SyncVL: Synchronizing Vision $\Longleftrightarrow$ Language Using Unsupervised Adaptation

Maria Marrium[1], Muhammad Haris Khan[2], Sajid Javed[3], and Arif Mahmood*[1]

[1] Information Technology University, Lahore, Pakistan
{phdcs22002,arif.mahmood}@itu.edu.pk
[2] MBZUAI, Abu Dhabi, UAE {muhammad.haris@mbzuai.ac.ae}
[3] Khalifa University of Science and Technology, Abu Dhabi, UAE
{sajid.javed@ku.ac.ae}

Abstract. Contrastive vision–language models (VLMs) such as CLIP, EVACLIP, ImageBind, and SigLIP exhibit impressive zero-shot generalization, yet their performance on specific data distributions can further be enhanced by improving the synchronization of the vision and language representations. For this purpose, we present `SyncVL`, a bi-directional VL synchronization framework that enables fully unsupervised adaptation of VLMs. SyncVL jointly refines visual and textual representations through two lightweight transformer-based encoder-decoder synchronizers (Vision-to-Text and Text-to-Vision) that progressively project VL representation into a shared latent subspace without requiring labeled data. To ensure robust synchronization, we form groups using multiple image-views and text templates. Both synchronizers are trained using a reward function motivated by group relative policy optimization while keeping the original VLMs frozen. Additionally, a mutual-distillation mechanism is proposed to iteratively reinforce each synchronizer. We extensively evaluate `SyncVL` on 24 benchmark datasets spanning unsupervised VL adaptation, out-of-distribution matching, unsupervised clustering, cross-modal retrieval, object detection, and segmentation. Across all tasks, `SyncVL` consistently improves over state-of-the-art baselines, establishing a new standard for unsupervised adaptation of VLMs.

Keywords: Vision Language Models· Unsupervised Adaptation· VL Synchronization· Reinforcement Learning· Group Relative Policy Optimization· Zero-Shot Classification· Out-of-distribution Adaptation

1 Introduction

Large-scale VLMs such as CLIP [57], EVACLIP [68], ImageBind [23], and SigLIP [72], learn joint multi-modal vision-language representations within a shared embedding space using a dual-encoder architecture. The VL synchronization serves as the foundation for many downstream applications, such as zero-shot

* Corresponding author

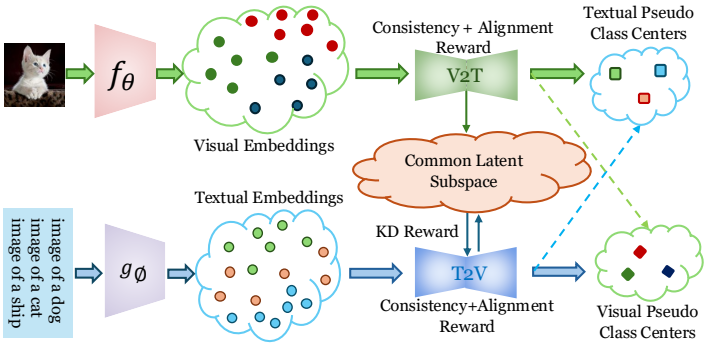
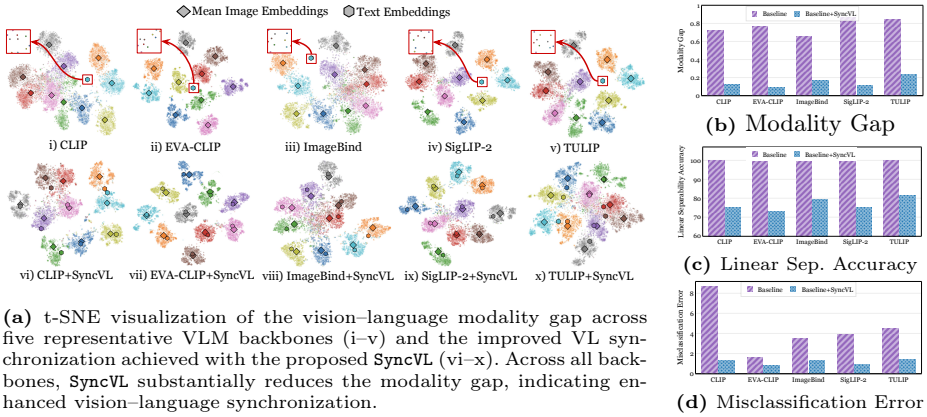


Fig. 1: Basic idea of the proposed label-free bi-directional vision-to-text (V2T) and Text-to-Vision (T2V) synchronizers (`SyncVL`).

[3, 46, 50, 71, 73, 94] and few-shot classification [61, 63, 66], image clustering [1, 18, 19], image–text understanding and retrieval [16, 25, 33, 42], VL reasoning and captioning [11, 37, 43], VL agents and robotics [13, 82, 97], content moderation [20, 96], tagging [32, 88], and recommendation [44, 101]. These VLMs exhibit good generalization performance in a number of applications; however, the inherent difference between visual and language modalities, as well as the shift within the visual domain, limits the transferability of learned representations to novel or specialized visual domains [4, 57]. Thus, the synchronization of vision and language representations can be further improved, resulting in a reduction of the modality gap.

Despite many advances in this direction, the VL synchronization issue is still pertinent and needs special attention. To address this limitation, several studies have explored techniques to enhance VL synchronization by leveraging labeled samples from target domains [34, 35, 92, 93, 99, 102]. These methods fine-tune CLIP or its prompts, thereby improving VL synchronization. However, by fine-tuning on a small dataset, the model can overfit to unwanted features or biases, resulting in loss of performance on unseen samples [22, 34, 63, 100]. In addition, collecting even a modest number of labeled samples may become infeasible in some real-world scenarios due to high annotation costs, data scarcity, or privacy constraints. Consequently, unsupervised VL synchronization, which seeks to map and align VL modalities using unlabeled data, has gained increasing attention as a promising alternative [2, 3, 30, 50, 71].

Several recent works have proposed unsupervised approaches to improve VL synchronization [3, 30, 38, 50, 55, 71]. A common strategy involves generating pseudo-labels, which are then used to fine-tune the model on the target domain in a self-supervised manner. Although effective to some extent, such fine-tuning-based adaptation methods may suffer from a few key drawbacks. First, updating the large-scale VLM encoders may erode the pre-trained multimodal knowledge, thereby weakening VLMs' inherent generalization ability [22, 31, 95, 100]. Second, pseudo-label noise may accumulate through iterative optimization, leading to confirmation bias and overfitting to incorrect labels [5, 50, 89].



(a) t-SNE visualization of the vision–language modality gap across five representative VLM backbones (i–v) and the improved VL synchronization achieved with the proposed `SyncVL` (vi–x). Across all backbones, `SyncVL` substantially reduces the modality gap, indicating enhanced vision–language synchronization.

Fig. 2: Impact of `SyncVL` on VL synchronization: (a–b) The modality gap [40] is significantly reduced across different VLM backbones. **(c)** Linear separability accuracy [15], measuring the percentage of vision and language embeddings separable by a linear classifier in the VLM embedding space. Lower values indicate stronger vision–language synchronization. **(d)** Misclassification error is reduced. Lower values are better for all measures. Results are reported on the CIFAR-10 validation split.

In this work, we propose SyncVL, a novel label-free bi-directional vision language synchronization framework that preserves the pretraining knowledge of VLMs while introducing two lightweight VL synchronizers: Vision-to-Text (V2T) and Text-to-Vision (T2V), as illustrated in Fig. 1. Unlike existing unidirectional adaptation strategies [3, 50], our bidirectional design enables mutual refinement between vision and language representations, leading to more stable and effective VL synchronization.

Furthermore, in contrast to prior methods that rely on supervised fine-tuning or generative feedback [31, 34, 100], SyncVL operates in a fully self-supervised manner using consistency and alignment rewards. The V2T and T2V synchronizers collaboratively adjust VL embeddings by encouraging intra-group consistency and inter-modal alignment, progressively narrowing the modality gap. To ensure stable optimization, we employ a relative reward-driven objective that evaluates synchronization quality within structured groups rather than relying on absolute loss values. This cooperative learning strategy effectively enhances adaptation robustness without requiring labeled data or encoder fine-tuning.

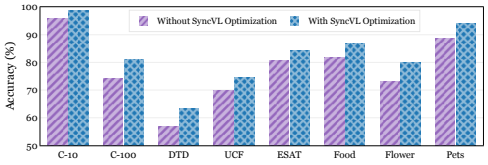


Fig. 3: Optimizing `SyncVL` using the proposed objective functions consistently improves performance.

The average similarity between the vision and language class-centers significantly improves by incorporating `SyncVL` with the original baselines. Fig. 2 shows

the improvement in VL synchronization for five different VLM backbones, including CLIP (B/32) [57], EVA-CLIP (B/16) [68], ImageBind (H/14) [23], SigLIP-2 (L/16) [72], and TULIP (B/16) [70]. The performance gain of `SyncVL` over baselines can be attributed to the synergy of gradients between alignment and consistency losses. The same losses if used without our proposed reward functions, conflict with each other, resulting in performance degradation as shown in Fig. 3. Rigorous experimental evaluations are performed on 21 benchmark datasets for unsupervised adaptation, out-of-distribution zero-shot classification, few-shot classification, unsupervised image clustering, cross-modal retrieval, object detection, and segmentation. Our results demonstrate that the proposed `SyncVL` outperforms existing State-Of-The-Art methods by a significant margin. Overall, our main **contributions** are summarized as follows.

1. We propose `SyncVL`, self-supervised bidirectional synchronization framework which maps Vision-to-Text (V2T) and Text-to-Vision (T2V) using consistency and alignment based rewards. Both synchronizers are trained using the objective function inspired by group relative policy optimization.
2. The consistency reward forces the synchronizers to map the full input group to the same pseudo-class in the output. The alignment reward forces the synchronizers to map input embeddings of one modality to the center of the pseudo-class embeddings of the second modality, significantly improving the VL synchronization.
3. The V2T is considered as a teacher for knowledge distillation to the T2V as a student to align its latent space for bidirectional coherence. The aligned latent space acts as a common subspace and used for inference.

2 Related Work

VL Models and Alignment. Large-scale vision–language models (VLMs) such as CLIP [57], EVA-CLIP [68], and ImageBind [23] learn a shared embedding space for images and text through contrastive training on large-scale paired datasets. This enables strong zero-shot generalization by synchronizing multi-modal representations without task-specific supervision. However, the synchronization learned during pretraining is often static and domain-dependent, which can lead to performance degradation under distribution shifts, corruptions, and domain generalization scenarios [17, 79, 83]. Recent works attempt to improve VLM's alignment and robustness using prompt learning [22, 51, 62, 100], adapter tuning [66, 91], or adversarial alignment [96], but these approaches require training-time updates and cannot adapt dynamically at inference. Another direction investigates unsupervised adaptation using unlabeled target data [38, 50, 55, 71]. However, many of these methods still rely on auxiliary supervision signals or incur significant computational overhead. More recent approaches, such as ReCLIP [30], DPA [3], and FAIR [2], address VL misalignment through source-free domain adaptation. Despite their effectiveness, these methods typically involve fine-tuning VLM components, which may degrade pretrained multi-

modal knowledge and suffer from pseudo-label noise accumulation during iterative optimization [5, 50, 89].

Test-Time Adaptation & Consistency Learning. Test-time adaptation (TTA) improves robustness to unseen domains by adapting models during inference [46, 63]. Methods such as TENT [75], MEMO [90], CoTTA [78], TTT++ [45], TTT-MAE [21], and SHOT [39] employ entropy minimization, self-supervised objectives, or pseudo-labeling for adaptation. However, these approaches primarily optimize surrogate objectives and often lack mechanisms to enforce prediction stability under augmentations. Consistency-based learning [65, 81] provides a natural signal for more robust adaptation. Our proposed `SyncVL` is applicable both using the training data or the test data without using image-labels. The consistency and alignment rewards are transformed into a reinforcement learning framework, allowing the model to self-optimize its representations online based on group relative policy optimization.

Reinforcement Learning for Vision-Language Alignment. Reinforcement learning (RL) has been used to optimize language or multi-modal models via human or model feedback. RLHF [9, 53] and RLAIF [6] have enabled preference-aligned language models, while in the multi-modal domain, GRPO (Group Relative Policy Optimization) [60] provides a unified policy optimization framework for aligning model outputs to preference signals. TTRV [64] applied this idea to decoder-based VLMs, showing that reinforcement signals at test time can directly improve downstream VL reasoning. However, applying reinforcement-based adaptation to dual-encoder architectures like CLIP is non-trivial because such models lack a token-level decoding policy. Previous work treats alignment as a fixed contrastive objective, limiting the use of RL-based optimization. Our proposed `SyncVL` bridges this gap by redefining the embedding generation process as a latent policy, where the reward is derived from prediction consistency rather than explicit labels or human feedback. This formulation generalizes the RL-based optimization idea from generative to contrastive VLMs and introduces a novel reinforcement-driven alignment mechanism.

3 Vision$\Longleftrightarrow$Language Synchronization Framework

In this work, we propose `SyncVL`, a bidirectional vision-language synchronization framework that enhances VL synchronization of the pre-trained dual-encoder VLMs. A schematic illustration of `SyncVL` is shown in Fig. 4. We propose Vision-to-Text (V2T) and Text-to-Vision (T2V) synchronizers trained using an objective function inspired by group reward policy optimization. The reward is based on the input images' augmentation consistency and pseudo-label-based contrastive loss. Together, these synchronizers enhance the vision-language alignment of dual-encoder VLMs.

3.1 Preliminaries

Dual-Encoder VLMs. Pre-trained dual-encoder VLMs employ an image encoder f_θ and a text encoder g_ϕ to project unlabeled images $\{\mathbf{x_j}\}_{j=1}^{n}$ and textual

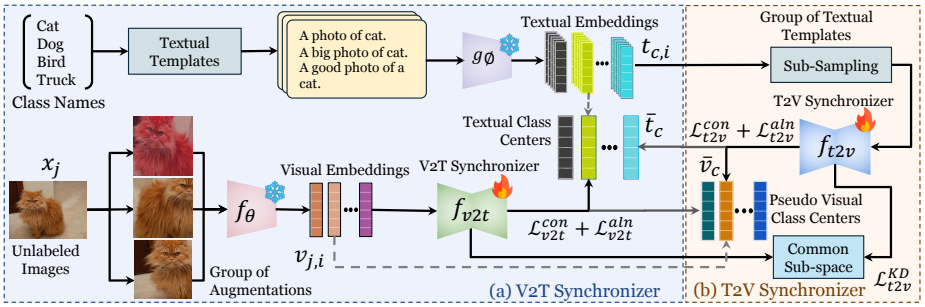


Fig. 4: Proposed bi-directional vision-language synchronization framework (`SyncVL`): (A) f_{v2t} maps visual embeddings to textual class centers, $\bar{\mathbf{t}}_c$. In later iterations $\bar{\mathbf{t}}_c$ is obtained from f_{t2v}. (B) f_{t2v} maps textual embeddings to visual pseudo class centers, $\bar{\mathbf{v}}_c$ obtained from trained f_{v2t}.

templates $\mathbf{T}_{c=1}^{C}$ into a shared embedding space, where C are the total classes. The encoders are pretrained with a contrastive loss to align corresponding image–text pairs and separate mismatched ones. During inference, zero-shot classification selects the class whose text embedding best matches the image embedding.

Group Relative Policy Optimization (GRPO). To improve VL synchronization, we get motivation from *GRPO*, a reinforcement learning algorithm originally proposed to stabilize fine-tuning of LLMs without requiring an explicit critic network [24,60]. GRPO reformulates the policy optimization process by evaluating and updating the model based on the *relative performance* of multiple responses generated for the same input rather than their absolute reward values. In our work, the policy π_θ represents a V2T or T2V synchronizer. For V2T, for each input image, we generate a group of augmented views, each producing a corresponding textual output. These samples collectively form a *group*, and the model is trained to prefer outputs that are more semantically consistent across augmentations and better aligned with the pseudo-positive textual embeddings. For T2V, a group is formed over a set of textual embeddings. Once V2T is trained, then under its supervision, the T2V is trained employing GRPO.

3.2 Formulating Vision$\Longrightarrow$Text Synchronizer Reward

Textual Class Centers. From a large set of text templates, we randomly select a subset of m templates per class $\{\mathbf{T}_{c,i}\}_{i=1}^{m}$ and compute textual embeddings using VLMs text encoder $\mathbf{t}_{c,i} = g_\phi(\mathbf{T}_{c,i})$ which are then averaged to obtain class-level textual representation $\bar{\mathbf{t}}_c = 1/m \sum_{i=1}^{m} \mathbf{t}_{c,i}$.

V2T Group Formation. Given an image $\mathbf{x}_j$, its visual embedding $\mathbf{v}_j = f_\theta(\mathbf{x}_j)$ is transformed into the text-aligned semantic space via the proposed V2T synchronizer. To make a group, we apply *k-augmentations* to the image $\mathbf{x}_j$ as $\{\mathbf{x}_{j,1}, \ldots, \mathbf{x}_{j,k}\}$. For these augmentations, we obtain visual embeddings $\{\mathbf{v}_{j,1}, ..., \mathbf{v}_{j,k}\}$ which are then mapped to the textual space using V2T: $\mathbf{v}_{j,i}^{t} = f_{v2t}(\mathbf{v}_{j,i})$ and $\mathbf{v}_{j_i}^{t} \in \mathbb{R}^d$. Together, the input visual embeddings $\{\mathbf{v}_{j,i}\}_{i=1}^{k}$ and the output

visual embeddings $\{\mathbf{v}_{j,i}^t\}_{i=1}^k$ form a group. After initial phase of training of f_{v2t}, the group definition is changed to all images $\mathbf{x}_j$ having the same pseudo-class label form a group for training in the later phase.

Consistency Reward. The similarity of each instance-level visual embedding $\mathbf{v}_{j,i}^t$ to each text embedding class-center $\bar{\mathbf{t}}_c$ is computed as: $s_{(j,i,c)} = \cos(\mathbf{v}_{j,i}^t, \bar{\mathbf{t}}_c)$ and normalized into class distributions using softmax with temperature τ.

$$p_{(j,i,c)} = \frac{\exp(s_{(j,i,c)}/\tau)}{\sum_{c'} \exp(s_{(j,i,c')}/\tau)}, \tag{1}$$

Thus, for each instance $\mathbf{v}_{j,i}^t$, we get a vector of probabilities over all classes $\mathbf{p}_{j,i} \in \mathcal{R}^C$. The objective is to achieve a stable similarity distribution in the group by minimizing the KL divergence between the individual predictions and the average group prediction $\bar{\mathbf{p}}_j = \frac{1}{k}\sum_{i=1}^k \mathbf{p}_{j,i}$:

$$\mathcal{L}_{v2t}^{\text{con}}(j,i) = D_{\text{KL}}(\bar{\mathbf{p}}_\mathbf{j} \,\|\, \mathbf{p}_{j,i}), \tag{2}$$

Alignment Reward It aims to improve alignment between the individual visual embeddings $\mathbf{v}_{j,i}^t$ in the group with a pseudo-class center text embedding $\bar{\mathbf{t}}_{c^+}$. The pseudo-class label is determined by finding the majority label across the group of $\mathbf{x}_j$ augmentations. We also define a pseudo-hard negative class $\bar{\mathbf{t}}_{c^-}$ as the second-best matching class in the group. We aim to bring $\mathbf{v}_{j,i}^t$ close to $\bar{\mathbf{t}}_{c+}$ and far from $\bar{\mathbf{t}}_{c,-}$ by using contrastive loss.

$$\mathcal{L}_{v2t}^{\text{aln}}(j,i) = -\log \frac{\exp(\frac{s_{(j,i,c^+)}}{\tau})}{\exp(\frac{s_{(j,i,c^+)}}{\tau}) + \exp(\frac{s_{(j,i,c^-)}}{\tau})}, \tag{3}$$

Overall Reward. Overall reward combines distributional consistency reward across augmentations and alignment-based reward between the two modalities:

$$\mathcal{L}_{v2t}(j,i) = \mathcal{L}_{v2t}^{\text{con}}(j,i) + \mathcal{L}_{v2t}^{\text{aln}}(j,i), \tag{4}$$

3.3 Formulating Text$\Longrightarrow$Vision Synchronizer Reward

The trained f_{v2t} is used to provide the supervision to the reverse synchronizer f_{t2v}, which maps textual templates to the visual space. This supervision is provided as image pseudo-labels and knowledge distillation in the latent space. The KD loss aims to align the latent space of f_{t2v} with the latent space of f_{v2t}.

Visual Pseudo-class Center. The trained f_{v2t} is used to assign a pseudo-class label c to each $\mathbf{v}_j$. For all $\mathbf{v}_j$ belonging to the same pseudo class c, a mean visual embedding is computed as a representative of that class $\bar{\mathbf{v}}_c = \frac{1}{n_c}\sum_{j=1}^{n_c} \mathbf{v}_j$, where n_c is the number of images having the same pseudo-label c. We employ this average class embedding as a visual anchor for aligning textual representations.

T2V Group Formation. For each text class c we have m templates $\{\mathbf{t}_{c,i}\}_{i=1}^m$. The proposed T2V synchronizer aims to map each of these to the visual space: $\mathbf{t}_{c,i}^v = f_{t2v}(\mathbf{t_{c,i}})$ and $\quad \mathbf{t}_{c,i}^v \in \mathbb{R}^d$. The set of m templates and their corresponding

outputs, $\{\mathbf{t}^v_{c,i}\}_{i=1}^m$ collectively form a T2V group. The model f_{t2v} is trained to prefer outputs that are consistent across all templates and better aligned with the visual representation of the corresponding pseudo positive classes using the following three rewards.

Consistency Reward $\mathcal{L}^{\text{con}}_{t2v}$. After mapping a textual embedding $\mathbf{t}_{c,i}$ to the visual space $\mathbf{t}^v_{c,i}$, the similarity is calculated with each class mean visual embedding $\bar{\mathbf{v}}_c$, $s_{c,i} = \cos(\mathbf{t}^v_{c,i}, \bar{\mathbf{v}}_c)$, and probability distributions $\mathbf{p}_{c,i}$ are computed using softmax with temperature as in Eq. (1). The KL divergence is minimized between the mean of class probabilities $\bar{\mathbf{p}}_c = 1/m \sum_{i=1}^m \mathbf{p}_{c,i}$, and each individual vector $\mathbf{p}_{c,i}$ within the group as in Eq. (2).

Alignment Reward $\mathcal{L}^{\text{aln}}_{t2v}$. The objective is to improve the synchronization between the class mean-visual-embedding $\hat{v}_c$ and the template-embedding $\mathbf{t}^v_{c,i}$ within the group. For the whole group of $\mathbf{t}^v_{c,i}$, one positive class c^+ is selected based on the majority vote and the second best is selected as the hard negative c^-, and then the contrastive loss is minimized as in Eq. (3).

Knowledge Distillation Reward $\mathcal{L}^{\text{KD}}_{t2v}$. The trained f_{v2t} also acts as a teacher and provides additional supervision to f_{t2v}, which acts as a student model. The KD reward minimizes the distance between the textual templates and the corresponding pseudo-class visual representations in the latent space and improves bidirectional coherence.

$$\mathcal{L}^{\text{KD}}_{t2v}(c,i) = \|f^e_{v2t}(\bar{\mathbf{v}}_c) - f^e_{t2v}(\mathbf{t_{c,i}})\|_1, \tag{5}$$

where f^e_{v2t} and f^e_{t2v} are the encoders of the two synchronizers.

Overall T2V Reward. T2V reward consists of distributional consistency reward across m text templates and alignment-based reward:

$$\mathcal{L}_{t2v}(c,i) = \mathcal{L}^{\text{con}}_{t2v}(c,i) + \mathcal{L}^{\text{aln}}_{t2v}(c,i) + \mathcal{L}^{\text{KD}}_{t2v}(c,i) \tag{6}$$

The f_{t2v} synchronizer is also trained using the group-relative optimization as discussed above.

3.4 Overall Training

Iterative Training In the initial phase, f_{v2t} is first trained and the training of f_{t2v} follows. In the next phase, visual class centers are calculated using f_{v2t}: $\bar{\mathbf{v}}_c = \frac{1}{n_c}\sum_{j=1}^{n_c} \mathbf{v}^t_j$, and the textual class centers are calculated using f_{t2v}: $\bar{\mathbf{t}}_c = \frac{1}{m}\sum_{i=1}^{m} \mathbf{t}^v_{c,i}$. Both f_{v2t} and f_{t2v} are trained sequentially for multiple iterations to obtain convergence of visual and textual embeddings.

Group-Relative Optimization. Instead of optimizing V2T or T2V rewards in absolute terms, GRPO computes the *relative advantage* of each sample within the group. For V2T: $r_{j,i} = \mathcal{L}_{v2t}(j,i)$, $A_{j,i} = (r_{j,i} - \bar{r}_j)/(\sigma_j + \epsilon)$, where $\bar{r}_j = \frac{1}{k}\sum_{i=1}^k r_{j,i}$, and $\sigma_j = \sqrt{\frac{1}{k}\sum_{i=1}^k (r_{j,i} - \bar{r}_j)^2}$. This normalization ensures that each output's contribution is evaluated relative to those generated from the same image. The policy is then updated using the clipped objective, similar to PPO [60]:

$$\mathcal{L}_{\text{G}} = \mathbb{E}_i\Big[\min\big(r_{j,i}(\theta)A_{j,i}, \text{clip}(r_{j,i}(\theta), 1-\epsilon, 1+\epsilon)A_{j,i}\big)\Big],$$

where $r_{j,i}(\theta) = \pi_\theta(\mathbf{v}^t_{j,i}|\mathbf{v}_{j,i})/\pi_{\theta_{\text{old}}}(\mathbf{v}^t_{j,i}|\mathbf{v}_{j,i})$ is the likelihood ratio ensuring stable updates. Conceptually, this means the synchronizer f_{v2t} learns to generate visual representations $\mathbf{v}^t_{j,i}$ that are not only individually accurate but also more consistent and better aligned relative to its own alternatives. The T2V synchronizer is also being trained using a similar GRPO-based methodology. By optimizing over visual and textual groups, GRPO naturally enforces invariance to visual and textual perturbations and improves fine-grained VL correspondence, pushing the policy toward a more stable and semantically robust alignment regime.

Inference. At inference time, we only use the encoders f^e_{v2t} and f^e_{t2v} to project vision-textual embeddings in the aligned latent subspace obtained using the `SyncVL` framework. Thus, our model enforces distributional stability across augmented inputs, directional contrastive alignment between modalities, and ensures latent symmetry between the two synchronizers. This bidirectional reward-guided optimization fosters stronger VL synchronization and improves VLM's zero-shot performance on diverse downstream tasks.

4 Experiments and Results

Datasets. For unsupervised adaptation, we evaluated the proposed `SyncVL` on 12 datasets consisting of: *General classification* datasets: CIFAR-10/100 [36], GTSRB [29], and ImageNet [12]; *Fine-grained classification* datasets: DTD [10], Food101 [8], Oxford-IIIT Pets [54], FGVC-Aircraft [47], Flowers102 [52], and Birdsnap [7]; *Action recognition* dataset: UCF101 [67]; *Satellite imagery* dataset: EuroSAT [26] and report results using the test splits provided by [100]. To evaluate the `SyncVL` adaptation abilities to distribution shift, we evaluated on three *Out-Of-Distribution* (OOD) datasets: ImageNet-Rendition [27], ImageNet-Sketch [76], and ImageNet-Adversarial. We evaluated `SyncVL` for *object detection and segmentation* using two widely used datasets: PASCAL-VOC [14] and MS-COCO [41]. We followed the experimental settings of [98] for object detection

Table 1: Comparison of SOTA unsupervised adaptation methods using the CLIP ViT-B/32 backbone.

Methods	CIFAR10	CIFAR100	DTD	UCF101	EuroSAT	Food101	Pets	Flowers	Aircraft	Cars	Caltech101	SUN397	ImageNet	Avg.
CLIP (B/32)	91.30	65.10	44.50	64.50	49.40	82.40	87.00	66.46	21.20	59.40	87.90	63.20	63.20	65.04
UPL† [31]	91.26	67.41	45.37	62.04	51.88	84.25	83.84	67.40	17.07	49.41	92.36	62.12	58.22	64.05
POUF† [71]	90.50	62.00	46.10	61.20	62.90	82.10	87.80	67.80	18.20	57.70	94.10	60.00	52.20	64.82
CPL [89]	94.68	77.30	56.30	71.00	82.20	83.18	87.62	76.70	19.76	-	-	-	63.74	68.54
TransCLIP [85]	91.18	67.38	50.40	59.00	81.50	89.00	74.40	68.70	20.30	63.57	91.72	67.44	64.90	68.42
LafTer [50]	95.80	74.60	46.10	73.90	82.45	84.93	71.00	68.20	19.86	-	93.30	64.50	64.20	69.90
ReCLIP† [30]	94.84	72.26	53.88	67.01	70.8	84.19	87.49	72.63	18.87	59.06	92.43	73.89	64.52	70.08
GTA-CLIP [59]	-	-	57.51	71.00	77.38	66.31	90.81	79.74	21.21	64.27	94.16	70.14	66.31	68.99
DPA [3]	95.97	76.47	55.69	68.49	80.04	84.76	90.71	75.56	20.67	62.62	96.06	68.13	64.64	72.29
`SyncVL` (ours)	**98.65**	**81.14**	**63.28**	**74.5**	**84.14**	**86.98**	**93.96**	**79.84**	**25.06**	**67.12**	**96.54**	**74.84**	**68.24**	**76.48**

†Results are adapted from DPA [3].

Table 2: Unsupervised adaptation results on additional backbones.

Method	C-10	C-100	DTD	UCF	ESAT	Food	Pets	Avg.
CLIP (B/16) [57]	90.80	68.22	44.70	69.10	49.00	88.70	89.00	71.36
+SyncVL	**98.95**	**82.31**	**60.68**	**81.87**	**83.24**	**90.86**	**96.02**	**84.85**
SigLIP (L/16) [87]	81.67	48.70	20.90	38.09	25.53	86.58	93.2	56.38
+SyncVL	**92.12**	**68.42**	**35.72**	**55.02**	**44.65**	**89.82**	**95.22**	**68.71**
ImageBind (H/14) [23]	96.53	78.54	51.28	76.95	61.02	90.73	93.46	78.36
+SyncVL	**98.72**	**82.62**	**53.14**	**78.10**	**65.01**	**92.61**	**96.01**	**80.89**
EVACLIP (B/16) [68]	98.40	87.70	53.10	83.35	67.00	89.40	92.20	81.59
+SyncVL	**99.24**	**90.78**	**54.24**	**85.82**	**70.45**	**96.98**	**96.53**	**84.86**
SigLIP-2 (L/16) [72]	96.11	83.14	68.79	81.26	51.12	95.64	95.69	81.68
+SyncVL	**99.05**	**85.83**	**69.85**	**83.57**	**70.68**	**96.29**	**96.12**	**85.91**
TULIP (B/16) [70]	95.51	78.75	64.36	74.68	45.28	92.47	94.88	77.99
+SyncVL	**98.63**	**83.09**	**67.26**	**78.25**	**63.49**	**95.26**	**96.18**	**83.17**

and [58] for segmentation. We also evaluated SyncVL for *cross-modal retrieval* using four datasets: Scene text retrieval datasets: Street View Text (SVT) [77], Total-Text Retrieval (TTR) [80], Phrase-level Scene Text Retrieval (PSTR) [86], and INQUIRE [74]. We employed the standard evaluation measure used by the SOTA methods [3, 50, 56, 57, 74]. More details are given in Supp. Section A.1.

Implementation and Training Details. Each V2T and T2V synchronizer comprises an encoder-decoder pair each having two transformer blocks. We adopt an iterative training strategy to progressively synchronize visual and textual representations. In the first iteration, V2T is trained for 100 epochs using groups of $k = 7$ augmentations per image to enforce intra-instance consistency and alignment. The resulting pseudo-labels are then used to train T2V for another 100 epochs. In subsequent iterations, pseudo-labels derived from image embeddings are used for re-grouping images, and both synchronizers are refined in an alternating manner. We use AdamW optimizer with 1×10^{-4} lr, and 0.01 weight decay for both synchronizers. Unless otherwise stated, all experiments use the CLIP (ViT-B/32) backbone. See Supp. Section A.2 for additional details. SyncVL algorithm and computational cost analysis are in Supp. Sections A.3 and A.4.

SyncVL Unsupervised Adaptation Performance. In Table 1, we compare our proposed SyncVL with eight SOTA unsupervised adaptation methods using CLIP (ViT-B/32) backbone. SyncVL consistently achieves the best performance across all 10 datasets, surpassing the previous best method (DPA) by a notable margin of 4.19%. Specifically, SyncVL demonstrates significant gains on CIFAR-100 (+4.67%) and DTD (+5.77%), indicating more effective VL synchronization compared to existing adaptation methods. These consistent improvements across diverse visual categories, from coarse to fine-grained recognition, highlight the robustness and gen-

Table 3: Comparisons on OOD datasets in transductive and unsupervised settings. CoOp is used with 16-shots.

Methods	INet-S	INet-R	INet-A
CLIP (B/32) [57]	42.10	68.60	32.10
CoOP (M=16) [100]	40.44	64.45	30.62
CLIP-Adapter (UA) [22]	32.04	54.75	20.12
Lafter (Trans.) [50]	42.70	<u>72.60</u>	31.50
DPA (Trans.) [3]	42.60	68.40	32.20
SyncVL (UA)	<u>45.96</u>	71.26	<u>35.14</u>
SyncVL (Trans.)	**47.24**	**74.18**	**36.98**

eralization ability of SyncVL in enhancing unsupervised adaptation performance. In Table 2, we present results on six additional backbones across 7 datasets. SyncVL consistently improves performance across the backbones and datasets, reaffirming the effectiveness of the proposed approach. See Supp. Section B.1 for additional comparisons. **SyncVL Adaptation to Distribution Shifts.** We evaluate SyncVL robustness to distribution shifts in both unsupervised adaptation (UA) and transductive (Trans.) settings, following the experimental settings of prior works [22, 50, 100]. In the UA setting, SyncVL is trained solely on the ImageNet training set without any access to samples from the target OOD datasets. For the transductive setting, we adopt the evaluation setup of [3,50]. As reported in Table 3, SyncVL (UA) achieves competitive or superior results compared to both few-shot and transductive baselines, demonstrating strong domain generalization capability. Further extending SyncVL to the transductive setting yields additional gains, consistently outperforming all compared methods. These results highlight SyncVL's generalization and its effectiveness in improving VL synchronization under significant distributional shifts.

SyncVL Comparison with Few-Shot Methods. We compare the proposed SyncVL (UA) with two SOTA few-shot methods under 8-shot and 16-shot settings using the CLIP (ViT-B/32) backbone. Table 4 summarizes the results across six datasets. SyncVL consistently outperforms 8-shot methods, while on 16-shot methods it obtain best performance on five out of six datasets. Unlike few-shot methods that depend on a limited set of annotated examples, SyncVL functions entirely in an unsupervised adaptation setting, requiring no labeled target data, while achieving superior accuracy. These results highlight the effectiveness and efficiency of SyncVL as a label-free alternative to few-shot adaptation techniques. See Supp. Section B.2 for more comparisons.

Table 4: Comparison of proposed SyncVL (UA) with few-shot (8-shot and 16-shot) methods using CLIP B/32 backbone.

Methods	C10	C100	DTD	ESAT	Food	Pets
CoOp (M:8) [69]	94.63	75.91	61.82	77.17	78.84	85.61
CoOp (M:16)	94.50	75.78	**67.55**	76.78	79.12	86.70
MaPLe (M:8) [34]	88.23	71.14	34.99	65.25	80.51	81.85
MaPLe (M:16)	91.94	71.74	53.43	71.49	81.01	86.54
SyncVL (UA)	**98.65**	**81.14**	63.28	**84.14**	**86.98**	**93.96**

Object Detection and Segmentation Using SyncVL. We evaluate SyncVL to test its performance beyond classification, focusing on object detection and segmentation using two backbones. Table 5 shows that SyncVL consistently improves performance across backbones, tasks, and datasets. For CLIP, SyncVL improves AP_{50} by 4.6 and 3.6 and mIoU by 11.45 and 1.61 on PASCAL VOC and MS COCO datasets. For TULIP, SyncVL improves AP_{50} by 3.44 and 3.72 and mIoU by 2.11 and 1.97 on PASCAL VOC and MS COCO datasets. This indicates that SyncVL effectively models spatial and structural information that are essential for precise localization and accurate segmentation. More results are in Supp. Section B.3.

Table 5: SyncVL improves Object Detection (AP_{50}) and Segmentation (mIOU).

Methods	PASCAL VOC		MS COCO	
	AP_{50}	mIOU	AP_{50}	mIOU
CLIP (B/16)	30.48	16.20	27.80	4.40
SyncVL (ours)	**35.09**	**27.65**	**31.40**	**6.01**
TULIP (B/16)	28.70	15.84	21.32	3.72
SyncVL (ours)	**32.14**	**17.95**	**25.04**	**5.69**

Table 6: Unsupervised clustering performance of SyncVL visual representations (CLIP B/32).

Methods	C-10	C-100	DTD	ESAT	Food	Pets
TEMI [1]	96.9	73.7	**62.3**	88.7	84.6	78.3
HUME [18]	89.2	55.5	52.3	66.7	66.7	64.4
TURTLE [19]	85.9	45.8	49.1	69.8	72.4	62.1
SyncVL (ours)	**98.8**	**79.3**	62.2	**90.1**	**87.3**	**80.7**

Table 7: Zero-shot cross-modal retrieval performance (mAP).

Methods	SVT	TTR	PSTR	INQUIRE
CLIP (B/16)	68.43	39.23	89.21	11.40
CYN-Ret [56]	75.11	62.96	-	-
SyncVL (ours)	**85.73**	**83.67**	**93.27**	**18.63**
TULIP (B/16)	86.37	42.37	51.90	23.52
SyncVL (ours)	**91.57**	**54.18**	**63.24**	**27.71**

Unsupervised Clustering Using SyncVL. To further assess the intra-class compactness and inter-class discrimination capability of SyncVL visual representation, we use k-means clustering. In Table 6, we compare SyncVL clusters with three recent SOTA methods. The number of clusters is set equal to the number of classes following [18, 19]. On average, SyncVL achieves superior performance across datasets, indicating that it effectively preserves semantic structure and intra-class cohesion of visual representations while enhancing VL synchronization even in a fully unsupervised setting. More results are in Supp. B.4.

Cross-Modal Retrieval Using SyncVL. In Table 7, we compare SyncVL cross-modal retrieval performance using two backbones and four datasets. To ensure fair comparison, all methods are evaluated under the same settings [56]. For SyncVL, the retrieval is done within the shared subspace of V2T and T2V synchronizers. SyncVL consistently improves retrieval performance over the SOTA methods across datasets. SyncVL shows an improvement of 10.62% and 20.71% on SVT and TTR dataset over CYN-Ret. Overall, SyncVL demonstrates effective unsupervised adaptation for cross-modal retrieval. See Supp. B.5 for details.

ML Safety Measures. We evaluate SyncVL on three ML safety measures: calibration (assessing how well predicted probabilities reflect actual correctness), consistency (stability of predictions under minor input changes), and corruption (robustness to input perturbations) under Gaussian (G), Shot (S), and Impulse (I) noises [28, 48, 49] using two datasets. For calibration, we report the expected calibration error using 15 bins. For consistency, we report the mean flip rate over a perturbation sequence of length 31. For corruptions, we report the mean classification error. Lower is better for all measures. As shown in Table 8, SyncVL leads to a notable reduction in calibration error, mean flip rate, and different types of corruptions across both datasets. These results demonstrate that SyncVL not only improves VL synchronization but also yields a safer and more dependable model behavior in challenging and noisy conditions. See Supp. Section B.6 for additional results.

Table 8: Evaluating SyncVL for ML Safety Measures including Calibration error (Calib.), Consistency flip-rate error (Cons.), classification error under Gaussian (G), Shot (S), and Impulse (I) feature noises.

Datasets	Methods	Calib.	Cons.	Corruptions		
				G	S	I
C-10	CLIP	78.0	7.4	22.6	27.9	37.6
	+SyncVL	**34.4**	**4.8**	**12.6**	**19.8**	**27.9**
C-100	CLIP	60.6	28.7	42.8	50.1	49.4
	+SyncVL	**27.1**	**20.6**	**31.1**	**36.5**	**41.3**

SyncVL Robustness to Poor Initialization. In Table 9, we evaluate robustness under large domain shifts using five datasets and two backbones where baseline performance is low. Despite poor initialization, SyncVL achieves consistent improvements, with average gains of +17.3 and +15.6, respectively. This

Table 9: SyncVL improves despite poor and noisy initialization.

Methods	Aircraft	DTD	ESAT	Birds	GTSRB
CLIP (RN50)	19.30	41.70	41.10	32.60	35.20
+SyncVL	**24.36**	**61.94**	**80.73**	**38.19**	**51.38**
CLIP (B/32)	21.20	44.50	49.40	37.80	32.20
+SyncVL	**25.06**	**63.28**	**84.14**	**40.76**	**49.65**

Table 10: Cosine similarity between Consistency ($\mathbf{g}^{con}$) and Alignment ($\mathbf{g}^{aln}$) rewards gradients. Positive values indicates alignment.

Cos ($\mathbf{g}^{con}$, $\mathbf{g}^{aln}$)	**CIFAR-10**		**CIFAR-100**		**DTD**	
	Init	Final	Init	Final	Init	Final
Without SyncVL	-0.17	-0.10	-0.31	-0.24	-0.15	-0.09
With SyncVL (ours)	0.44	0.63	0.59	0.65	0.47	0.56

robustness stems from its bidirectional vision–language synchronization, which progressively refines pseudo-labels by enforcing VL consistency in a shared subspace, thereby mitigating the impact of initial noise. This makes SyncVL effective even when initial zero-shot labels are unreliable. See Supp. Section B.7 for iterative pseudo-label refinement performance.

SyncVL Rewards Gradient Analysis. We adopt the definition of gradient conflict from [84] to analyze the interaction between the gradients of the consistency reward and the alignment reward. Specifically, we measure the cosine similarity between their gradients to quantify the degree of interference during optimization. Negative cosine similarity indicates conflicting update directions, whereas positive values suggest cooperative optimization. As shown in Table 10, GRPO substantially reduces gradient conflict compared to the baseline formulation. By effectively decoupling competing objectives, GRPO mitigates destructive interference between the two reward signals, leading to more stable optimization dynamics. This reduction in gradient conflict correlates with consistent and significant performance improvements across datasets and tasks.

Ablation Studies. We conduct multiple ablation studies to assess different aspects of the proposed SyncVL using three representative datasets: CIFAR-10, CIFAR-100, and DTD. First, Table 11 evaluates the **contribution of each SyncVL component** to the final performance. We compare baseline performance with several configurations: using only V2T, with and without the proposed SyncVL objective function (OF); using both V2T and T2V, with and without OF. Using only V2T yields moderate accuracies, which improve when combined with the OF (Fig. 3). Adding the T2V synchronizer further boosts performance, and the full bidirectional setup with the OF achieves the best results. This demonstrates the complementary role of each component in enhancing VL synchronization. Next, we investigate the contributions of the consistency ($\mathcal{L}^{con}$), alignment ($\mathcal{L}^{aln}$), and KD ($\mathcal{L}^{KD}$) rewards. Table 12 shows that using only $\mathcal{L}^{con}$ or $\mathcal{L}^{aln}$ for V2T and T2V yields similar accuracies, while combining them improves performance. Adding the KD reward to f_{t2v} achieves the best results,

Table 11: Performance breakdown of each SyncVL module.

V2T	T2V	OF	C-10	C-100	DTD
×	×	×	91.30	65.10	44.50
✓	×	×	86.37	68.54	50.18
✓	×	✓	93.06	71.86	53.24
✓	✓	×	95.86	74.18	56.72
✓	✓	✓	**98.65**	**81.14**	**63.28**

Table 12: Performance breakdown across each reward.

$\mathcal{L}^{con}$	$\mathcal{L}^{aln}$	$\mathcal{L}^{Lat}$	C-10	C-100	DTD
×	×	×	91.30	65.10	44.50
✓	×	×	94.18	73.69	55.24
×	✓	×	94.64	73.24	55.40
✓	✓	×	96.32	77.16	57.06
✓	✓	✓	**98.65**	**81.14**	**63.28**

Table 13: Impact of synchronizers architecture design on performance.

Sync Arch.	C-10	C-100	DTD
MLP-AE	84.72	69.31	50.12
Transformer ×1 (AE)	97.98	79.89	62.57
Transformer ×2 (AE)	98.65	**81.14**	63.28
Transformer ×4 (AE)	**98.74**	79.91	**64.13**

Table 14: Significance of augmentation groups (Aug.), pseudo-label regrouping, and iterative synchronizer training (Itr.).

Aug.	Re-Grouping	Itr.	C-10	C-100	DTD
✓	×	×	97.14	79.93	58.86
✓	✓	×	97.89	80.63	59.92
✓	✓	✓	**98.65**	**81.14**	**63.28**

confirming that all three rewards complement each other. See Supp. Section B.7 for results on additional datasets. We then evaluate the **impact of synchronizer architecture** by comparing an autoencoder [512, 256, 128, 256, 512] with transformer-based variants using ×1, ×2, or ×4 transformer blocks in both the encoder and decoder. Table 13 shows that transformer-based synchronizers significantly outperform the fully connected design, improving accuracy by up to +13.93%, while increasing depth beyond two blocks offers little improvement but higher computational cost. Therefore, we select the *Transformer ×2* as an efficient performance–cost trade-off. We evaluate the **impact of different grouping schemes** in GRPO, where effective group formation is essential for training. We first construct groups using image augmentations to train f_{v2t}. After the initial stage, f_{v2t} is further optimized using groups formed based on pseudo-class labels. Table 14 shows the performance gains obtained through this re-grouping strategy. In subsequent iterations, the output of each synchronizer is used to refine the other synchronizer, resulting in further improvements. This ablation study highlights the importance of appropriate grouping schemes for effectively optimizing `SyncVL`. See Supp. Section B.7 for additional ablations and Section B.8 for failure case analysis.

5 Conclusion

In this work, we introduced `SyncVL`, a bi-directional vision–language synchronization framework designed to enable fully unsupervised adaptation of dual-encoder VLMs under distribution shift. `SyncVL` simultaneously learns Vision-to-Text (V2T) and Text-to-Vision (T2V) synchronizers, which collaboratively refine one another through reward-driven optimization. This bi-directional strategy allows visual and textual embeddings to be iteratively synchronized, promoting stronger VL consistency and more stable representations in challenging, unseen domains. Central to `SyncVL` is the proposed objective function inspired from group relative policy optimization, which leverages group-based relative rewards to guide both synchronizers toward more coherent alignment objectives. Comprehensive experiments across 21 datasets demonstrate that `SyncVL` achieves substantial performance gains in unsupervised adaptation, OOD generalization, cross-modal retrieval, object detection, and segmentation, all without requiring labeled data or updates to the pre-trained backbone. These results establish `SyncVL` as a powerful and scalable solution for adapting vision–language models to real-world, distribution-shifted environments, and highlight the promise of bi-directional, policy-optimized synchronization as a paradigm for future research.

References

1. Adaloglou, N., Michels, F., Kalisch, H., Kollmann, M.: Exploring the limits of deep image clustering using pretrained models. In: BMVC (2023)
2. Ali, E., Silva, S., Arora, C., Khan, M.H.: Towards fine-grained adaptation of clip via a self-trained alignment score. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5875–5885 (2026)
3. Ali, E., Silva, S., Khan, M.H.: Dpa: Dual prototypes alignment for unsupervised adaptation of vision-language models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6083–6093. IEEE (2025)
4. An, B., Zhu, S., Panaitescu-Liess, M.A., Mummadi, C.K., Huang, F.: Perceptionclip: Visual classification by inferring and conditioning on contexts. In: ICLR (2024)
5. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2020)
6. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)
7. Berg, T., Liu, J., Woo Lee, S., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2011–2018 (2014)
8. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 446–461. Springer International Publishing, Cham (2014)
9. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)
10. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. CoRR **abs/1311.3618** (2013), `http://arxiv.org/abs/1311.3618`
11. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), `https://openreview.net/forum?id=vvoWPYqZJA`
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). `https://doi.org/10.1109/CVPR.2009.5206848`
13. Duan, J., Yuan, W., Pumacay, W., Wang, Y.R., Ehsani, K., Fox, D., Krishna, R.: Manipulate-anything: Automating real-world robots using vision-language models. In: Conference on Robot Learning. pp. 5326–5350. PMLR (2025)
14. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
15. Fahim, A., Murphy, A., Fyshe, A.: It's not a modality gap: Characterizing and addressing the contrastive gap. arXiv preprint arXiv:2405.18570 (2024)

16. Fang, K., Song, J., Gao, L., Zeng, P., Cheng, Z.Q., Li, X., Shen, H.T.: Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17292–17301 (2024)
17. Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. Advances in neural information processing systems **34**, 7068–7081 (2021)
18. Gadetsky, A., Brbic, M.: The pursuit of human labeling: a new perspective on unsupervised learning. Advances in Neural Information Processing Systems **36**, 60527–60546 (2023)
19. Gadetsky, A., Jiang, Y., Brbic, M.: Let go of your labels with unsupervised transfer. In: International Conference on Machine Learning. pp. 14382–14407. PMLR (2024)
20. Gampa, P., Valsangkar, A.A., Choubey, S., A., P.: Prioritised moderation for online advertising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2004–2012 (June 2023)
21. Gandelsman, Y., Sun, Y., Chen, X., Efros, A.: Test-time training with masked autoencoders. Advances in Neural Information Processing Systems **35**, 29374–29385 (2022)
22. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
23. 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: CVPR (2023)
24. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025)
25. Guo, W., Zhang, L., Zhang, K., Liu, Y., Mao, Z.: Visual-linguistic dependency encoding for image-text retrieval. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 17384–17396 (2024)
26. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. CoRR **abs/1709.00029** (2017), `http://arxiv.org/abs/1709.00029`
27. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. ICCV (2021)
28. Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., Steinhardt, J.: Pixmix: Dreamlike pictures comprehensively improve safety measures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16783–16792 (2022)
29. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In: The 2013 international joint conference on neural networks (IJCNN). pp. 1–8. Ieee (2013)
30. Hu, X., Zhang, K., Xia, L., Chen, A., Luo, J., Sun, Y., Wang, K., Qiao, N., Zeng, X., Sun, M., et al.: Reclip: Refine contrastive language image pre-training with source free domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2994–3003 (2024)

31. Huang, T., Chu, J., Wei, F.: Unsupervised prompt learning for vision-language models. arXiv preprint arXiv:2204.03649 (2022)
32. Huang, X., Zhang, Y., Ma, J., Tian, W., Feng, R., Zhang, Y., Li, Y., Guo, Y., Zhang, L.: Tag2text: Guiding vision-language model via image tagging. arXiv preprint arXiv:2303.05657 (2023)
33. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. In: International Conference on Learning Representations. vol. 2024, pp. 16926–16941 (2024)
34. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
35. Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15190–15200 (2023)
36. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. Rep. TR-2009, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada (2009)
37. Kuang, J., Shen, Y., Xie, J., Luo, H., Xu, Z., Li, R., Li, Y., Cheng, X., Lin, X., Han, Y.: Natural language understanding and inference with mllm in visual question answering: A survey. ACM Computing Surveys **57**(8), 1–36 (2025)
38. Li, J., Savarese, S., Hoi, S.: Masked unsupervised self-training for label-free image classification. In: International Conference on Learning Representations (2023)
39. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International conference on machine learning. pp. 6028–6039. PMLR (2020)
40. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems **35**, 17612–17625 (2022)
41. 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. pp. 740–755. Springer (2014)
42. Lin, Z., Chen, X., Pathak, D., Zhang, P., Ramanan, D.: Revisiting the role of language priors in vision-language models. In: Proceedings of the 41st International Conference on Machine Learning. pp. 29914–29934 (2024)
43. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
44. Liu, Y., Wang, Y., Sun, L., Yu, P.: Rec-gpt4v: Multimodal recommendation with large vision-language models. arXiv preprint arXiv:2402.08670 (2024), `https://arxiv.org/abs/2402.08670`
45. Liu, Y., Kothari, P., Van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: Ttt++: When does self-supervised test-time training fail or thrive? Advances in Neural Information Processing Systems **34**, 21808–21820 (2021)
46. Ma, X., Zhang, J., Guo, S., Xu, W.: Swapprompt: Test-time prompt adaptation for vision-language models. Advances in Neural Information Processing Systems **36**, 65252–65264 (2023)
47. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
48. Marrium, M., Mahmood, A., Khan, M.H., Shakeel, M.S., Kang, W.: Diffusion-guided graph data augmentation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)

49. Marrium, M., Mahmood, A., Khan, M.H., Shakeel, M.S., Kang, W.: Enhancing gnn learning with node augmentation. Neural Networks p. 109228 (2026)
50. Mirza, M.J., Karlinsky, L., Lin, W., Possegger, H., Kozinski, M., Feris, R., Bischof, H.: Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. Advances in Neural Information Processing Systems **36** (2024)
51. Nayak, N.V., Yu, P., Bach, S.: Learning to compose soft prompts for compositional zero-shot learning. In: The Eleventh International Conference on Learning Representations (2023)
52. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics and image processing. pp. 722–729. IEEE (2008)
53. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
54. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (2012)
55. Qian, Q., Xu, Y., Hu, J.: Intra-modal proxy learning for zero-shot visual categorization with clip. Advances in Neural Information Processing Systems **36**, 25461–25474 (2023)
56. Qin, X., Zhang, P., Yang, J.J.O., Zeng, G., Li, Y., Wang, Y., Zhang, W., Dai, P.: Clip is almost all you need: Towards parameter-efficient scene text retrieval without ocr. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24873–24883 (2025)
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
58. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18082–18091 (2022)
59. Saha, O., Lawrence, L., Van Horn, G., Maji, S.: Generate, transduct, adapt: Iterative transduction with vlms. In: International Conference on Computer Vision (ICCV) (2025)
60. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
61. Shi, C., Yang, S.: Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2932–2941 (2023)
62. Shi, Z., Lipani, A.: Dept: Decomposed prompt tuning for parameter-efficient fine-tuning. In: The Twelfth International Conference on Learning Representations (2024)
63. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems **35**, 14274–14289 (2022)
64. Singh, A., Marjit, S., Lin, W., Gavrikov, P., Yeung-Levy, S., Kuehne, H., Feris, R., Doveh, S., Glass, J., Mirza, M.J.: Ttrv: Test-time reinforcement learning for vision language models. arXiv preprint arXiv:2510.06783 (2025)

65. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
66. Song, L., Xue, R., Wang, H., Sun, H., Ge, Y., Shan, Y., et al.: Meta-adapter: An online few-shot learner for vision-language model. Advances in Neural Information Processing Systems **36**, 55361–55374 (2023)
67. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR **abs/1212.0402** (2012), `http://arxiv.org/abs/1212.0402`
68. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2023)
69. Sun, X., Hu, P., Saenko, K.: Dualcoop: Fast adaptation to multi-label recognition with limited annotations. Advances in Neural Information Processing Systems **35**, 30569–30582 (2022)
70. Tang, Z., Lian, L., Eisape, S., Wang, X., Herzig, R., Yala, A., Suhr, A., Darrell, T., Chan, D.M.: Tulip: Towards unified language-image pretraining. arXiv preprint arXiv:2503.15485 (2025)
71. Tanwisuth, K., Zhang, S., Zheng, H., He, P., Zhou, M.: Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In: International conference on machine learning. pp. 33816–33832. PMLR (2023)
72. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
73. Udandarao, V., Gupta, A., Albanie, S.: Sus-x: Training-free name-only transfer of vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2725–2736 (2023)
74. Vendrow, E., Pantazis, O., Shepard, A., Brostow, G., Jones, K., Mac Aodha, O., Beery, S., Van Horn, G.: Inquire: A natural world text-to-image retrieval benchmark. Advances in Neural Information Processing Systems **37**, 126500–126514 (2024)
75. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (ICLR) (2021), `https://openreview.net/forum?id=uXl3bZLkr3c`
76. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Advances in Neural Information Processing Systems. pp. 10506–10518 (2019)
77. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International conference on computer vision. pp. 1457–1464. IEEE (2011)
78. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)
79. Wang, R., Zheng, H., Duan, X., Liu, J., Lu, Y., Wang, T., Xu, S., Zhang, B.: Few-shot learning with visual distribution calibration and cross-modal distribution alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23445–23454 (2023)
80. Wen, L., Wang, Y., Zhang, D., Chen, G.: Visual matching is enough for scene text retrieval. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. pp. 447–455 (2023)

81. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. Advances in neural information processing systems **33**, 6256–6268 (2020)
82. Yang, Y., Zhou, T., Li, K., Tao, D., Li, L., Shen, L., He, X., Jiang, J., Shi, Y.: Embodied multi-modal agent trained by an llm from a parallel textworld. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26275–26285 (2024)
83. Yao, H., Zhang, R., Xu, C.: Tcp: Textual-based class-aware prompt tuning for visual-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23438–23448 (2024)
84. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. NeurIPS **33**, 5824–5836 (2020)
85. Zanella, M., Gérin, B., Ayed, I.: Boosting vision-language models with transduction. Advances in Neural Information Processing Systems **37**, 62223–62256 (2024)
86. Zeng, G., Zhang, Y., Wei, J., Yang, D., Zhang, P., Gao, Y., Qin, X., Zhou, Y.: Focus, distinguish, and prompt: Unleashing clip for efficient and flexible scene text retrieval. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2525–2534 (2024)
87. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
88. Zhang, A., Yu, Y., Li, S., Gao, R., Zhang, L., Gao, S.: Contrastive learning-based personalized tag recommendation. Sensors **24**(18), 6061 (2024)
89. Zhang, J., Wei, Q., Liu, F., Feng, L.: Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. In: International Conference on Machine Learning. pp. 60004–60020. PMLR (2024)
90. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. Advances in neural information processing systems **35**, 38629–38642 (2022)
91. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
92. Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., Li, H.: Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15211–15222 (2023)
93. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European conference on computer vision. pp. 493–510. Springer (2022)
94. Zhang, S., Khan, S., Shen, Z., Naseer, M., Chen, G., Khan, F.S.: Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In: CVPR. pp. 3479–3488 (2023)
95. Zhang, Y., Zhu, W., Tang, H., Ma, Z., Zhou, K., Zhang, L.: Dual memory networks: A versatile adaptation approach for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28718–28728 (2024)
96. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. Advances in Neural Information Processing Systems **36**, 54111–54138 (2023)

97. Zheng, S., Liu, J., Feng, Y., Lu, Z.: Steve-eye: Equipping llm-based embodied agents with visual perception in open worlds. arXiv preprint arXiv:2310.13255 (2023)
98. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16793–16803 (2022)
99. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
100. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
101. Zhou, P., Liu, C., Ren, J., Zhou, X., Xie, Y., Cao, M., Rao, Z., Huang, Y.L., Chong, D., Liu, J., et al.: When large vision language models meet multimodal sequential recommendation: An empirical study. In: Proceedings of the ACM on Web Conference 2025. pp. 275–292 (2025)
102. Zhu, X., Zhang, R., He, B., Zhou, A., Wang, D., Zhao, B., Gao, P.: Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2605–2615 (2023)