Omniview-Tuning: Boosting Viewpoint Invariance of Vision-Language Pre-training Models

Shouwei Ruan[1], Yinpeng Dong[2,5], Hanqing Liu[3], Yao Huang[1], Hang Su[2,4], and Xingxing Wei[1,3]*

[1] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
[2] Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center, THBI Lab, Tsinghua-Bosch Joint ML Center, Tsinghua University, Beijing, 100084, China
[3] Hangzhou Innovation Institute, Beihang University, Hangzhou 311228, China
[4] Zhongguancun Laboratory, Beijing, 100080, China
[5] RealAI
{shouweiruan,hqliu,y_huang,xxwei}@buaa.edu.cn, {dongyinpeng,suhangss}@tsinghua.edu.cn

Abstract. Vision-Language Pre-training (VLP) models like CLIP have achieved remarkable success in computer vision and particularly demonstrated superior robustness to distribution shifts of 2D images. However, their robustness under 3D viewpoint variations is still limited, which can hinder the development for real-world applications. This paper successfully addresses this concern while keeping VLPs' original performance by breaking through two primary obstacles: 1) the scarcity of training data and 2) the suboptimal fine-tuning paradigms. To combat data scarcity, we build the Multi-View Caption (MVCap) dataset — a comprehensive collection of over four million multi-view image-text pairs across more than 100K objects, providing more potential for VLP models to develop generalizable viewpoint-invariant representations. To address the limitations of existing paradigms in performance trade-offs and training efficiency, we design a novel fine-tuning framework named Omniview-Tuning (OVT). Specifically, OVT introduces a Cross-Viewpoint Alignment objective through a minimax-like optimization strategy, which effectively aligns representations of identical objects from diverse viewpoints without causing overfitting. Additionally, OVT fine-tunes VLP models in a parameter-efficient manner, leading to minimal computational cost. Extensive experiments on various VLP models with different architectures validate that OVT significantly improves the models' resilience to viewpoint shifts and keeps the original performance, establishing a pioneering standard for boosting the viewpoint invariance of VLP models. The code and dataset are available via https://github.com/Heathcliff-saku/Omniview_Tuning

Keywords: Vision-Language Pre-training · Viewpoint Invariance

[*] Corresponding author.

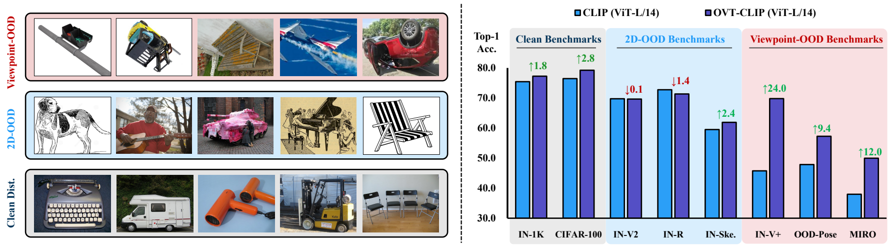


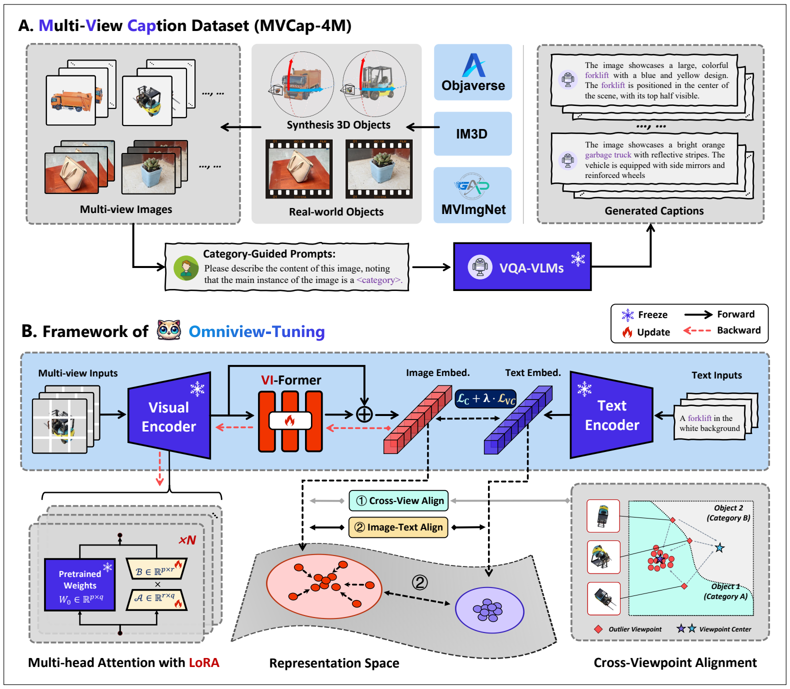
Fig. 1: The Challenge of Viewpoint Invariance in VLP. We selected benchmarks representing clean distributions (ImageNet-1K [14], CIFAR-100 [27]), common 2D-OOD (ImageNet-V2 [42], ImageNet-R(endition) [21], ImageNet-Sketch [55]), and viewpoint-OOD (ImageNet-V(iewpoint)+ [46], OOD-CV(Pose) [61], MIRO [7]). We display samples from these data distributions (*left*) and report the Top-1 accuracy of the original CLIP (ViT-L/14) and our improved OVT-CLIP (ViT-L/14) (*right*).

1 Introduction

Vision-Language Pre-training (VLP) models, such as CLIP [40] and BLIP [29], have shown great promise in learning transferable representations across various tasks. By aligning images and texts in a joint embedding space with a large corpus of paired image-text data, VLP models exhibit exceptional representation and generalization capabilities that surpass traditional task-specific models. Owing to this, the VLP models serve as foundation models for numerous tasks, including visual recognition [40], visual question answering [1,33], and image generation [41,47]. Moreover, these models can effectively integrate real-world visual inputs with humankind instructions, leading to their increasing use in physical-world applications, like autonomous driving [62] and embodied robotics [28, 57].

Besides their expressive power, VLP models have also shown excellent robustness under out-of-distribution (OOD) data [17, 40, 54, 60], including common corruptions [6, 15, 22], stylistic changes [21, 55], and natural distribution shifts [21, 23, 42]. However, a recent study [45] identifies that although VLP models excel at handling 2D-OOD samples, they suffer significant performance degradation under *3D viewpoint changes*, revealing a notable shortcoming of the existing VLP models. As demonstrated in Fig. 1, when dealing with the benchmarks concerned with 3D viewpoint shifts [16, 46, 61], CLIP's performance is obviously lower than that on 2D-OOD benchmarks. This large gap likely stems from limited coverage of diverse viewpoints in the training datasets [18, 49, 53], which is crucial for learning viewpoint-invariant representations. As VLP models are increasingly deployed in real-world environments where viewpoint shifts often occur, enhancing their resilience to such changes is urgent and essential.

To address this problem, this paper sets out to *enhance the viewpoint invariance of VLP models while preserving the original performance as much as possible*. However, achieving this goal meets the following challenges: (1) *Data scarcity*: acquiring VLP training data that covers a wide range of viewpoint variations is particularly challenging compared to conventional image-text pair data. Although some datasets introduced for task-specific models do include viewpoint



The image showcases a large, colorful forklift with a blue and yellow design. The forklift is positioned in the center of the scene, with its top half visible.

The image showcases a bright orange garbage truck with reflective stripes. The vehicle is equipped with side mirrors and reinforced wheels

Please describe the content of this image, noting that the main instance of the image is a <category>.

A forklift in the white background

Fig. 2: Method Overview. (A) We create the first multi-view image caption dataset by collecting multi-view samples from existing 3D object and video datasets, and generating category-guided descriptions using VLLMs. **(B)** The proposed Omniview-Tuning takes multi-view image caption data as input, employs the cross-view alignment objective to encourage the model to learn viewpoint-invariant representations, and achieves efficient fine-tuning by updating VIformer and LoRA parameters.

variations [4, 23, 35, 61], they often lack the textual descriptions vital for VLP. Even the largest available multi-view datasets [10, 46, 59] fall short in terms of scale, category coverage, and viewpoint diversity, thereby limiting the potential for VLP models to develop generalizable viewpoint-invariant representations. (2) *Inappropriate paradigms*: traditional approaches, which often regard viewpoint changes as adversarial attacks and employ adversarial training paradigms for enhancing invariance [2, 45, 46], are not entirely suitable for VLP models. Such frameworks typically entail a trade-off between robustness and accuracy—a balance that requires more careful consideration for foundation VLP models, where our aim is not solely to improve viewpoint invariance but, more importantly, to bridge the gap between it and the original performance. Furthermore, these approaches necessitate extra 3D reconstruction and neural rendering to capture adversarial viewpoints, leading to prohibitive computational costs for large-scale VLP models. For instance, tuning ResNet-50 with VIAT [46] under a dataset of just 1K objects demands around 400 GPU hours. Therefore, it is important to make training more efficient and less resource-intensive.

Based on the above discussions, this paper conducts a pioneering exploration of the viewpoint invariance of VLP models. Specifically, we address the aforementioned challenges by making the following contributions:

Million-scale multi-view image-text training set. We introduce a large-scale **M**ulti-**V**iew **Cap**tion (**MVCap**) dataset tailored for viewpoint invariance of VLP models, comprising over 4.6 million multi-view image-text pairs across more than 100K objects. To assemble a diverse collection of multi-view image-text pairs, we amalgamate various 3D assets with real-world multi-view data. This process involves an extensive selection and rendering of multi-view images from existing datasets. We then utilize a Vision Large Language Model (VLLM) for automated caption generation to obtain semantically rich textual descriptions without extensive manual efforts. To ensure category consistency across varying viewpoints in the generated captions, we implement a category-guided prompting strategy, which maintains accuracy in textual descriptions for different viewpoints of the same object or scene (details in Sec. 3).

Effective framework for enhancing VLP's viewpoint invariance. We propose **Omniview-Tuning (OVT)**, a novel framework designed to enhance the viewpoint invariance of prevalent VLP models. As illustrated in Fig. 2, OVT employs multi-view image-text pairs for training additional learnable components. To amplify the model's proficiency in learning viewpoint-invariant representations, we introduce a **Cross-viewpoint Alignment** objective, ensuring that representations of the same object from different viewpoints are close and unified in the high-dimensional feature space. To prevent performance trade-offs due to the concept drift from aggressive viewpoint alignment, we innovatively construct the optimization paradigm of OVT in a **minimax-like form**. The optimization process includes identifying extreme outlier viewpoints during the maximization step, while optimizing the model's invariant representation for these outlier samples in the minimization step. This strategy enables the model to focus more on the worst-case viewpoint samples, thereby maximally preserving the original embedding distribution and avoiding performance degradation while saving computational costs. Moreover, OVT is designed in a **Parameter-Efficient** Fine-Tuning manner to improve efficiency, and creatively incorporates two trainable parameter modules: an embedding transformation module named VIFormer and the Low-Rank Adaptation (LoRA [25]) weights, to acquire additional viewpoint invariance capabilities efficiently.

Extensive experiments across various VLP architectures and tasks. We conduct extensive experiments to show the efficacy of the OVT framework in improving the viewpoint invariance for VLP models while maintaining performance on clean data and 2D-OOD samples. For example, by fine-tuning CLIP with OVT on different architectures (ViT-B/32, ViT-B/16, and ViT-L/14), the Top-1 accuracy on viewpoint-OOD benchmarks increased by an average of **9.6%**, **10.2%**, and **8.9%**, respectively, with only a minimal sacrifice on 2D-OOD benchmarks by an average of 2.6%, 1.4%, and 0.2%. Furthermore, serving as the visual encoder in VLLMs (*e.g.*, LLaVa [33]), OVT-CLIP also effectively improves viewpoint invariance in image captioning and visual question answering tasks.

2 Related Work

2.1 Viewpoint Invariance and Robustness

Viewpoint invariance is a key property of human vision [5] but is usually lacking in computer vision models [2, 16]. Addressing viewpoint invariance and robustness involves strategies like data augmentation and adversarial learning. Early efforts aim to enhance viewpoint robustness by incorporating datasets enriched with viewpoint variations [4, 23, 35, 61]. For example, Madan *et al.* encourage models to learn viewpoint-robust representations by incorporating object-pose combinations [35]. However, these methods often falter under malicious viewpoint perturbations due to their inability to capture the worst-case viewpoint samples. Recently, achieving viewpoint invariance within the adversarial training paradigm has shown promise [2,16,20,46]. By treating viewpoint variations as an adversarial attack, Alcorn *et al.* employ a differentiable renderer to train models against adversarial viewpoints optimized from a limited 3D objects set [2]. Recent studies, such as Viewfool [16] and VIAT [45, 46], have introduced neural radiance field (NeRF) [38,39], enabling the characterization of adversarial viewpoint distributions from 2D multi-view inputs. Besides, studies in the self-supervised domain employ a label-free paradigm to improve viewpoint invariance, but these methods often require the introduction of complex structures, such as large-scale graph structures [56] and viewpoint generators [12]. Distinct from previous studies, our work pioneers the improvement of viewpoint invariance representation within large-scale VLP models, which is facilitated through suitable training data and refined fine-tuning methodologies.

2.2 Vision-Language Pre-training

In the realm of VLP, significant strides have been made in understanding and bridging the semantic gap between visual and textual information. Despite the variety of existing VLP paradigms, such as single-stream encoder (*e.g.*, VisualBERT [31] and UNITER [8], *etc.*) or dual-stream encoder equipped with diverse training objectives, the dual-stream contrastive learning architecture exemplified by ALIGN [30] and OpenAI's CLIP [40] dominates the field. CLIP, in particular, has gained widespread attention for its ability to perform zero-shot classification tasks by adopting a vast corpus of internet-collected image-text pairs, demonstrating the power of large-scale contrastive pre-training. Thus, Our investigation primarily focuses on these VLP architectures. Building upon these foundational works, subsequent iterations like open-CLIP [26], EVA-CLIP [51, 52], and Meta-CLIP [58] have introduced nuanced enhancements. These refinements, ranging from the incorporation of more expansive high-quality image-text datasets and improved training methodologies, have collectively contributed to performance uplifts. BLIP [29], meanwhile, introduces a bootstrapping mechanism by the proposed captioner and filter module that achieve significant performance improvements on various downstream tasks.

Table 1: Comparison of current large-scale multi-view datasets. *"Spherical"* indicates whether the viewpoints cover spherical space, *"Diversity"* assesses the diversity of viewpoints, and *"Caption"* indicates whether textual descriptions are provided.

Dataset	Year	#Obj.	#Cat.	#Avg. View.	#Sample	Image Domain	Spherical	Diversity	Caption
OOWL [24]	2019	500	25	240	120K	Real	✗	★★	✗
CO3D [43]	2021	18.6k	50	$\sim$80	1.5M	Real	✗	★	✗
ABO [10]	2022	7.9k	63	30	238K	Synthetic	✗	★	✗
IM3D [46]	2023	1.0k	100	100	100K	Synthetic	✓	★★★	✗
MVImgNet [59]	2023	219.0k	238	$\sim$30	6.5M	Real	✗	★★	✗
MVCap (Ours)	2024	94.6k	1600	100/$\sim$30	4.6M	Synthetic+Real	✓	★★★	✓

3 Multi-view Caption Dataset

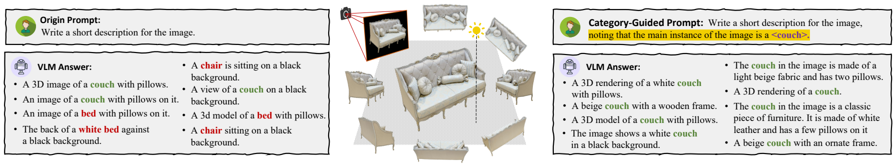
We recognize that one of the key challenges in achieving viewpoint invariance for VLP is the scarcity of training data that offer comprehensive viewpoints. As shown in Tab. 1, existing large-scale multi-view datasets [10, 24, 43, 46, 59] typically lack in either sample diversity, category breadth, or textual descriptions, limiting their effectiveness for supporting VLP models to achieve viewpoint invariance. To address these limitations, we next introduce the MVCap dataset.

3.1 Multi-View Image Collection

We commence by gathering a multi-view image collection $\mathcal{D} = \{I_{ij} \mid i = 1, 2, ..., N; j = 1, 2, ..., M_i\}$, where N and M_i represent the counts of objects and their viewpoints, respectively. To cover various categories from virtual to real-world scenes, we integrate samples from Objaverse [13], IM3D [46], and MVImgNet [59]. Since the original 3D dataset includes a fair share of noisy and semantically indistinct objects, we leverage semantic embeddings provided by OpenShape [34] to conduct cosine similarity sorting based on the embeddings of customized labels. Finally, we filter 24,495 virtual 3D objects endowed with distinct semantic clarity and cover over 1,600 categories. For each chosen 3D object, we employ Blender to render 100 random viewpoint images from the upper hemisphere, ensuring a comprehensive and varied viewpoint representation in our collected samples. We also incorporate objects from MVImgNet with over 30 valid viewpoints (video frames), thereby acquiring a substantial number of real-world multi-view samples to enrich the dataset's content and quality further.

3.2 Category-Guided Caption Generation

The granularity and precision of textual descriptions are pivotal in VLP training, as they influence the model's generalization capabilities and the variety of visual concepts learned. Relying solely on simple prompt engineering, such as "*a photo of [category]*," may introduce biases and limit the model's generalizability, whereas manual annotation is costly. To circumvent this, we utilize InstructBLIP-flant5xl [11], a leading VLLM, to create multi-view captions automatically. However, such VLLMs also grapple with viewpoint invariance, where the model's responses to different viewpoints can often be category-inconsistent,



Origin Prompt:
Write a short description for the image.

VLM Answer:

- A 3D image of a **couch** with pillows.
- An image of a **couch** with pillows on it.
- An image of a **bed** with pillows on it.
- The back of a **white bed** against a black background.
- A **chair** is sitting on a black background.
- A view of a **couch** on a black background.
- A 3d model of a **bed** with pillows.
- A **chair** sitting on a black background.

Category-Guided Prompt: Write a short description for the image, noting that the main instance of the image is a **<couch>.**

VLM Answer:

- A 3D rendering of a white **couch** with pillows.
- A beige **couch** with a wooden frame.
- A 3D model of a **couch** with pillows.
- The image shows a white **couch** in a black background.
- The **couch** in the image is made of a light beige fabric and has two pillows.
- A 3D rendering of a **couch**.
- The **couch** in the image is a classic piece of furniture. It is made of white leather and has a few pillows on it
- A beige **couch** with an ornate frame.

Fig. 3: Generated multi-view captions with common and category-guided prompts.

as depicted in Fig. 3. This situation presents a "chicken or egg" dilemma: we hope to use a viewpoint-invariant model to supply data for viewpoint invariance training. We address this by the design of category-guided prompting. Specifically, we use prompts containing ground-truth category information to eliminate the hallucination of large VLLMs in response to viewpoint-shifted inputs, thereby generating category-consistent multi-view captions. Formally, the forward process for generating captions can be represented as follows:

$$\begin{aligned} T_{ij} &= \mathcal{G}[I_{ij}, \mathrm{Prompt}(c_i)]; \\ \mathrm{Prompt}(c_i) &= \text{"}\textit{Write a short description for the image,} \\ &\textit{noting that the main instance of the image is a} < c_i > .\text{"}, \end{aligned} \tag{1}$$

where $c_i \in C$ denotes the category label for the i-th object, and $\mathcal{G}$ denotes the forward process of InstructBLIP. This yields the multi-view image-text pairs $\tilde{\mathcal{D}} = \{\langle I_{ij}, T_{ij}\rangle \mid i = 1, 2, ..., N; j = 1, 2, ..., M_i\}$, which can be utilized for the viewpoint invariance fine-tuning of VLP models.

4 Omniview-Tuning

4.1 Preliminaries: Contrastive Vision-Language Pre-training

Despite the variety of existing VLP paradigms, such as single-stream or dual-stream architectures equipped with diverse training objectives, the dual-stream contrastive learning architecture exemplified by CLIP *et al.* [30, 40] dominates the field. Thus, Our investigation primarily focuses on this VLP architecture.

Without the loss of generality, these VLP models are composed of a visual encoder $E_{\mathbf{W_v}} : I \to z^I \in \mathbb{R}^d$ and a text encoder $E_{\mathbf{W_t}} : T \to z^T \in \mathbb{R}^d$, which maps visual and textual inputs to a unified high-dimensional feature space $\mathbb{R}^d$, respectively, where $\mathbf{W_v}$ and $\mathbf{W_t}$ are weight matrices of two encoders. Given a large corpus of image-text pairs $\{\langle I_i, T_i\rangle\}_{i=1}^N$, VLP models typically employ an image-text contrastive (ITC) loss as the training objective:

$$\mathcal{L}_{ITC} = \tfrac{1}{2}(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}), \tag{2}$$

which is composed of an image-to-text and a text-to-image terms formulated as:

$$\begin{aligned} \mathcal{L}_{I\to T} &= -\tfrac{1}{N}\textstyle\sum_{i=1}^N \log \frac{\exp(d(z_i^I, z_i^T)/\tau)}{\sum_{k=1}^N \exp(d(z_i^I, z_k^T)/\tau)}, \\ \mathcal{L}_{T\to I} &= -\tfrac{1}{N}\textstyle\sum_{i=1}^N \log \frac{\exp(d(z_i^T, z_i^I)/\tau)}{\sum_{k=1}^N \exp(d(z_i^T, z_k^I)/\tau)}, \end{aligned} \tag{3}$$

where τ represents a learnable temperature parameter, z^I and z^T denote the image and text embeddings, respectively. The $\mathcal{L}_{ITC}$ maximizes the similarity between matched image-text pairs while minimizing the similarity for mismatched pairs, thus enabling the alignment of visual and textual information to the same feature space, bringing the embeddings of matched pairs closer. Following [29, 30, 40], the proposed Omniview-Tuning implements $\mathcal{L}_{ITC}$ for aligning multi-view images with text modalities, which is explained in the next section.

4.2 Problem Formulation

Viewpoint Invariance of Vision-Language Pre-training. In computer vision scenario, viewpoint invariance implies that model $f(\cdot)$ can provide consistent predictions or representations given any different views of the identical object or scene [46]. Formally, given a collection of multi-view images $\mathcal{D} = \{I_{ij} \mid i = 1, 2, ..., N; j = 1, 2, ..., M_i\}$, viewpoint invariance is required:

$$f(I_{ij}) = f(I_{ij'}), \quad \forall i, j, j' \text{ with } j \neq j', \tag{4}$$

where i is the index of the object/scene, j and j' are indexes of two viewpoint samples. However, in the context of dual-stream VLP models, this concept requires a more refined interpretation. For VLP models, viewpoint invariance necessitates that the visual representations (*i.e.*, the embeddings inferred from the visual encoder) from different viewpoints be sufficiently close in the feature space. Assuming I_{ij} and $I_{ij'}$ are images from different viewpoints of the same object, this requirement can be formulated as follows:

$$d\Big[E_{\mathbf{W_v}}(I_{ij}), E_{\mathbf{W_v}}(I_{ij'})\Big] \leq \epsilon, \tag{5}$$

where $d(\cdot)$ denotes a distance metric in the representation space, such as cosine distance, ϵ represents the maximum variance allowed.

Optimization Objectives of Omniview-Tuning. Although images from different viewpoints often correspond to slightly varying textual descriptions, influenced by context, grammatical structure, and linguistic ambiguity, this variation could be significantly amplified in the high-dimensional representation space [9, 44]. Therefore, relying solely on image-text alignment may not suffice to adequately align embeddings from different viewpoints. Starting from the definition of viewpoint invariance, we introduce a cross-viewpoint alignment objective within the $\mathcal{L}_{ITC}$ to directly encourage the model to learn invariant representations between different viewpoints, rather than relying on the indirect alignment through textual descriptions. This can be seen as a regularization that forces the model to obtain viewpoint invariance, even when such invariance is not explicitly articulated in the textual descriptions. With this consideration, given a multi-view training set $\mathcal{D}$, the optimization problem is defined as follows:

$$\min_{\mathbf{W_v}, \mathbf{W_t}} \big[\mathcal{L}_{ITC} + \lambda \cdot \underbrace{\textstyle\sum_i \sum_{j \neq j'} d(z^I_{ij}, z^I_{ij'})}_{\mathcal{L}_{VC}}\big], \tag{6}$$

where the first term represents the image-text alignment used in the pre-training process, while the second term signifies the cross-viewpoint alignment goal mentioned above, referred to as Viewpoint Consistency loss ($\mathcal{L}_{VC}$), which aims to minimize the cosine distance between embeddings from different viewpoints. λ is a hyperparameter that balances the importance of two loss terms.

4.3 Optimization Strategy

In summary, the naive way to achieve viewpoint invariance is to calculate the loss terms in Eq. (6) based on the forward process of encoders, then update the encoders' weight using gradient descent. However, it has a relatively high time complexity to solve Eq. (6) because current $\mathcal{L}_{VC}$ requires iterating over every possible combination of viewpoints. Therefore, we endeavor to provide a more effective implementation for the original optimization problem. Drawing from the advantages of adversarial training [36, 46], we frame the optimization of the $\mathcal{L}_{VC}$ in a minimax format, rewriting the original problem Eq. (6) as:

$$\min_{\mathbf{W_v},\mathbf{W_t}} \left[\mathcal{L}_{ITC} + \lambda \cdot \underbrace{\max_{\mathcal{O}=\{O_i\}_{i=1}^{N},|O_i|=K} \textstyle\sum_{i=1}^{N} \sum_{j\in\mathcal{O}} l(z^I_{ij}, z^I_{C_i})}_{\mathcal{L}_{VC}} \right], \quad \text{where} \;\; l(z^I_{ij}, z^I_{C_i}) = \max\left[d(z^I_{ij}, z^I_{C_i}) + m, 0\right], \tag{7}$$

where $\mathcal{O} = \{O_i\}_{i=1}^{N}$ is the outlier viewpoints set, $z^I_{C_i}$ are anchor viewpoint embeddings of each object, and $l(\cdot)$ is the cosine distance with a margin m. During the optimization, The maximization step first identifies the collection of top-K outlier viewpoints $\mathcal{O}$, which are the viewpoint samples with the highest degree of representational deviation. Then, the minimization step encourages the outlier viewpoint embeddings to converge towards corresponding anchor viewpoint embeddings $z^I_{C_i}$. We obtain $z^I_{C_i}$ by calculating the **nearest-neighbor weighted** embedding centroid of each object:

$$z^I_{C_i} = \textstyle\sum_{j=1}^{M_i} \tilde{\omega}_{ij} \cdot z^I_{ij}, \quad \text{where} \;\; \tilde{\omega}_{ij} = \omega_{ij} / \sum_j \omega_{ij}, \;\; \omega_{ij} = 1 / \sum_{z^I_{ih} \in \mathcal{Q}_{ij}} d(z^I_{ij}, z^I_{ih}) \tag{8}$$

where $\mathcal{Q}_{ij} = \{z^I_{ih}\}_{h=1}^{5}$ is the top-5 nearest neighbours of each viewpoint embedding z^I_{ij}. As for the outlier viewpoints set, we define them as the viewpoints with the top-K farthest cosine distances from $z^I_{C_i}$.

The adoption of this strategy offers dual advantages: *Firstly*, it allows the model to focus solely on extreme outlier viewpoints, preventing concept drift and potential overfitting to the fine-tuning dataset that results from excessive alignment. *Secondly*, this approach reduces computational overhead and significantly enhances optimization efficiency.

4.4 Parameter-Efficient Modules

To mitigate the impact of full parameters update on the original performance and enhance training efficiency, we achieve viewpoint invariance by efficiently

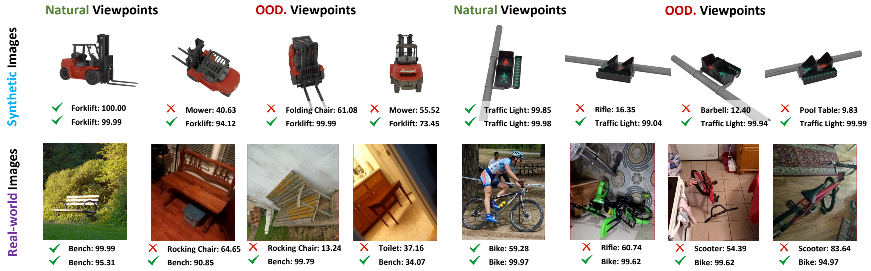


Fig. 4: Visualization results for zero-shot classification results.

fine-tuning the parameters of the visual encoder while keeping the text encoder frozen. Inspired by LoRA [25], we perform low-rank decomposition on the weights of the visual encoder $\mathbf{W_v} \in \mathbb{R}^{n \times m}$ to substitute full-parameter update:

$$\tilde{\mathbf{W_v}} = \mathbf{W_v} + \mathit{\Delta} W = \mathbf{W_v} + \mathbf{BA}, \quad \text{where } \mathbf{B} \in \mathbb{R}^{m \times r}, \mathbf{A} \in \mathbb{R}^{r \times n}, r \ll \min(n, m), \tag{9}$$

where $\mathbf{A}$ and $\mathbf{B}$ are two learnable low-rank parameter matrices, which we apply to the self-attention layers of the visual encoder and update them during fine-tuning while freezing the original pre-trained weights. Compared to directly adjusting the original network weights (i.e., full-parameter fine-tuning manner in TeCoA [37] and FARE [48]), this approach enables us to improve the model's viewpoint invariance representation capability while better preserving the original performance. Drawing inspiration from the success of CLIP-Adapter [19], which improves CLIP's performance in few-shot scenarios by introducing linear layers after the encoder, we propose a similar module called VIformer:$f_{\boldsymbol{\theta}} : z^I \in \mathbb{R}^d \to s^I \in \mathbb{R}^d$, where $\boldsymbol{\theta}$ is the weight. Unlike CLIP-Adapter, VIformer transforms the original embeddings z^I by introducing self-attention layers in a learnable manner to extract and retain specific viewpoint-invariant key components s^I. Combining LoRA and VIformer modules, the forward process of image encoding can be represented as follows:

$$\begin{aligned} \tilde{z}^I &= \alpha \cdot f_{\boldsymbol{\theta}}(z^I) + (1 - \alpha) \cdot z^I \\ &= \alpha \cdot f_{\boldsymbol{\theta}}(\mathbf{W_v} \cdot I + \mathbf{BA} \cdot I) + (1 - \alpha) \cdot (\mathbf{W_v} \cdot I + \mathbf{BA} \cdot I), \end{aligned} \tag{10}$$

where the constant value α denotes the residual ratio to balance achieving original performance and viewpoint invariance performance. Therefore, for Eq. (6), we now only need to update $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\theta}$, rather than the entire weights of VLP. To facilitate the understanding of the OVT training process, we provide the pseudocode for OVT as shown in Appendix D

5 Experiments

Our evaluation of Omniview-Tuning spans several downstream tasks, including zero-shot classification, image captioning, and vision question answering.

Table 2: Configurations of OVT and zero-shot Top-1 accuracy (%) on ImageNet-1K with ImageNet-V+. The number in parentheses shows the performance change relative to the pre-trained weights. Through OVT training, each model maintains the performance on ImageNet-1K (IN-1K) while significantly improving the performance on ImageNet-V+ (IN-V+.), narrowing the performance gap.

Model	Pretrain Weight	Pretrain Data	Total #Param.	Trainable #Param.	Image Size	Batch Size	#Iter.	IN-1K	IN-View+
OVT-OpenCLIP ViT-B/32	OpenCLIP-ViT-B/32	LAION (2B)	151M	6.6M	224	512	35k	67.8 (↑1.3)	59.5 (↑22.4)
OVT-OpenCLIP ViT-B/16	OpenCLIP-ViT-B/16	LAION (2B)	149M	6.6M	224	512	35k	69.7 (↑2.1)	61.7 (↑17.5)
OVT-OpenCLIP ViT-L/14	OpenCLIP-ViT-L/14	LAION (2B)	428M	11.8M	224	256	20k	77.3 (↑ 2.1)	69.8 (↑16.6)
OVT-MetaCLIP ViT-B/32	MetaCLIP-ViT-B/32	Common Crawl (2.5B)	151M	6.6M	224	512	40k	69.7 (↑2.1)	54.8 (↑13.8)
OVT-MetaCLIP ViT-B/16	MetaCLIP-ViT-B/16	Common Crawl (2.5B)	149M	6.6M	224	512	40k	73.8 (↑1.7)	64.8 (↑15.2)
OVT-MetaCLIP ViT-L/14	MetaCLIP-ViT-L/14	Common Crawl (2.5B)	428M	11.8M	224	256	20k	77.7 (↓1.4)	75.4 (↑9.0)
OVT-BLIP ViT-B/16	Salesforce-BLIP-ViT-B/16	COCO *et al.* (129M)	234M	4.3M	224	256	20k	61.7 (↑8.8)	54.8 (↑18.0)

For zero-shot classification (Sec. 5.1), we conduct evaluations for CLIP [40] and BLIP [29] architectures. For image captioning and vision question answering (Sec. 5.2), we replace the visual encoders in Vision Large Language Models (VLLMs) with our fine-tuned versions. We adopt LLaVA-1.5 [32,33], and OpenFlamingo [3], the most advanced open-source VLLMs available. Additionally, we present the ablation study and convergence analysis of our approach in Sec. 5.3.

5.1 Evaluation of Zero-Shot Classification

Baselines. We adopt the official CLIP (OpenAI CLIP [40]) and the community open-source version (OpenCLIP [26]) as baselines. Additionally, we include the current state-of-the-art Eva02-CLIP [51] and MetaCLIP [58] as another set of baselines to compare CLIP trained with improved techniques and extensive training data. For BLIP, we use the official implementation [29] as the baseline.
Settings. We train two series of CLIP models using our OVT framework and MVCap dataset, each series comprising three different visual encoder architectures (ViT-B/32, ViT-B/16, and ViT-L/14). OVT-OpenCLIP are fine-tuned on the original weights of OpenCLIP, while OVT-MetaCLIP are based on MetaCLIP. The fine-tuning settings are detailed in Tab. 2. We standardized the λ=1.0, α=0.1, and the outlier viewpoints number K=5 and set the LoRA rank at 8.
Datasets and Metrics. We employ a various set of benchmarks for evaluation, including clean data distributions [14, 27], common 2D-OOD [21, 23, 42, 55, 61], and most importantly, viewpoint-OOD [7, 16, 46, 61]. For each benchmark, we report Top-1 and Top-5 accuracy and average accuracy across all benchmarks. The evaluations follow the standard prompting engineering of CLIP [40].
Results and Discussions. Tab. 3 summarizes the performance of our OVT-trained VLP models against various VLP versions, including their accuracy across different benchmarks and average accuracy across clean, common-OOD, and Viewpoint-OOD domains. We can draw the following conclusions:

(1) OVT significantly enhances the models' invariance to Viewpoint-OOD samples. Across different VLP architectures and visual encoders, OVT-trained models perform best on almost all viewpoint-OOD benchmarks. On the average

Table 3: Top-1/Top-5 zero-shot accuracy (%) under different benchmarks

	Clean				Common-OOD						Viewpoint-OOD					
Model	ImageNet-100 [46]	ImageNet-1K [14]	Cifar-100 [27]	**Avg. Acc.**	ImageNet-V2 [42]	ImageNet-Ske. [55]	ImageNet-OOD. [23]	ImageNet-Ren. [21]	OOD-CV [61]	**Avg. Acc.**	ImageNet-View. [16]	ImageNet-View.+ [46]	OOD-CV-Pose [61]	MIRO [7]	**Avg. Top-1**	**Total Avg. Acc.**
A. Comparisons with ViT-B/32 baselines																
OpenAI CLIP	77.5/93.9	63.3/88.8	64.3/88.1	68.4/90.2	55.8/83.4	42.2/70.3	33.4/62.2	50.7/75.4	50.2/82.6	46.5/74.8	44.5/65.4	27.5/52.4	47.2/84.5	26.5/59.4	36.4/65.4	48.6/75.5
Open CLIP	**81.1**/95.3	66.5/89.9	**75.8/94.0**	**74.5/93.0**	**58.1**/83.9	**53.6/79.3**	34.8/64.4	**61.0/81.9**	**53.5/81.9**	**52.2/78.3**	54.4/72.1	37.1/63.2	46.9/81.6	33.0/69.2	42.8/71.5	54.6/79.7
OVT-OpenCLIP	80.9/95.6	67.8/90.8	65.0/89.3	71.2/91.9	58.0/84.2	45.8/73.4	42.8/75.0	50.3/71.4	51.7/79.5	49.7/76.7	**61.9/81.2**	**59.5/85.6**	**52.8/82.5**	**35.4/80.1**	**52.4/82.4**	**56.0/82.4**
MetaCLIP	80.7/95.6	67.6/90.5	**77.7/95.2**	**75.3/93.8**	59.5/85.4	**55.9/81.4**	32.4/62.5	**63.2/83.8**	**52.0/84.2**	**52.6/79.5**	61.4/76.7	41.0/67.8	48.9/**87.9**	34.8/73.2	46.5/76.4	56.3/82.0
OVT-MetaCLIP	**80.7/95.6**	**69.7/92.0**	71.8/93.0	74.0/93.5	**60.6/85.8**	47.8/75.8	**43.5/73.8**	49.0/70.8	50.1/80.1	50.2/77.2	**64.0/79.2**	**54.8/80.4**	**55.1**/84.8	**35.6/77.0**	**52.4/80.3**	**56.9/82.3**
B. Comparisons with ViT-B/16 baselines																
OpenAI CLIP	82.1/95.7	68.3/91.9	67.2/89.4	72.5/92.3	61.8/87.4	48.2/76.3	27.7/55.7	59.1/83.0	52.2/**84.6**	49.8/77.4	51.6/68.9	36.9/63.8	53.4/86.8	30.1/66.1	43.0/71.4	53.2/79.1
Open CLIP	83.2/96.2	70.1/91.8	**77.0/94.8**	**76.8/94.3**	62.2/87.0	**56.0/82.0**	30.7/59.8	**64.9/85.6**	54.3/82.7	**53.6/79.4**	58.1/74.4	44.2/70.9	48.5/84.0	34.6/74.6	46.4/76.0	57.0/82.0
OVT-OpenCLIP	**83.9/97.0**	**71.9/93.1**	69.0/90.7	74.9/93.6	**64.0/88.6**	50.5/77.9	**36.8/68.9**	57.0/77.2	**56.3**/84.5	52.9/**79.4**	**65.4/80.7**	**61.7/85.8**	**56.9/87.4**	**42.4/84.9**	**56.6/84.7**	**59.6/84.7**
EVA-CLIP	**85.3**/96.5	**74.6/94.2**	87.5/**98.0**	**82.5/96.3**	**67.0/89.8**	57.6/82.3	21.3/47.3	69.6/87.5	**53.1**/83.1	53.7/78.0	61.8/76.6	44.3/69.4	53.9/87.4	32.9/73.2	48.2/76.6	59.1/82.1
MetaCLIP	84.3/97.2	72.1/93.4	**78.9**/95.4	78.4/95.3	65.0/89.3	**60.1/84.8**	26.2/56.4	**70.2/89.3**	52.3/85.4	**54.8/81.0**	64.2/79.4	49.6/76.1	48.9/**90.9**	38.5/78.7	50.3/81.2	59.2/84.7
OVT-MetaCLIP	83.4/**97.4**	73.8/94.1	73.9/93.6	77.0/95.0	65.9/89.4	53.6/81.0	**36.2/66.8**	59.0/79.6	51.6/**83.8**	53.2/80.1	**69.7/84.0**	**64.8/87.3**	**55.2**/87.8	**39.2/82.9**	**57.2/85.5**	**60.5/85.6**
C. Comparisons with ViT-L/14 baselines																
OpenAI CLIP	86.5/97.4	75.4/94.6	76.5/93.3	79.5/95.1	**69.8**/90.9	59.5/84.3	18.6/43.8	72.8/91.4	52.9/88.8	54.7/79.8	60.3/75.6	45.8/71.5	47.9/88.2	38.0/74.1	48.0/77.3	58.6/82.8
Open CLIP	86.8/97.8	75.2/94.3	**83.7/96.7**	**81.9/96.2**	67.7/90.2	**63.2/86.4**	24.0/50.5	**74.5/91.2**	54.5/85.0	56.8/80.6	65.7/78.1	53.2/76.7	52.4/90.5	42.3/83.0	53.4/82.1	61.9/85.0
OVT-OpenCLIP	**89.0/97.8**	**77.3/95.3**	79.2/95.3	81.8/96.1	69.6/**91.5**	61.9/86.0	**27.5/55.4**	71.3/88.7	**56.4/87.0**	**57.3/81.7**	**72.2/86.6**	**69.8/89.7**	**57.3/94.1**	**50.0/89.3**	**62.3/89.9**	**65.1/88.1**
EVA-CLIP	88.5/97.9	**79.6/96.0**	**90.6/98.6**	**86.3/97.5**	**72.8/92.7**	68.0/89.1	16.3/40.0	**82.8/95.7**	54.7/87.4	58.9/81.0	71.5/82.3	61.1/81.7	54.4/**94.5**	39.6/86.1	56.6/86.1	65.0/86.8
MetaCLIP	88.3/**98.3**	79.1/95.9	84.1/96.9	83.8/97.0	72.5/92.6	**68.9/89.8**	17.0/40.6	81.8/95.1	**56.6**/87.5	**59.3**/81.1	77.3/89.3	66.4/87.0	**58.9**/93.3	**48.1**/89.6	62.7/89.8	**66.6**/88.0
OVT-MetaCLIP	**88.8**/97.5	77.7/95.9	84.0/96.9	83.5/96.8	70.8/92.2	64.4/87.9	**20.8/47.0**	77.0/92.7	56.3/**89.3**	57.8/**81.8**	**79.3/90.6**	**75.4/93.0**	57.0/94.4	46.4/**93.8**	**64.5/92.9**	66.5/**89.3**
D. Comparisons with BLIP ViT-B/16 baselines																
BLIP	76.6/93.3	52.9/80.2	**67.0**/88.3	65.5/87.3	47.3/74.7	**51.0/76.6**	25.6/53.4	**64.3/83.8**	53.9/**87.6**	48.4/75.2	55.2/68.2	36.8/63.3	50.8/**89.9**	27.0/66.1	42.4/71.9	50.7/77.1
OVT-BLIP	**82.2/97.0**	**61.7/88.8**	66.6/**88.9**	**70.2/91.5**	**53.7/82.9**	46.5/74.2	**33.8/62.7**	57.4/77.9	**56.4**/87.3	**49.6/77.0**	**62.6/79.0**	**54.8/79.9**	**55.2**/89.5	**31.5/73.2**	**51.0/80.4**	**55.2/81.8**

accuracy of viewpoint-OOD datasets, OVT-OpenCLIP with ViT-B/32, ViT-B/16, and ViT-L/14 shows improvements of 9.6%, 10.2%, and 8.9% over OpenCLIP, respectively. OVT-BLIP demonstrated an average improvement of 8.6%. **(2)** While enhancing viewpoint invariance, OVT maintains performance on clean samples and 2D-OOD without significant performance trade-offs. For 2D-OOD benchmarks, OVT-OpenCLIP with ViT-B/32, ViT-B/16, and ViT-L/14 sacrifice only 2.6%, 1.4%, and 0.2% accuracy. **(3)** Compared to earlier CLIP baselines, the recently developed MetaCLIP exhibits better zero-shot performance. Based on this, OVT further enhances its performance under viewpoint-OOD samples.

Visualization. We showcase OVT-OpenCLIP and the original OpenCLIP prediction on several viewpoint-OOD samples. As illustrated in Fig. 4, OVT-CLIP successfully predicts the categories of images from various unusual viewpoints in all cases, whereas the original CLIP is prone to make incorrect predictions.

5.2 Performance on Other Tasks

Settings. As LLaVA and Openflamingo use the OpenAI CLIP (ViT-L/14) to encode vision inputs, we applied OVT to this model in this section and comparing with other OpenAI CLIP (ViT-L/14) versions in image captioning tasks. The training setup remains consistent with the OVT-OpenCLIP described in Tab. 2.

Baselines. In addition to comparing with the original OpenAI CLIP version, we also select $TeCoA^4$ [37] and $FARE^4$ [48], robust CLIP models based on adversarial training, as baselines, which have been proven to possess good resistance to adversarial samples in image captioning task.

Table 4: Image captioning performance under clean distribution samples and viewpoint-OOD samples from Real-world and Synthetic domains. We utilize the MP-Net [50] to calculate the similarity between generated descriptions and ground-truth labels, considering predictions successful if they exceed the similarity threshold β.

		Real-world Domain						**Synthetic Domain**					
		OOD-CV (iid) [61]			OOD-CV (Pose) [61]			IM3D [46]			ImageNet-V+ [46]		
Model	**Visual Encoder**	β@1.0	β@0.5	β@*Adp.*	β@1.0	β@0.5	β@*Adp.*	β@1.0	β@0.5	β@*Adp.*	β@1.0	β@0.5	β@*Adp.*
LLaVa-7b	OpenAI CLIP(ViT-L/14)	44.1	61.1	67.5	46.4	**53.6**	58.7	46.7	53.3	58.8	20.4	25.5	32.1
	TeCoA^4 [37](ViT-L/14)	41.9	58.9	65.5	36.1	41.6	49.2	26.3	30.1	42.6	8.7	11.6	22.6
	FARE^4 [48](ViT-L/14)	42.1	58.9	65.2	40.2	45.9	50.8	35.2	39.2	49.2	12.7	15.8	23.1
	OVT-CLIP(ViT-L/14)	43.5	59.5	65.9	**46.5**	**53.6**	**59.1**	49.4	54.0	61.8	**26.4**	**31.9**	**41.0**
LLaVa-13b	OpenAI CLIP(ViT-L/14)	45.4	68.0	70.6	**48.6**	**58.6**	60.8	48.7	56.7	60.8	21.2	28.4	32.5
	TeCoA^4 [37](ViT-L/14)	42.4	67.0	72.2	37.4	48.9	51.3	25.0	28.6	41.5	8.4	10.9	21.8
	FARE^4 [48](ViT-L/14)	43.9	66.7	71.1	41.9	52.1	54.8	36.1	41.4	48.6	12.1	15.9	20.8
	OVT-CLIP(ViT-L/14)	45.7	67.3	70.8	48.2	**58.6**	**61.9**	50.4	58.9	63.2	**26.4**	**36.2**	**40.9**

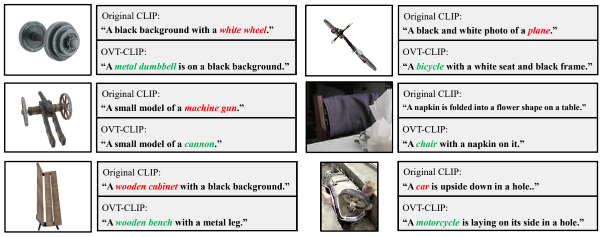

Original CLIP:
"A black background with a white wheel."
OVT-CLIP:
"A metal dumbbell is on a black background."
Original CLIP:
"A black and white photo of a plane."
OVT-CLIP:
"A bicycle with a white seat and black frame."
Original CLIP:
"A small model of a machine gun."
OVT-CLIP:
"A small model of a cannon."
Original CLIP:
"A napkin is folded into a flower shape on a table."
OVT-CLIP:
"A chair with a napkin on it."
Original CLIP:
"A wooden cabinet with a black background."
OVT-CLIP:
"A wooden bench with a metal leg."
Original CLIP:
"A car is upside down in a hole.."
OVT-CLIP:
"A motorcycle is laying on its side in a hole."

Fig. 5: The image descriptions generated by LLaVa-13B using our OVT-CLIP and the original OpenAI CLIP as vision encoder, where ***red texts*** indicates incorrect category descriptions, and ***green texts*** represents correct.

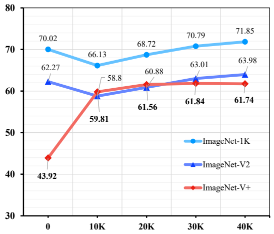


Fig. 6: The Top-1 accuracy of OVT-OpenCLIP (ViT-B/16) with the iterations increases.

Datasets and Metrics. Given the absence of caption benchmarks that include viewpoint-changing OOD samples, we conduct evaluations using existing viewpoint-OOD datasets, including real-world datasets (using OOD-CV (iid) to represent clean distribution and OOD-CV (Pose) for viewpoint-OOD) and synthetic datasets (using IM3D [46] for clean distribution and ImageNet-V+ for viewpoint-OOD). We adopt word embedding distance to calculate the accuracy of the captioning task. By adopting MPNet [50], a state-of-the-art textual embedding model, we measure the similarity between keywords in the generated description and the ground-truth categories. Then assess the accuracy by counting the number of samples that exceed a specific similarity threshold β.

Results and Discussions. Tab. 4 shows the image captioning accuracy of CLIP models under different training strategies, considering β at 1.0 (indicating predictions involve ground-truth categories), 0.5, and *Adp.* (meaning β is equal to the average similarity in the clean distribution). We found that OVT-CLIP improves the accuracy of descriptions generated by LLaVa for viewpoint-OOD samples while maintaining its performance on corresponding clean distributions. When used as the visual encoder for the LLaVa-7B model, OVT-CLIP achieved an 8.9% increase in accuracy compared to the original CLIP model weights under $\beta = Adp.$ Besides, we find that although robust CLIP versions maintain perfor-

Table 5: Average Top-1/Top-5 zero-shot accuracy (%) under different data distributions within various ablation settings.

$\mathcal{L}_{ITC}$	VIFormer	$\mathcal{L}_{VC}$	**Total Avg.**		**Clean Avg.**		**Common-OOD Avg.**		**Viewpoint-OOD Avg.**	
✗	✗	✗	61.9	85.0	81.9	96.2	56.8	80.6	53.4	82.1
✓	✗	✗	61.9	85.7 (↑0.7)	79.4 (↓2.5)	95.6 (↓0.6)	56.2 (↓0.6)	81.4 (↑0.8)	56.0 (↑2.6)	83.8 (↑1.7)
✓	✓	✗	62.2 (↑0.3)	86.2 (↑1.2)	79.9 (↓2.0)	95.4 (↓0.8)	55.4 (↓1.4)	81.6 (↑1.0)	57.5 (↑4.1)	85.2 (↑3.1)
✓	✓	✓	**65.1 (↑3.2)**	**88.1 (↑3.1)**	**81.8 (↓0.1)**	**96.1 (↓0.1)**	**57.3 (↑0.5)**	**81.7 (↑1.1)**	**62.3 (↑8.9)**	**89.9 (↑7.8)**

mance on clean distribution samples, they experience a significant performance decline when facing viewpoint-OOD samples. We select some examples with the generated description in Fig. 5. For the visual question-answering task, we used OpenFlamingo as the VLLMs. The results are reported in the Appendix A.

5.3 Ablation Studies and Additional Results

Our ablation studies focus on the VIFormer and the $\mathcal{L}_{VC}$ within the Omniview-Tuning framework. Tab. 5 shows the Top-1/Top-5 acc of OVT-OpenCLIP (ViT-L/14) across various data distributions and different ablation settings. Beyond the original OpenCLIP, we set a baseline that only uses $\mathcal{L}_{ITC}$ for fine-tuning. Keeping other training settings fixed, reliance solely on $\mathcal{L}_{ITC}$ led to a more significant performance decline in clean and 2D-OOD samples while achieving limited viewpoint OOD performance improvement (2.6%/1.7%). The integration of VIFormer led to further improvements in viewpoint OOD accuracy (4.1%/3.1%). With the further addition of $\mathcal{L}_{VC}$, the improvement in viewpoint OOD performance is most significant (8.9%/7.8%), and it also reduces performance sacrifices in other data distributions. More ablation analyses are available in Appendix B. Furthermore, we report OVT's training convergence, depicted in Fig. 6. We display the Top-1 accuracy evolution for OVT-OpenCLIP (ViT-B/16) across various training iterations. We observe that around 40K iterations, with a batch size of 512, are sufficient for effective convergence, thus achieving a balance in performance across different data distributions.

6 Conclusions

To tackle the challenge of 3D viewpoint invariance in VLP models, this paper first introduced the MVCap dataset, a million-scale collection of image-text pairs with diverse viewpoint variations. Building upon this groundwork, we then proposed the Omniview-Tuning framework, which incorporates a novel Cross-Viewpoint Alignment objective in a parameter-efficient manner, effectively enhancing the VLP models' ability to generate viewpoint-invariant representations. Moreover, through extensive experiments, we successfully verified that Omniview-Tuning could bring significant improvements in viewpoint invariance while preserving the original performance. These advancements provide valuable insights and a standard for future research on viewpoint invariance in foundation models.

Acknowledgments

This work was supported by the Project of the National Natural Science Foundation of China (No. 62076018, 92370124, 62276149, 92248303), the Fundamental Research Funds for the Central Universities, and Tsinghua-Alibaba Joint Research Program. Y. Dong was also supported by the China National Postdoctoral Program for Innovative Talents and CCF-BaiChuan-Ebtech Foundation Model Fund.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 23716–23736 (2022)
2. Alcorn, M.A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.S., Nguyen, A.: Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4845–4854 (2019)
3. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
4. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019)
5. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological review **94**(2), 115 (1987)
6. Calian, D.A., Stimberg, F., Wiles, O., Rebuffi, S.A., Gyorgy, A., Mann, T., Gowal, S.: Defending against image corruptions through adversarial augmentations. arXiv preprint arXiv:2104.01086 (2021)
7. Cha, J., Lee, K., Park, S., Chun, S.: Domain generalization by mutual-information regularization with pre-trained models. European Conference on Computer Vision (ECCV) (2022)
8. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
9. Chuang, Y.S., Dangovski, R., Luo, H., Zhang, Y., Chang, S., Soljačić, M., Li, S.W., Yih, W.t., Kim, Y., Glass, J.: Diffcse: Difference-based contrastive learning for sentence embeddings. arXiv preprint arXiv:2204.10298 (2022)
10. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21126–21136 (2022)
11. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)

12. Das, S., Ryoo, M.S.: Viewclr: Learning self-supervised video representation for unseen viewpoints. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5573–5583 (2023)
13. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
14. 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. Ieee (2009)
15. Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., Zhu, J.: Benchmarking robustness of 3d object detection to common corruptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1022–1032 (2023)
16. Dong, Y., Ruan, S., Su, H., Kang, C., Wei, X., Zhu, J.: Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. Advances in Neural Information Processing Systems **35**, 36789–36803 (2022)
17. Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., Schmidt, L.: Data determines distributional robustness in contrastive language image pre-training (CLIP). In: Proceedings of the 39th International Conference on Machine Learning. pp. 6216–6234 (2022)
18. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108 (2023)
19. 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 pp. 1–15 (2023)
20. Hamdi, A., Ghanem, B.: Towards analyzing semantic robustness of deep neural networks. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 22–38. Springer (2020)
21. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8349 (2021)
22. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
23. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
24. Ho, C.H., Leung, B., Sandstrom, E., Chang, Y., Vasconcelos, N.: Catastrophic child's play: Easy to perform, hard to defend adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9229–9237 (2019)
25. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
26. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773, if you use this software, please cite it as below.

27. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
28. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)
29. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
30. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
31. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
32. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2023)
34. Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., Su, H.: Openshape: Scaling up 3d shape representation towards open-world understanding. Advances in Neural Information Processing Systems **36** (2024)
35. Madan, S., Henry, T., Dozier, J., Ho, H., Bhandari, N., Sasaki, T., Durand, F., Pfister, H., Boix, X.: When and how cnns generalize to out-of-distribution category-viewpoint combinations. arXiv preprint arXiv:2007.08032 (2020)
36. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
37. Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: Understanding zero-shot adversarial robustness for large-scale models. arXiv preprint arXiv:2212.07016 (2022)
38. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
39. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
40. 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)
41. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8821–8831 (2021)
42. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
43. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10901–10911 (2021)
44. Rong, X.: word2vec parameter learning explained. arXiv preprint arXiv:1411.2738 (2014)

45. Ruan, S., Dong, Y., Su, H., Peng, J., Chen, N., Wei, X.: Improving viewpoint robustness for visual recognition via adversarial training. arXiv preprint arXiv:2307.11528 (2023)
46. Ruan, S., Dong, Y., Su, H., Peng, J., Chen, N., Wei, X.: Towards viewpoint-invariant visual recognition via adversarial training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4709–4719 (2023)
47. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Gontijo-Lopes, R., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems (2022)
48. Schlarmann, C., Singh, N.D., Croce, F., Hein, M.: Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. arXiv preprint arXiv:2402.12336 (2024)
49. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
50. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems **33**, 16857–16867 (2020)
51. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
52. 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 (2024)
53. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
54. Tu, W., Deng, W., Gedeon, T.: A closer look at the robustness of contrastive language-image pre-training (CLIP). In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
55. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in Neural Information Processing Systems **32** (2019)
56. Wang, X., He, K., Gupta, A.: Transitive invariance for self-supervised visual representation learning. In: Proceedings of the IEEE international conference on computer vision. pp. 1329–1338 (2017)
57. Wu, Z., Wang, Z., Xu, X., Lu, J., Yan, H.: Embodied task planning with large language models. arXiv preprint arXiv:2307.01848 (2023)
58. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023)
59. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., et al.: Mvimgnet: A large-scale dataset of multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9150–9161 (2023)
60. Zhang, Y., Huang, Y., Sun, Y., Liu, C., Zhao, Z., Fang, Z., Wang, Y., Chen, H., Yang, X., Wei, X., et al.: Benchmarking trustworthiness of multimodal large language models: A comprehensive study. arXiv preprint arXiv:2406.07057 (2024)
61. Zhao, B., Yu, S., Ma, W., Yu, M., Mei, S., Wang, A., He, J., Yuille, A., Kortylewski, A.: Ood-cv: A benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. In: ICML 2022 Shift Happens Workshop (2022)

62. Zhou, X., Liu, M., Zagar, B.L., Yurtsever, E., Knoll, A.C.: Vision language models in autonomous driving and intelligent transportation systems. arXiv preprint arXiv:2310.14414 (2023)