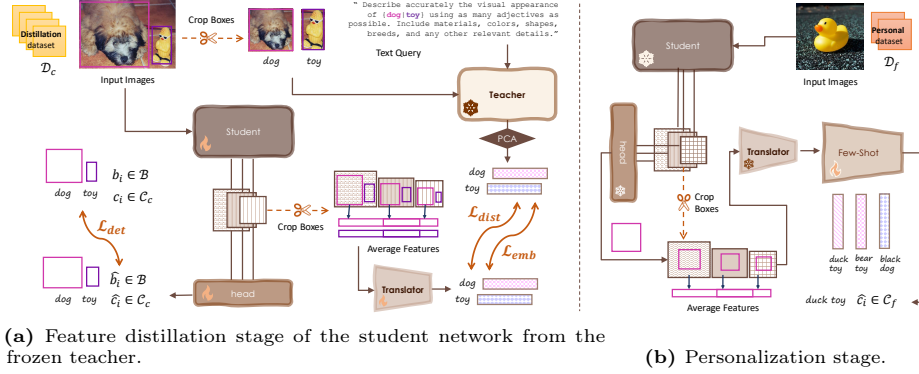
We validate MOCHA on personal object detection datasets, where the goal is to adapt a general detector to user-specific personal classes (*e.g.*, distinguishing a generic *dog* from a particular *user’s dog*). MOCHA achieves significant improvements over state-of-the-art methods while retaining minimal inference cost, making it practical for resource-constrained deployment.

## 2 Related Work

*Cross-Architecture Knowledge Distillation.* Traditional knowledge distillation methods [9, 12, 21, 22, 29, 31, 34, 38, 40, 43, 47] typically assume that teacher and student networks share similar architectures. The adoption of large transformer-based models [5, 6, 27, 44] has raised new challenges in transferring knowledge into efficient CNNs, which offer better trade-offs for deployment. Recent works [11, 26] address this by projecting multi-scale features from teacher and student into a shared space for supervision. These methods primarily aim to improve performance on the same task shared by teacher and student. In contrast, our approach aligns more closely with AuXFT [3], which leverages teacher features to guide a student on a related auxiliary task without degrading performance on the original one. Specifically, AuXFT addresses few-shot instance-level personalized object detection by injecting high-level teacher features into a lightweight detector to support prototype-based classification. To the best of our knowledge, AuXFT is the only prior work exploring a similar setup. Our method differs in three ways: (i) we leverage multimodal supervision from both visual and textual signals, unlike purely visual clues used in prior art; (ii) we distill compact region-level embeddings rather than dense intermediate activations, improving efficiency; and (iii) we explicitly regularize the embedding space to encourage geometric and relational alignment between individual instances.

*Cross-Modal Adaptation.* Several cross-modal learning strategies have been proposed to exploit complementary information across modalities [2, 46] or to enforce consistency between predictions from different modalities [14, 15]. In our work, the VLM is used only offline as a teacher: after distillation, MOCHA performs personalization and detection with a compact vision-only detector, without text prompts or VLM inference at test time.

*Personalized Scene Understanding.* Personalization was first explored in Natural Language Processing (NLP) [5, 24, 44]. Later, it was extended to computer vision, with early applications in personalized semantic segmentation [49] and object detection [33]. The advent of foundation models such as CLIP [39] and SAM [18] has shifted personalization strategies toward prompt-based control of generalist architectures [32]. Building on this trend, several methods leverage strong visual or multimodal representations to guide instance-aware predictions through textual or visual prompts, including open-vocabulary detectors like ViLD [10] and prompt-driven approaches such as SwissDINO [37], PerSAM [48], Matcher [25], and SegGPT [45]. While these approaches allow for zero-shot recognition, they are computationally demanding and do not directly address the few-shot instance-level recognition scenario considered here. Our work instead distills rich semantic features into a compact detector during server-side training, avoiding reliance on textual prompts at test time (that would add significant computational costs) or online adaptation, and enabling efficient few-shot personalization for resource-constrained deployment.



**Fig. 2: MOCHA system.** (a) *Feature distillation*: Student detector is trained on dataset  $\mathcal{D}_c$  aligning multiscale region-level features to PCA-pruned multimodal embeddings from a frozen vision-language teacher via the translation module  $t_S(\cdot)$ . (b) *Personalization*: Student backbone and Translator are frozen and used to compute semantic prototypical features from a personal dataset  $\mathcal{D}_f$ . These are then used to train a prototype-based few-shot learner  $p(\cdot)$  for user-specific object classification.

### 3 Methodology

MOCHA (Multi-modal Objects-aware Cross-arcHitecture Alignment) distills multimodal semantics from a vision-language foundation model into a lightweight student detector by aligning region-level embeddings and regularizing the student feature space. MOCHA enhances feature separability and generalization for few-shot personalized detection. Our method is architecture-agnostic and combines multimodal supervision, feature translation, and a dual loss enforcing both local and relational alignment.

#### 3.1 Problem Setup

MOCHA consists of three stages, as described next.

*Base Pretraining.* We begin with a standard detection model  $m_S(\cdot) = l_S \circ g_S$ , where  $l_S(\cdot)$  is the detection head,  $g_S(\cdot)$  is the student backbone, and  $\circ$  is the composition operator. We train it on a pretraining detection dataset using standard objectives  $\mathcal{L}_{det}$  [16]. This stage results in a strong generic detector that identifies broad object categories but lacks personalized semantics or fine-grained embedding separation and serves as initialization for the student’s weights.

*Feature Distillation.* After initialization, we perform feature distillation over an object detection dataset  $\mathcal{D}_c$ , using a frozen large vision-language model as the teacher (Fig. 2a). Each image  $X \in \mathcal{D}_c$  in the dataset is labeled by a set of tuples  $\{(b_i, c_i)\}_{i=1}^n \subset \mathcal{B} \times \mathcal{C}_c$ , where  $c_i \in \mathcal{C}_c$  are class labels and  $b_i \in \mathcal{B} \subset \mathbb{R}^4$  are box coordinates identifying  $n$  image regions  $X_i$ . The teacher is assumed to



comprise a visual encoder  $g_T(\cdot)$ , and a large language model  $f_T(\cdot)$ , enabling the joint processing of images and text. Note that, in general, the student and teacher features may not be aligned. For this reason, we use a feature translation module  $t_S(\cdot) : \mathbb{R}^{d_s} \mapsto \mathbb{R}^{d_t}$  (where  $d_s$  and  $d_t$  are the dimensions of the student and teacher features, respectively), to allow knowledge distillation from the student’s to the teacher’s space. Importantly, teacher embeddings are precomputed and cached, avoiding VLM inference during training iterations and deployment.

*Few-Shot Personalization.* After feature distillation, the student backbone is frozen and used to extract multiscale region-level features  $\{F_j\}_j = g_S(X)$  from a personal detection dataset  $\mathcal{D}_f$ , labeled with bounding boxes  $\mathcal{B}$  and class labels  $\mathcal{C}_f$ . The features are translated into the shared embedding space via the frozen translation module (Fig. 2b) and fed to a prototype-based few-shot learner  $p(\cdot)$ , like nearest class mean [41], which is trained and tested as in [3]. More in detail, during personalization, a user provides a few labeled samples of personal instances of objects to the system (*e.g.*, 1-5 samples). Later, during inference, the system automatically overrides the coarse predictions with the personal labels. This strategy enables the student to adapt to novel visual concepts while retaining its compact architecture. The prototype-based classifier computes class-specific prototypes from distilled embeddings without teacher inference, prompt engineering, or detector fine-tuning, enabling low-overhead personalization.

### 3.2 Multimodal Supervision Extraction

Without loss of generality, in our experiments, we instantiate the teacher with a LLaVa [24] model, which includes a shared projection matrix  $W_T$ , in addition to  $g_T(\cdot)$  and  $f_T(\cdot)$ . Each cropped image region  $X_i$  is processed by the teacher’s visual encoder to obtain a visual embedding  $Z_{V,i}$ , projected into a joint text-vision representation space as  $H_{V,i} = W_T Z_{V,i}$ . Simultaneously, the class label  $c_i$  is embedded as a textual query  $Z_{Q,i}$  and projected via  $H_{Q,i} = W_T Z_{Q,i}$ . The pair  $(H_{V,i}, H_{Q,i})$  is passed to the frozen language model  $f_T$ , which computes a sequence of tokens describing the interaction between the region and the class. We average the output along the temporal dimension to obtain a single semantic embedding  $h_i \in \mathbb{R}^{d_h}$  representing the fused visual-linguistic description of the region  $X_i$ .

*Dimensionality Reduction.* While the teacher’s multimodal feature  $h_i$  captures rich semantics, it does not necessarily retain appearance-level cues. To combine both semantic and visual signals, we concatenate the multimodal representation with the  $d_z$ -dimensional class token  $z_{V,i} \in Z_{V,i} \subset \mathbb{R}^{d_z}$  of its corresponding visual embedding:

$$u_i = \text{concat}(\gamma z_{V,i}, h_i) \in \mathbb{R}^d, \quad d = d_z + d_h. \quad (1)$$

with  $\gamma = \|h_i\|$  as a rescaling factor that normalizes visual features to textual ones. In practice,  $h_i$  is obtained by mean-pooling the hidden states of the final decoder layer of the VLM, while  $z_{V,i}$  corresponds to the  $\ell_2$ -normalized class token

of the visual encoder. The scaling factor  $\gamma$  balances the norms of visual and textual features, preventing the higher-magnitude modality from dominating the fused representation. This combined vector represents both the raw visual content of the object and its semantic interpretation conditioned on the label. However, the resulting representation is high-dimensional and could include redundant information. To reduce dimensionality and improve efficiency, we apply Principal Component Analysis (PCA) to  $u_i$ , offline on the same dataset used for distillation, and we take the first  $d_t$  elements:

$$\hat{u}'_i = \text{PCA}(u_i) \in \mathbb{R}^{d_t}. \quad (2)$$

Finally, to equalize the contribution of each channel in the PCA representation, we estimate the standard deviation of each channel activation  $\sigma_c$  in the distillation dataset, and use it to rescale the channels as:

$$u'_i[c] = \hat{u}'_i[c]/\sigma_c \quad \text{for } c=1, \dots, d. \quad (3)$$

Remarkably, the deviations closely follow a hyperbolic shape, allowing us to estimate their value based only on the channel index (detailed discussion in the Appendix).

### 3.3 Student Detector and Feature Aggregation

To perform object-level supervision, we apply the same cropping strategy used for the teacher: for each ground truth bounding box  $b_i$ , we first resize the feature maps to a common size to account for resolution differences, ensuring consistent spatial alignment. Then, we extract the corresponding region-aligned features  $F_{A,j,i}$  from each resolution level  $j$ , and apply spatial average pooling to obtain fixed-size descriptors  $f_{A,j,i} \in \mathbb{R}^{d_s}$ .

The final aggregated feature for region  $i$  is formed by concatenating the pooled representations at each level:

$$f_{A,i} = \text{concat}_j(f_{A,j,i}) \in \mathbb{R}^{d_s}. \quad (4)$$

We denote with  $F_A$  the dense equivalent of these features, that is, the concatenation without region pooling.

*Feature Translation Module.* The translation module  $t_S$  is modelled after a transformer encoder block and consists of a channel-wise multi-head self-attention block followed by a lightweight multi-layer perceptron (MLP). The attention mechanism enhances expressiveness by modelling inter-channel dependencies, while the MLP adapts the representation to the required dimensionality:

$$f'_{A,i} = t_S(f_{A,i}), \quad (5)$$

where  $f'_{A,i} \in \mathbb{R}^{d_t}$  is the translated student feature with the same dimension as the PCA-compressed teacher target  $u'_i$ .

Dimensionality matching alone is insufficient due to the heterogeneity between detector and VLM latent spaces. The translation module therefore performs a semantic mapping rather than a simple projection, learning to bridge the structural gap between region-level CNN features and multimodal transformer representations. In our implementation (Fig. 2a), the translation module is trained jointly with the student detector, allowing the feature alignment to evolve progressively during distillation.

### 3.4 Training Objectives and Strategy

To align the translated student features with the teacher targets, we define two complementary objectives: a pointwise distillation loss and a relational embedding loss.

*Distillation Loss.* The primary objective enforces direct alignment between translated student features  $f'_{A,i}$  and their corresponding teacher vector  $u'_i$ . We define the distillation loss as the average of  $\ell_1$  and  $\ell_2$  distances, as in [3]:

$$\mathcal{L}_{\text{dist}} = \frac{1}{n} \sum_{i=1}^n (\|f'_{A,i} - u'_i\|_1 + \|f'_{A,i} - u'_i\|_2). \quad (6)$$

This formulation captures both robustness to outliers (via the  $\ell_1$  term) and fine-grained magnitude alignment (via the  $\ell_2$  term), promoting stable and discriminative alignment.

*Embedding Loss.* Beyond individual alignment, we encourage the student to preserve the global geometry of the teacher’s embedding space. To this end, we define a relational embedding loss based on pairwise Euclidean distances between all student and teacher features. Given the translated student features  $\{f'_{A,i}\}_i$  and the corresponding teacher targets  $\{u'_i\}_i$ , we compute pairwise distance matrices  $D_{ff}$ ,  $D_{uu}$  using normalized Euclidean distances. We remove self-distances by discarding diagonal entries and convert into probability distributions using softmax:

$$P_{ff} = \text{softmax}(-D_{ff}), \quad P_{uu} = \text{softmax}(-D_{uu}). \quad (7)$$

The embedding loss is then defined as the cross-entropy between the two distributions:

$$\mathcal{L}_{\text{emb}} = -\frac{1}{n} \sum_{i=1}^n \sum_{j \neq i} P_{uu}[i, j] \log P_{ff}[i, j]. \quad (8)$$

This term ensures that the student preserves not only semantic content but also the relative structure of the embedding space, reflecting inter-class and intra-class relationships encoded by the teacher.

Although the relational loss involves pairwise interactions and is theoretically  $\mathcal{O}(N^2)$ , in practice it is computed over a small set of ground-truth object regions per image, making the overhead negligible compared to standard convolutional operations (on OpenImages, we observe on average  $N \approx 8$  regions per image).

*Final Objective.* The final objective combines the detection loss with the distillation and embedding terms:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{emb}}\mathcal{L}_{\text{emb}}, \quad (9)$$

where the  $\lambda$  coefficients regulate the contribution of each auxiliary term. This combined objective ensures that the student learns both task-relevant representations and the semantic structure transferred from the teacher.

## 4 Experiments

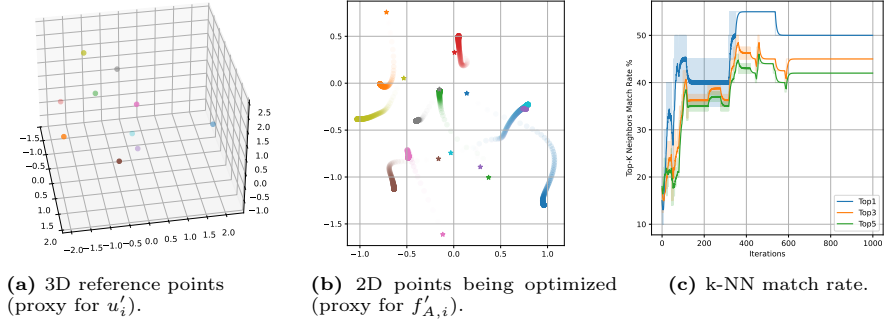
We evaluate MOCHA in the context of few-shot personalized object detection, assessing its ability to transfer multimodal knowledge from a large VLM teacher into lightweight student detectors. Our experiments cover four personal datasets and analyze performance from three complementary perspectives: general behavior and design insights, comparison with state-of-the-art methods, and detailed ablation studies. Results are reported after introducing the experimental setup and presenting preliminary analyses.

### 4.1 Experimental Setup

*Datasets.* For pretraining, we adopt COCO [23] and OpenImagesV7 [20]. Also, we consider starting from AuXFT pretrained weights (which in turn start from COCO and fine-tune on OpenImages). Distillation is performed on OpenImages. Considered personalization datasets are: PerSeg [48], POD [3], CORE50 [28], and iCubWorld [7].

*Models.* Our approach uses YOLOv8n ( $\approx 3.2\text{M}$  params) as the student model and LLaVa-1.5-7B ( $\approx 7.2\text{B}$  params) as the teacher, which includes CLIP with ViT-B/32 as the visual encoder and LLaMA 7B as the text encoder. For comparison, we also include: (i) the text encoder from LLaVa, *i.e.*, LLaMA 7B ( $\approx 7\text{B}$  params) (ii) the visual encoder from LLaVa, *i.e.*, CLIP ( $\approx 0.2\text{B}$  params), and (iii) DINOv2 [35] ( $\approx 22\text{M}$  params). Moreover, we compare against other distillation-based approaches: DRKD [34], OFA [11], knowledge distillation at the output space via KL divergence (KL) [42] and its combination with embedding space distillation via MSE (MSE + KL) [40], open-vocabulary detectors (VILD) [10], vision-language grounding models (GLIP) [21], similarity-based feature distillation SKDF [31], and the most relevant personalization method to date (AuXFT). Finally, we perform ablations on the student model including YOLOv11n [17] ( $\approx 2.6\text{M}$  params) and RT-DETR-l [30] ( $\approx 45\text{M}$  params).

*Metrics.* For consistency and fair comparison, we follow the same implementation details as in AuXFT, except when distilling from the AuXFT-pretrained model. In that case, the number of distillation epochs is reduced from 50 to 20, as the convergence time decreases. We report standard end-to-end mAP<sup>50-95</sup> for



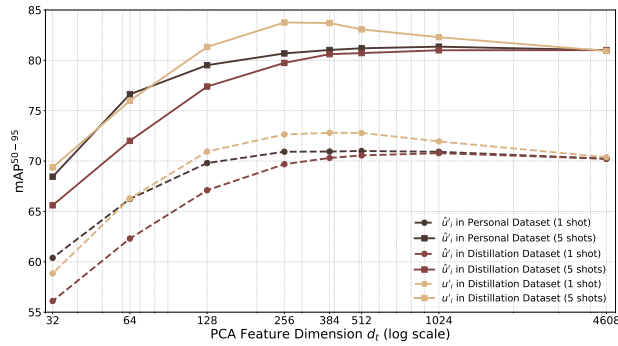
**Fig. 3: Effect of  $\mathcal{L}_{\text{emb}}$  on a set of ten 2D points optimized with respect to 3D ones.** (a) 3D reference points, proxy for the teacher embeddings  $u'_i$ . (b) Evolution of the 2D points (proxy for student embeddings  $f'_{A,i}$ ) updated via  $\mathcal{L}_{\text{emb}}$  from the 3D reference points  $u'_i$ .  $\star$  marks the original location, timesteps increase with color saturation. (c) Percent rate of 2D top- $k$  nearest neighbors (k-NN) that match those of the reference 3D distribution.

PerSeg and POD, and retrieval accuracy for CORE50 and iCubWorld following AuXFT [3], where a prediction is considered correct if the target object is retrieved among the top-ranked detector proposals. Few-shot personalized detection is evaluated under 1- and 5-shot regimes.

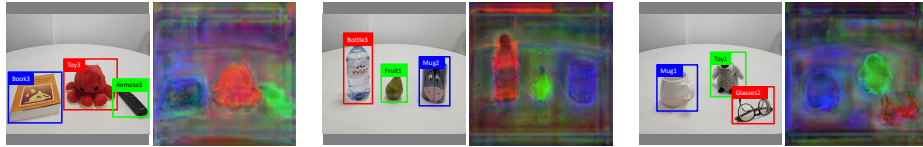
## 4.2 Analyses

*Embedding Loss.* Fig. 3 illustrates the impact of our relational embedding loss  $\mathcal{L}_{\text{emb}}$  on preserving the geometric structure of the teacher features onto student ones. We consider a toy setup, where we optimize the coordinates of a set of 2D points (proxies for student embeddings  $f'_{A,i}$ ) to match the pairwise distance distribution of a fixed 3D reference configuration (proxy for teacher embeddings  $u'_i$ ). Fig. 3a shows the 3D reference points sampled uniformly in space. Fig. 3b visualizes the optimization trajectory in the 2D space: each path starts from an initial position (marked by  $\star$ ) and progressively adapts under the guidance of  $\mathcal{L}_{\text{emb}}$ , preserving local neighborhood relationships while globally rearranging to approximate the original distance geometry (color saturation indicates temporal evolution). Fig. 3c quantitatively tracks the alignment between neighborhood structures. We report the percentage rate of 2D top- $k$  nearest neighbors that match those of the reference 3D distribution. We observe a steady improvement across all  $k$  with a rapid convergence. This analysis confirms the role of relational supervision in regularizing the student’s embedding space beyond pointwise alignment, ultimately promoting feature diversity and semantic separability.

*Impact of Teacher Feature Dimensionality.* Fig. 4 evaluates the role of teacher supervision dimensionality ( $d_t$ ) by comparing features: (i) of the personal datasets  $\mathcal{D}_f$  (as an upper bound), (ii) of the distillation dataset  $\mathcal{D}_c$  unnormalized and (iii) PCA normalized (additional details and validation in Appendix). Overall,



**Fig. 4:  $\text{mAP}^{50-95}$  at different PCA dimensionality.** Average score across all evaluation datasets varying feature dimension  $d_t$ .



**Fig. 5: Feature similarity between the  $F_A$  embeddings and the teacher target  $u_i'$  encoded in the R/G/B channels, one for each object in input scenes from POD.**

performance increases steadily up to  $d_t = 512$ , which strikes a good balance between compactness and expressivity. After that, it plateaus or slightly degrades up to the maximum dimension  $d = 4608$ . This indicates that most of the relevant semantics are preserved in a compact subspace, and that further dimensions may introduce redundancy or noise. We further analyze modality contributions to the retained PCA components in the Appendix. Normalized features  $u_i'$  consistently outperform their unnormalized counterparts  $\hat{u}_i'$  (especially in the 1-shot setting), obtaining better performance even if compared to personal datasets' features.

*Feature Similarity.* Fig. 5 shows the similarity between the teacher vision-language descriptors  $u_i'$  of the three objects in each scene and the features output by the student,  $F_A$ . In each bottom image, the three similarity matrices are visualized in R/G/B channels, showing that MOCHA successfully maps sparse VLM object-level features (inside object boxes) into the dense features of an efficient detector.

### 4.3 Main Results

We begin our evaluation in the oracle setup, where models are assessed using ground truth bounding box coordinates (one per evaluation) during the few-shot personalization stage. This setting removes the detection head and focuses solely on the classification capability of each approach, allowing us to compare different teacher models and types of supervision embeddings (linguistic, visual, and

**Table 1: Oracle results using ground truth bounding box coordinates on personal datasets.** Top block: performance when using frozen teacher embeddings directly in the personalization stage. Bottom block: performance of student distilled with the teacher signals.

| Teacher                                          | Student | PerSeg         | POD            |                |                | CRe50          |                | iCubWorld      |             | Avg |
|--------------------------------------------------|---------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-------------|-----|
|                                                  |         | 1 SHOT         | 1 SHOT         | 5 SHOT         | 1 SHOT         | 5 SHOT         | 1 SHOT         | 5 SHOT         |             |     |
| DINO                                             | —       | 90.5 $\pm$ 2.7 | 55.1 $\pm$ 4.0 | 66.8 $\pm$ 0.0 | 34.8 $\pm$ 7.2 | 47.9 $\pm$ 9.8 | 50.6 $\pm$ 2.9 | 73.5 $\pm$ 2.2 | 59.9        |     |
| CLIP ( $z_{v,i}$ )                               | —       | 95.7 $\pm$ 1.4 | 64.5 $\pm$ 3.5 | 77.8 $\pm$ 0.0 | 58.1 $\pm$ 4.6 | 79.3 $\pm$ 4.4 | 55.5 $\pm$ 3.2 | 77.9 $\pm$ 2.5 | 72.7        |     |
| LLaVa ( $h_i$ )                                  | —       | 82.2 $\pm$ 3.3 | 59.0 $\pm$ 4.1 | 72.7 $\pm$ 0.5 | 44.1 $\pm$ 3.4 | 63.1 $\pm$ 4.2 | 46.7 $\pm$ 3.1 | 67.6 $\pm$ 2.7 | 62.2        |     |
| LLaVa ( $h_i$ ) + CLIP ( $z_{v,i}, \gamma = 1$ ) | —       | 93.9 $\pm$ 1.8 | 66.4 $\pm$ 3.8 | 80.7 $\pm$ 0.2 | 59.4 $\pm$ 4.6 | 80.4 $\pm$ 4.3 | 58.0 $\pm$ 3.2 | 80.6 $\pm$ 2.4 | 74.2        |     |
| LLaVa ( $h_i$ ) + CLIP ( $z_{v,i}$ )             | —       | 95.1 $\pm$ 1.7 | 70.0 $\pm$ 3.4 | 80.6 $\pm$ 0.0 | 57.0 $\pm$ 4.0 | 79.7 $\pm$ 3.7 | 56.4 $\pm$ 3.6 | 81.2 $\pm$ 2.5 | <b>74.3</b> |     |
| —                                                | YOLOv8n | 69.6 $\pm$ 3.4 | 35.4 $\pm$ 3.4 | 41.3 $\pm$ 0.0 | 31.5 $\pm$ 5.5 | 42.8 $\pm$ 7.8 | 33.3 $\pm$ 2.7 | 48.6 $\pm$ 2.4 | 43.2        |     |
| CLIP ( $z_{v,i}$ )                               | YOLOv8n | 75.5 $\pm$ 3.5 | 40.9 $\pm$ 3.2 | 51.7 $\pm$ 0.0 | 35.0 $\pm$ 3.9 | 47.3 $\pm$ 5.7 | 40.6 $\pm$ 2.9 | 60.0 $\pm$ 3.3 | 50.7        |     |
| LLaVa ( $h_i$ )                                  | YOLOv8n | 73.0 $\pm$ 3.6 | 37.1 $\pm$ 4.2 | 47.6 $\pm$ 0.0 | 33.9 $\pm$ 3.7 | 45.9 $\pm$ 5.4 | 37.4 $\pm$ 3.1 | 55.8 $\pm$ 2.6 | 48.0        |     |
| LLaVa ( $h_i$ ) + CLIP ( $z_{v,i}$ )             | YOLOv8n | 82.0 $\pm$ 3.4 | 45.7 $\pm$ 3.5 | 58.1 $\pm$ 0.0 | 34.5 $\pm$ 4.1 | 47.8 $\pm$ 5.9 | 42.9 $\pm$ 3.1 | 62.7 $\pm$ 2.2 | <b>54.0</b> |     |

multimodal). Tab. 1 is organized into two blocks. The top block reports results obtained by using the teacher embeddings directly during personalization, while the bottom block shows the performance of YOLO students distilled with the same supervision signals. In the first block, visual teachers show strong performance. Notably, CLIP outperforms DINO (used in AuXFT). On the other hand, LLaVa’s features ( $h_i$ ) perform worse, but combining them with CLIP features ( $z_{V,i}$ ) — *i.e.*, using  $u_i$  for supervision — yields notable gains at no added cost (as CLIP is already part of LLaVa), demonstrating the complementarity of visual and textual cues. Normalizing the visual features also leads to a slight improvement over their unnormalized version ( $\gamma = 1$ ). In the second block, we evaluate distillation using either visual supervision (CLIP,  $z_{V,i}$ ), textual supervision (LLaVa,  $h_i$ ), or their combination (LLaVa,  $u_i$ ). All models are distilled through MOCHA supervision, showing consistent gains of our approach over the non-distilled baseline (YOLO only). Still, using a single modality alone proves less effective than leveraging full multimodal supervision. Note that in Tab. 1 we evaluate using the ground truth boxes, so the retrieval accuracy corresponds to standard detection accuracy, with each image containing a single annotated instance. In Tab. 2, instead, the retrieval protocol [3] allows multiple candidate boxes per image; we denote an object as correct if *any* predicted box matches the target. As this metric accepts multiple candidates, the scores in Tab. 2 are naturally slightly higher than those in Tab. 1. These results indicate that improvements do not stem merely from using a stronger multimodal teacher, but from enforcing structured relational alignment in the embedding space.

*Comparisons.* We evaluate the complete approach, including both backbone and detection head, in the standard few-shot personalization setting (predicting bounding box coordinates). Tab. 2 compares MOCHA with recent approaches for personalized detection. To ensure fair comparisons, we re-implemented representative distillation objectives within our pipeline, modifying only the supervision loss while keeping teacher, student, crops, PCA compression, and personalization protocol fixed. We consider output-space transfer (KL, KL+MSE), cross-architecture

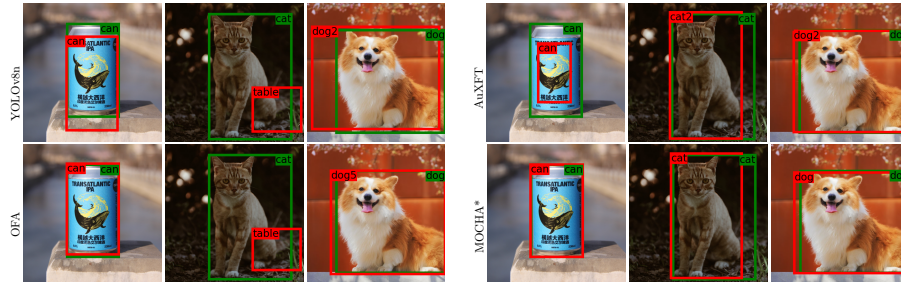
**Table 2: Few-shot personalized detection results on personal datasets.** First row: student baseline with no distillation. First block: MOCHA compared against prior state-of-the-art supervision methods. Second block: evaluation of different MOCHA variants.

| Teacher            | Supervision    | Student | PerSeg         | POD            |                | CRe50          |                | iCubWorld      |                | Avg  |
|--------------------|----------------|---------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|------|
|                    |                |         | 1 SHOT         | 1 SHOT         | 5 SHOT         | 1 SHOT         | 5 SHOT         | 1 SHOT         | 5 SHOT         |      |
| –                  | –              | YOLOv8n | 41.2 $\pm$ 2.6 | 23.6 $\pm$ 2.4 | 30.4 $\pm$ 0.0 | 57.8 $\pm$ 6.3 | 67.3 $\pm$ 7.7 | 51.2 $\pm$ 3.0 | 68.4 $\pm$ 2.4 | 48.6 |
| DINO               | AuXFT          | YOLOv8n | 48.8 $\pm$ 3.4 | 31.5 $\pm$ 2.7 | 38.8 $\pm$ 0.0 | 58.8 $\pm$ 6.4 | 69.3 $\pm$ 6.0 | 55.0 $\pm$ 3.0 | 74.5 $\pm$ 2.5 | 53.8 |
| CLIP ( $z_{V,i}$ ) | ViLD           | YOLOv8n | 44.4 $\pm$ 3.2 | 24.8 $\pm$ 2.8 | 33.3 $\pm$ 0.0 | 54.6 $\pm$ 5.2 | 64.9 $\pm$ 4.9 | 57.7 $\pm$ 3.0 | 70.4 $\pm$ 2.4 | 50.9 |
| LLaVa ( $u_i$ )    | KL div.        | YOLOv8n | 43.7 $\pm$ 2.9 | 23.4 $\pm$ 2.5 | 27.8 $\pm$ 0.0 | 54.4 $\pm$ 5.0 | 65.2 $\pm$ 5.0 | 54.7 $\pm$ 3.0 | 69.5 $\pm$ 2.3 | 49.0 |
| LLaVa ( $u_i$ )    | KL div. + MSE  | YOLOv8n | 50.1 $\pm$ 3.0 | 27.6 $\pm$ 2.4 | 30.5 $\pm$ 0.0 | 53.3 $\pm$ 4.5 | 65.7 $\pm$ 4.6 | 61.1 $\pm$ 2.9 | 78.5 $\pm$ 1.9 | 52.1 |
| LLaVa ( $u_i$ )    | OFA            | YOLOv8n | 47.2 $\pm$ 3.4 | 23.3 $\pm$ 2.6 | 28.0 $\pm$ 0.0 | 55.3 $\pm$ 4.3 | 63.6 $\pm$ 4.3 | 60.4 $\pm$ 2.5 | 72.5 $\pm$ 2.4 | 50.1 |
| LLaVa ( $u_i$ )    | GLIP           | YOLOv8n | 52.1 $\pm$ 2.8 | 27.3 $\pm$ 3.0 | 32.7 $\pm$ 0.0 | 58.7 $\pm$ 5.5 | 64.6 $\pm$ 5.7 | 61.2 $\pm$ 2.4 | 71.1 $\pm$ 2.3 | 52.2 |
| LLaVa ( $u_i$ )    | SKDF           | YOLOv8n | 47.0 $\pm$ 2.9 | 25.8 $\pm$ 2.6 | 28.4 $\pm$ 0.0 | 58.4 $\pm$ 4.8 | 67.3 $\pm$ 5.6 | 62.7 $\pm$ 2.9 | 74.1 $\pm$ 2.3 | 52.0 |
| LLaVa ( $u_i$ )    | DRKD           | YOLOv8n | 53.5 $\pm$ 2.7 | 24.3 $\pm$ 3.0 | 24.7 $\pm$ 0.0 | 57.7 $\pm$ 5.8 | 64.8 $\pm$ 6.2 | 59.7 $\pm$ 2.7 | 69.7 $\pm$ 2.4 | 50.6 |
| CLIP ( $z_{V,i}$ ) | MOCHA (visual) | YOLOv8n | 53.0 $\pm$ 2.8 | 26.7 $\pm$ 3.1 | 37.5 $\pm$ 0.0 | 57.2 $\pm$ 4.7 | 67.2 $\pm$ 4.8 | 62.6 $\pm$ 2.8 | 76.1 $\pm$ 2.3 | 54.6 |
| LLaVa ( $h_i$ )    | MOCHA (text)   | YOLOv8n | 52.3 $\pm$ 2.8 | 28.8 $\pm$ 2.6 | 33.7 $\pm$ 0.0 | 56.2 $\pm$ 4.7 | 67.1 $\pm$ 4.7 | 62.8 $\pm$ 2.8 | 76.8 $\pm$ 2.0 | 54.0 |
| LLaVa ( $u_i$ )    | MOCHA (COCO)   | YOLOv8n | 55.4 $\pm$ 3.3 | 33.9 $\pm$ 2.8 | 38.9 $\pm$ 0.0 | 55.8 $\pm$ 4.7 | 65.4 $\pm$ 4.7 | 60.6 $\pm$ 3.3 | 77.6 $\pm$ 2.6 | 56.0 |
| LLaVa ( $u_i$ )    | MOCHA (AuXFT)  | YOLOv8n | 59.1 $\pm$ 3.3 | 36.3 $\pm$ 3.3 | 45.9 $\pm$ 0.0 | 60.9 $\pm$ 4.4 | 70.6 $\pm$ 4.5 | 61.4 $\pm$ 3.1 | 77.0 $\pm$ 2.3 | 58.7 |

feature projection (OFA), similarity/feature regression objectives (GLIP, SKDF), and relational matching (DRKD) as alternative supervision paradigms in our multimodal region-level setting. Overall, KL-based and OFA-style supervision provide limited gains, while similarity- and relation-based objectives (GLIP, SKDF, DRKD) improve performance but remain below MOCHA. This suggests that neither pointwise regression nor relation matching alone is sufficient: combining local alignment with explicit relational regularization is key to effectively transferring multimodal semantics. ViLD struggles in highly personalized scenarios, despite its open-vocabulary capabilities. AuXFT achieves the strongest average performance among prior methods, but still underperforms in more challenging domains (*e.g.*, PerSeg and POD), where fine-grained semantic alignment is critical. MOCHA surpasses all competing approaches, showing consistently strong results across all datasets and settings, with a +10.1 average improvement over the YOLOv8n baseline and +4.9 over the strongest competitor, AuXFT. When initialized from COCO-pretrained weights, MOCHA already exceeds prior state-of-the-art methods; starting from AuXFT-pretrained weights yields further gains, confirming the compatibility and effectiveness of the initialization. MOCHA introduces only minimal computational overhead: on YOLOv8n on PerSeg, it adds just 3 ms/image, corresponding to a  $\sim$ 10% relative increase (see Appendix for details).

*Qualitative Results.* Fig. 6 presents a qualitative comparison on the PerSeg dataset, featuring three sample images: a *can*, a *cat*, and a *dog*. Green bounding boxes indicate ground truth annotations, while red boxes show the highest-confidence predictions. MOCHA (AuXFT) consistently demonstrates superior performance across all samples, with correct class predictions and precise bounding boxes. In contrast, competing methods exhibit occasional misclassifications and less accurate detections. These qualitative examples reflect the trends observed in the quantitative results, with MOCHA offering more reliable predictions.





**Fig. 6: Qualitative results on PerSeg.** Ground truth in green, prediction with the highest confidence in red (class names shown refer to personal class labels in PerSeg). YOLOv8n refers to baseline. \*: refers to MOCHA (AuXFT).

#### 4.4 Ablation Studies

*Generalization to Other Student Models.* Tab. 3 reports an ablation on the student model. Results show that MOCHA consistently improves few-shot detection across all student architectures, demonstrating strong generalization beyond the YOLOv8n model used in the main experiments. The largest gains are obtained with YOLOv8n, where the combination of MOCHA’s region-level supervision and YOLO’s multi-scale feature hierarchy yields substantial improvements over both the baseline and AuXFT. Upgrading the student to YOLOv11n preserves this trend, albeit with smaller margins. This suggests that newer backbones, while stronger on their own, leave less room for improvement but still benefit from MOCHA supervision. On the other hand, the transformer-based RT-DETR-l shows the smallest but still consistent gains. This is because RT-DETR-l is much larger ( $\approx 10$  times) and architecturally very different from YOLO detectors, making the translation module less effective and limiting its suitability for on-device use. Despite this, MOCHA still provides consistent gains across datasets, confirming its robustness and ability to offer complementary supervision across detector families. Nevertheless, performance variations across architectures reflect their different inductive biases and the magnitude of the semantic gap with respect to the VLM teacher. CNN-based detectors (*e.g.*, YOLO variants) benefit more from multimodal alignment, while transformer-based detectors (*e.g.*, RT-DETR), being architecturally closer to the teacher, exhibit smaller but consistent improvements.

*Components Design.* Tab. 4 reports an extensive ablation study, isolating the contributions of key design and training components of MOCHA across the PerSeg and iCubWorld datasets. We first assess the impact of the PCA dimensionality  $d_t$  used to compress teacher embeddings. As previously highlighted in Fig. 4, compressing the teacher’s embedding improves performance compared to using full-resolution features. A PCA dimension of  $d_t = 512$  represents the best balance between compactness and expressiveness. For this reason, we selected it as the embedding size. More aggressive compression (*e.g.*, 256 or 384) leads to reduced

**Table 3: Ablation studies on the student model.** Comparison between AuXFT and MOCHA (AuXFT) supervisions using different student models: YOLOv8n, YOLOv11n, RT-DETR-l.

| Supervision  | Student   | PerSeg<br>1 SHOT      | iCubWorld<br>1 SHOT   | 5 SHOT                |
|--------------|-----------|-----------------------|-----------------------|-----------------------|
| –            | YOLOv8n   | 41.2 $\pm$ 2.6        | 51.2 $\pm$ 3.0        | 68.4 $\pm$ 2.4        |
| AuXFT        | YOLOv8n   | 48.8 $\pm$ 3.4        | 55.0 $\pm$ 3.0        | 74.5 $\pm$ 2.5        |
| <b>MOCHA</b> | YOLOv8n   | <b>59.1</b> $\pm$ 3.3 | <b>61.4</b> $\pm$ 3.1 | <b>77.0</b> $\pm$ 2.3 |
| –            | YOLOv11n  | 49.3 $\pm$ 3.8        | 52.7 $\pm$ 4.1        | 72.8 $\pm$ 2.7        |
| AuXFT        | YOLOv11n  | 52.8 $\pm$ 3.8        | 53.9 $\pm$ 4.1        | 73.3 $\pm$ 2.4        |
| <b>MOCHA</b> | YOLOv11n  | <b>56.2</b> $\pm$ 3.2 | <b>57.5</b> $\pm$ 2.5 | <b>74.1</b> $\pm$ 2.4 |
| –            | RT-DETR-l | 43.5 $\pm$ 3.5        | 45.0 $\pm$ 3.8        | 63.5 $\pm$ 2.5        |
| AuXFT        | RT-DETR-l | 45.2 $\pm$ 3.4        | 46.1 $\pm$ 3.7        | 64.0 $\pm$ 2.4        |
| <b>MOCHA</b> | RT-DETR-l | <b>47.0</b> $\pm$ 3.1 | <b>47.8</b> $\pm$ 3.2 | <b>64.4</b> $\pm$ 2.3 |

**Table 4: Ablation studies on MOCHA components.** Blocks: PCA dimension  $d_t$ , MOCHA components, pretrainings.

| Pretr. | $d_t$ | Case                  | PerSeg<br>1 SHOT      | iCubWorld<br>1 SHOT   | 5 SHOT                |
|--------|-------|-----------------------|-----------------------|-----------------------|-----------------------|
| OI     | –     | (YOLOv8n)             | 41.2 $\pm$ 2.6        | 51.2 $\pm$ 3.0        | 68.4 $\pm$ 2.4        |
| AuXFT  | 256   | –                     | <u>56.4</u> $\pm$ 3.3 | 61.2 $\pm$ 2.6        | 75.7 $\pm$ 2.1        |
| AuXFT  | 384   | –                     | 56.2 $\pm$ 3.5        | 59.6 $\pm$ 3.0        | 74.7 $\pm$ 2.0        |
| AuXFT  | 1024  | –                     | 53.8 $\pm$ 3.5        | 58.4 $\pm$ 3.1        | <u>77.0</u> $\pm$ 2.3 |
| AuXFT  | 512   | $\mathcal{L}_{emb}$   | 56.6 $\pm$ 3.3        | 59.8 $\pm$ 2.9        | 74.9 $\pm$ 2.0        |
| AuXFT  | 512   | $t_S$ no attn.        | 56.8 $\pm$ 3.3        | 57.5 $\pm$ 3.1        | 73.2 $\pm$ 2.1        |
| AuXFT  | 512   | $t_S$ no fn           | 55.7 $\pm$ 3.4        | <b>61.4</b> $\pm$ 3.0 | 76.6 $\pm$ 2.3        |
| AuXFT  | 512   | $\mathcal{D}_c$ augm. | 53.3 $\pm$ 3.9        | 58.4 $\pm$ 2.9        | 73.4 $\pm$ 2.1        |
| AuXFT  | 512   | 2 $\times$ lr (enc)   | 53.6 $\pm$ 2.9        | 59.0 $\pm$ 3.0        | 75.8 $\pm$ 2.5        |
| AuXFT  | 512   | 2 $\times$ lr (dec)   | 53.9 $\pm$ 2.9        | 59.7 $\pm$ 3.1        | 75.9 $\pm$ 2.5        |
| OI     | 512   | (MOCHA)               | 53.8 $\pm$ 3.2        | 56.8 $\pm$ 2.7        | 70.4 $\pm$ 2.3        |
| COCO   | 512   | (MOCHA)               | 55.4 $\pm$ 3.3        | 60.6 $\pm$ 3.3        | <b>77.6</b> $\pm$ 2.5 |
| AuXFT  | 512   | (MOCHA)               | <b>59.1</b> $\pm$ 3.3 | <b>61.4</b> $\pm$ 3.1 | <u>77.0</u> $\pm$ 2.3 |

performance, indicating a loss of relevant semantic content, while larger ones (*e.g.*, 1024) increase computational cost without significant gains. Next, we examine the role of core MOCHA components. Removing the embedding alignment loss ( $\mathcal{L}_{emb}$ ) degrades accuracy on both datasets, underscoring its importance for preserving global teacher structure. Disabling attention mechanisms in the translator ( $t_S$  no attn.) also leads to a consistent drop, suggesting that attention layers play a key role in adapting multimodal features. A further linear-translator ablation is found in the Appendix. Likewise, removing the translator’s final layer ( $t_S$  no fn) reduces performance. Enabling data augmentation (standard YOLO augmentations only varying color,  $\mathcal{D}_c$  augm.) lowers robustness, while manipulating learning rates — doubling them for encoder or decoder separately — slightly harms performance. Finally, we compare different pretraining strategies. When pretraining is performed on OpenImages only, performance remains below that of MOCHA configurations. COCO-pretrained weights yield a substantial boost, but AuXFT-pretrained weights lead to the best results overall, thanks to their strong initial performance and architectural compatibility with MOCHA.

## 5 Conclusion

In this work, we introduced MOCHA, an effective multimodal objects-aware cross-architecture alignment technique that transfers region-level knowledge from large multimodal foundation models into compact vision detectors, enabling robust few-shot personalized object detection. Across four personal benchmarks (PerSeg, POD, COr50, and iCubWorld), MOCHA consistently surpasses prior approaches, achieving an average improvement of +10.1 over the YOLOv8n baseline and outperforming the best competitor, AuXFT, by +4.9. Our method generalizes well across different student architectures, including YOLOv11n and RT-DETR-l. MOCHA transfers knowledge from large multimodal models into compact visual detector which are well suited for edge deployment in resource-constrained scenarios, such as mobile and robotic platforms.

## References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millicah, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: International Conference on Neural Information Processing Systems (NeurIPS). Curran Associates Inc., Red Hook, NY, USA (2022)
2. Barbato, F., Camuffo, E., Milani, S., Zanuttigh, P.: Continual road-scene semantic segmentation via feature-aligned symmetric multi-modal network. In: IEEE International Conference on Image Processing (ICIP). pp. 722–728. IEEE (2024)
3. Barbato, F., Michieli, U., Moon, J., Zanuttigh, P., Ozay, M.: Cross-architecture auxiliary feature space translation for efficient few-shot personalized object detection. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2024)
4. Camuffo, E., Michieli, U., Milani, S.: Learning from mistakes: Self-regularizing hierarchical representations in point cloud semantic segmentation. *IEEE Transactions on Multimedia* (2023)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019)
6. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Fanello, S.R., Ciliberto, C., Santoro, M., Natale, L., Metta, G., Rosasco, L., Odone, F.: iCub World: Friendly Robots Help Building Good Vision Data-Sets. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 700–705 (6 2013). <https://doi.org/10.1109/CVPRW.2013.106>
8. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (2023)
9. Gou, J., Chen, Y., Yu, B., Liu, J., Du, L., Wan, S., Yi, Z.: Reciprocal teacher-student learning via forward and feedback knowledge distillation. *IEEE Transactions on Multimedia* **26**, 7901–7916 (2024)
10. Gu, J., Ghiasi, G., Cui, Y., Wang, Z., et al.: Vild: Open-vocabulary object detection via vision and language knowledge distillation. *International Conference on Learning Representations (ICLR)* (2022)
11. Hao, Z., Guo, J., Han, K., Tang, Y., Hu, H., Wang, Y., Xu, C.: One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In: IEEE International Conference on Neural Information Processing Systems (NeurIPS) (2023)
12. Hao, Z., Luo, Y., Wang, Z., Hu, H., An, J.: CDFKD-MFS: Collaborative Data-Free Knowledge Distillation via Multi-Level Feature Sharing. *IEEE Transactions on Multimedia* **24**, 4262–4274 (2022)
13. Huh, M., Cheung, B., Wang, T., Isola, P.: The platonic representation hypothesis. *Proceedings of Machine Learning Research (PMLR)* (2024)
14. Jaritz, M., Vu, T.H., Charette, R.d., Wirbel, E., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12605–12614 (2020)

15. Jaritz, M., Vu, T.H., De Charette, R., Wirbel, É., Pérez, P.: Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **45**(2), 1533–1544 (2022)
16. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 [computer software]. <https://github.com/ultralytics/ultralytics>. accessed july 2024 (2023), <https://github.com/ultralytics/ultralytics>
17. Jocher, G., Qiu, J.: Ultralytics yolo11 (2024), <https://github.com/ultralytics/ultralytics>
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 4015–4026. IEEE (2023)
19. Kothapalli, V.: Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research* (2022)
20. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)* **128**(7), 1956–1981 (2020)
21. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10965–10975 (June 2022)
22. Li, M., Zhang, L., Zhu, M., Huang, Z., Yu, G., Fan, J., Chen, T.: Lightweight model pre-training via language guided knowledge distillation. *IEEE Transactions on Multimedia* **26**, 10720–10730 (2024)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *European Conference on Computer Vision (ECCV)*. pp. 740–755. Springer (2014)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in Neural Information Processing Systems* **36**, 34892–34916 (2023)
25. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. *International Conference on Learning Representations (ICLR)* (2023)
26. Liu, Y., Cao, J., Li, B., Hu, W., Ding, J., Li, L.: Cross-architecture knowledge distillation. In: *IEEE/CVF Proceedings of the Asian Conference on Computer Vision (ACCV)*. pp. 3396–3411 (2022)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 10012–10022. IEEE (2021)
28. Lomonaco, V., Maltoni, D.: Core50: a new dataset and benchmark for continuous object recognition (2017)
29. Luo, S., Wang, X., Fang, G., Hu, Y., Tao, D., Song, M.: Knowledge amalgamation from heterogeneous networks by common feature learning. *International Joint Conference on Artificial Intelligence (IJCAI)* (2019)
30. Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., Liu, Y.: Rtdetr2: All-in-one detection transformer beats yolo and dino (2024)
31. Ma, S., Wang, Y., Wei, Y., Zhang, E., Fan, J., Sun, X., Chen, P.: Skdf: A simple knowledge distillation framework for distilling open-vocabulary knowledge to open-world object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(12), 11738–11752 (2025). <https://doi.org/10.1109/TPAMI.2025.3600435>

32. Meng, Z., Xia, X., Ma, J.: Toward foundation models for inclusive object detection: Geometry- and category-aware feature extraction across road user categories. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **54**(11), 6570–6580 (2024)
33. Michieli, U., Moon, J., Kim, D., Ozay, M.: Object-conditioned bag of instances for few-shot personalized instance recognition. In: *IEEE/SPS International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 7885–7889. IEEE (2024)
34. Ni, Z.L., Yang, F., Wen, S., Zhang, G.: Dual relation knowledge distillation for object detection (2023)
35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
36. Papayan, V., Han, X.Y., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences* **117**(40), 24652–24663 (2020). <https://doi.org/10.1073/pnas.2015509117>
37. Paramonov, K., Zhong, J.X., Michieli, U., Moon, J., Ozay, M.: Swiss dino: Efficient and versatile vision framework for on-device personal object search. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 2564–2571. IEEE (2024)
38. Passalis, N., Tzelepi, M., Tefas, A.: Heterogeneous knowledge distillation using information flow modeling. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2339–2348. IEEE (2020)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research (PMLR)* (2021)
40. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: *International Conference on Learned Representations (ICLR)* (2015)
41. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: *Advances in Neural Information Processing Systems*. vol. 30 (2017)
42. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: *International Conference on Machine Learning (ICML)*. *Proceedings of Machine Learning Research (PMLR)*, vol. 139, pp. 10347–10357 (18–24 Jul 2021)
43. Tu, Z., Zhou, W., Qian, X., Yan, W.: Hybrid knowledge distillation network for RGB-D co-salient object detection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* pp. 1–12 (2025)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in Neural Information Processing Systems* **30** (2017)
45. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023)
46. Xia, R., Zhao, C., Zheng, M., Wu, Z., Sun, Q., Tang, Y.: Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 21572–21581 (2023)
47. Zhang, L., Shi, Y., Shi, Z., Ma, K., Bao, C.: Task-oriented feature distillation. *Advances in Neural Information Processing Systems* **33**, 14759–14771 (2020)

- 48. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Qiao, Y., Gao, P., Li, H.: Personalize segment anything model with one shot. In: International Conference on Learning Representations (ICLR) (2024)
- 49. Zhang, Y., Zhang, C.B., Jiang, P.T., Cheng, M.M., Mao, F.: Personalized image semantic segmentation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10549–10559. IEEE (2021)