Parameter-efficient Fine-tuning in Hyperspherical Space for Open-vocabulary Semantic Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Open-vocabulary semantic segmentation seeks to label each pixel in an image with arbitrary text descriptions. Vision-language foundation models, especially CLIP, have recently emerged as powerful tools for acquiring open-vocabulary capabilities. However, fine-tuning CLIP to equip it with pixel-level prediction ability often suffers three issues: 1) high computational cost, 2) misalignment between the two inherent modalities of CLIP, and 3) degraded generalization ability on unseen categories. To address these issues, we propose H-CLIP, a symmetrical parameter-efficient fine-tuning (PEFT) strategy conducted in hyperspherical space for both of the two CLIP modalities. Specifically, the PEFT strategy is achieved by a series of efficient block-diagonal learnable transformation matrices and a dual cross-relation communication module among all learnable matrices. Since the PEFT strategy is conducted symmetrically to the two CLIP modalities, the misalignment between them is mitigated. Furthermore, we apply an additional constraint to PEFT on the CLIP text encoder according to the hyperspherical energy principle, i.e., minimizing hyperspherical energy during fine-tuning preserves the intrinsic structure of the original parameter space, to prevent the destruction of the generalization ability offered by the CLIP text encoder. Extensive evaluations across various benchmarks show that H-CLIP achieves new SOTA open-vocabulary semantic segmentation results while only requiring updating approximately 4% of the total parameters of CLIP.

1 Introduction

The aim of open-vocabulary semantic segmentation is to create a segmentation model capable of labeling each pixel in an image with categories that are not limited to a specific closed set according to text descriptions. Vision-language foundation models [43, 5, 34, 39, 11, 17, 26, 21, 27, 13, 18, 29, 10, 28, 45], especially CLIP [39], are often utilized to endow open-vocabulary recognition capabilities. Consequently, open-vocabulary semantic segmentation essentially boil down to transferring these vision-language foundation models, originally trained with image-level supervision, to perform pixel-level predictions.

To this end, current methods [52, 48, 7, 50] typically fine-tune CLIP on a benchmark dataset with segmentation annotations, i.e., COCO [2], to equip it with the segmentation ability. However, this often leads to three main issues. First, fine-tuning CLIP on limited categories would affect its generalization ability, resulting in significant performance degradation on unseen categories. Second, current fine-tuning strategies are usually asymmetrical, which inevitably causes a misalignment between the two inherent modalities of CLIP, i.e., image and text [52], which may lead to sub-optimal performance. Third, although remarkable performance gains, these approaches often rely on computationally extensive full fine-tuning, which raises concerns about scalability and affordability.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

To address these issues, we propose a symmetric parameter-efficient fine-tuning (PEFT) strategy for CLIP, dubbed H-CLIP. Specifically, we implement this PEFT through a partial orthogonal fine-tuning (POF) strategy, which introduces a series of efficient block-diagonal learnable transformation matrices into the hyperspherical space. Then, to preserve CLIP's generalization ability, we leverage the hyperspherical energy principle [32, 38], which suggests that maintaining the same hyperspherical energy during fine-tuning preserves the intrinsic structure, i.e., generalization ability. In light of this, we upgrade our POF by incorporating orthogonal constraints in the learnable matrices for updating CLIP's text encoder, as orthogonal transformations keep the hyperspherical energy unchanged during fine-tuning. Subsequently, we introduce a dual cross-relation communication (DCRC) module to explicitly encourage cross-modal and cross-layer communications within all learnable matrices. This communication not only preserves the hyperspherical energy but also further mitigating the misalignment problem.

Extensive results demonstrate that H-CLIP achieves new state-of-the-art open-vocabulary semantic segmentation results across three benchmarks by fine-tuning CLIP with approximately 4% of the total parameters of CLIP.

2 Related Work

2.1 Open-vocabulary Semantic Segmentation

Prior open-vocabulary semantic segmentation works typically perform this task through leveraging CLIP [39]. initial efforts like [56] directly fine-tune CLIP on mainstream segmentation datasets, e.g., COCO [2]. However, they claim that fine-tuning CLIP's encoder significantly reduces its ability to generalize to unseen classes. To address this issue, some methods [15, 8, 51, 49] swing to the opposite extreme, fine-tuning an additional mask generator [6] for segmentation while keeping CLIP frozen to maintain generalization-oriented recognition. However, this frozen parameter space lacks segmentation awareness, resulting in a misalignment between regions and text descriptions [30]. Other studies [52, 50, 7] propose an advanced solution that fine-tunes only selected parameters, e.g., certain layers of CLIP, to enable pixel-level predictions while keeping most of CLIP's parameters fixed, thus minimizing losing of generalization. Although the advantages are remarkable, these methods often work with a very small learning rate, implicitly encouraging a small deviation from the pre-trained CLIP, limiting the segmentation performance. In a nutshell, the trade-off between preserving CLIP's generalization and learning segmentation knowledge persists, hindering the final performance. Based on the paradigm of existing fine-tuning-based methods, our method explores a better trade-off from a fresh viewpoint: hyperspherical space.

2.2 Large-scale Model Fine-tuning

Along with the improvement of large-scale foundation models [26, 34, 28, 23, 53, 42, 41, 40, 60], e.g., segment anything model [23], numerous fine-tuning works [37, 36, 4, 57, 58, 14, 54, 47, 31, 61] are proposed to adapt these models to various downstream scenarios. The core of these approaches lies in updating only limited parameters to capture the specific characteristics of different scenarios, while keeping most parameters fixed to maintain generalization. In contrast, fine-tuning CLIP for open-vocabulary semantic segmentation often meets a dilemma. On the one hand, limited parameters typically fall short in facilitating the transition from a classification model, i.e., CLIP, to a segmentation task. On the other hand, directly increasing the number of trainable parameters risks undermining CLIP's ability to generalize to unseen classes, as experimented in CAT-Seg [7]. Most methods [52, 48] solve this issue by simply freezing CLIP's text encoder and fine-tuning its image encoder, inevitably causing misalignment between the two modalities of CLIP. In this paper, we shed light on how to preserve generalization in a symmetric parameter-efficient fine-tuning manner and strive to explore an appropriate fine-tuning method for open-vocabulary semantic segmentation.

3 Preliminaries

3.1 Hyperspherical Energy

Existing fine-tuning methods implicitly assume that a smaller Euclidean distance between the fine-tuned model and the pre-trained model indicates better preservation of the pre-trained ability. However, the Euclidean difference is unable to fully capture the degree of semantic preservation. According to the inspiration from Thomson problem[44] which is to determine the minimum electrostatic potential energy configuration of N mutually-repelling electrons on the surface of a unit sphere, we adopt the *Hyperspherical Energy* to characterize the diversity of the model. The hyperspherical energy function of a fully connected layer $\boldsymbol{W}$ is defined as $\mathrm{HE}(\boldsymbol{W}) := \sum_{i \neq j} \|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^{-1}$, where $\hat{\boldsymbol{w}}_i := \boldsymbol{w}_i / \|\boldsymbol{w}_i\|$ denotes the i-th normalized neuron. The power of the model representation can be characterized by the hyperspherical energy of its neurons. Higher energy implies higher redundancy, while lower energy indicates that these neurons of the model are more diverse. For the original semantic information not to be destroyed in the case of fine-tuning, we hypothesize that a good fine-tuning model should have a minimal difference in hyperspherical energy compared to the pre-trained model:

$$\min_{\boldsymbol{W}} \left\|\mathrm{HE}(\boldsymbol{W}) - \mathrm{HE}(\boldsymbol{W}^0)\right\| \quad \Leftrightarrow \quad \min_{\boldsymbol{W}} \left\| \sum_{i \neq j} \|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^{-1} - \sum_{i \neq j} \|\hat{\boldsymbol{w}}_i^0 - \hat{\boldsymbol{w}}_j^0\|^{-1} \right\|. \tag{1}$$

One can easily observe that the attainable minimum is zero for Eq. (1). In this case, the hyperspherical energy should satisfy an invariance property (the application of the same orthogonal transformation for all neurons demonstrates the pairwise hyperspherical similarity). Based on the hyperspherical energy invariance property, the minimum of zero can be achieved as long as $\boldsymbol{W}$ and $\boldsymbol{W}^0$ differ only up to a rotation or reflection, i.e., $\boldsymbol{W} = \boldsymbol{R}\boldsymbol{W}^0$ in which $\boldsymbol{R} \in \mathbb{R}^{d \times d}$ is an orthogonal matrix (The determinant 1 or -1 means rotation or reflection, respectively).

3.2 Notation of Tensor Product

In this section, we introduce the fundamental concept underlying our DCRC (Sec. 4.3): tensor product. A p-order tensor is indexed by p indices and can be represented as a multidimensional array of data. Formally, a p-order tensor $\mathcal{A}$ can be written as $\mathcal{A} = (a_{i_1, i_2, \cdots, i_p}) \in \mathbb{R}^{n_1 \times n_2 \times \cdots n_p}$. Slices of a tensor are matrices defined from the tensor by holding all but two indices constant. For a 3-order tensor, $\mathcal{A}(:, :, k)$ corresponds the k^{th} frontal slice. For p-order tensors, matrix slices of p-order tensors can be referenced using linear indexing by reshaping the tensor into an $n_1 \times n_2 \times n_3 n_4 \cdots n_p$ 3-order tensor and referring to the k^{th} frontal slice as $\mathcal{A}(:, :, k)$. $\mathcal{A}_i \in \mathbb{R}^{n_1 \times n_2 \times \cdots n_{p-1}}$ for $i = 1, \cdots, n_p$ denotes the $(p-1)$-order tensor created by holding the pth index of $\mathcal{A}$ fixed at i. It is possible to create a tensor in a block circulant pattern, where each block is a tensor of $(p-1)$-order:

$$\mathrm{circ}(\mathcal{A}) = \begin{bmatrix} \mathcal{A}_1 & \mathcal{A}_{n_p} & \mathcal{A}_{n_p - 1} & \cdots & \mathcal{A}_2 \\ \mathcal{A}_2 & \mathcal{A}_1 & \mathcal{A}_{n_p} & \cdots & \mathcal{A}_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}_{n_p} & \mathcal{A}_{n_p - 1} & \mathcal{A}_{n_p - 2} & \cdots & \mathcal{A}_1 \end{bmatrix},$$

where $\mathrm{circ}(\cdot)$ creates a block circulant tensor and the size of $\mathrm{circ}(\mathcal{A})$ is $(n_1 n_p \times n_2 n_p \times \cdots \times n_{p-2} n_p \times n_{p-1})$. define $\mathrm{unfold}(\cdot)$ to take an $n_1 \times \cdots \times n_p$ tensor $\mathcal{A}$ and return an $n_1 n_p \times n_2 \times \cdots n_{p-1}$ block tensor in the following way:

$$\mathrm{unfold}(\mathcal{A}) = \begin{bmatrix} \mathcal{A}_1 & \mathcal{A}_2 & \cdots & \mathcal{A}_{n_p} \end{bmatrix}^T.$$

The operation that takes unfold back to tensor form is the "fold" command. Specially, $\mathrm{fold}(\cdot, n_p)$ takes an $n_1 n_p \times n_2 \times \cdots \times n_{p-1}$ block tensor and returns an $n_1 \times \cdots \times n_p$ tensor, defined as:

$$\mathrm{fold}(\mathrm{unfold}(\mathcal{A}), n_p) = \mathcal{A}.$$

4 Methodology

4.1 Overview of H-CLIP

Fig. 2 illustrates the proposed H-CLIP framework, which is based on two core components: (1) POF updates the pre-trained parameter space of CLIP using a series of block-diagonal transformation matrices. According to analysis in Sec. 1, each parameter matrix in CLIP's text encoder is orthogonal to preserve generalization. (2) DCRC incorporates cross-modal and cross-layer communication within all tunable matrices, facilitating alignment between different modalities.

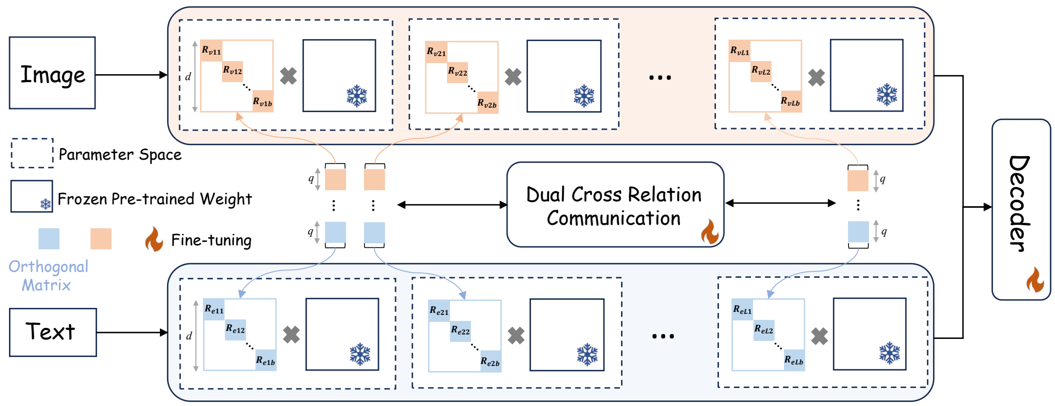

Figure 1: **A schematic representation of H-CLIP.** In the H-CLIP framework, we propose a partial orthogonal fine-tuning strategy, where each pre-trained weight matrix is paired with a tuned block-diagonal transformation matrix, some of which are orthogonal to preserve generalization. Then, we introduce a dual cross-relation communication mechanism to facilitate communication among all matrices, enabling alignment between different modalities.

4.2 Partial Orthogonal Fine-tuning

The core idea of partial orthogonal fine-tuning (POF) is to introduce the concept of hyperspherical space for fine-tuning CLIP. In this hyperspherical space, we fine-tune CLIP's text encoder under an orthogonality design principle from OFT [38] to preserve the hyperspherical energy of the pre-trained parameter space. Similarly, we use Cayley parameterization [3] to ensure a tunable matrix $\boldsymbol{R}$ is strictly orthogonal, formally as:

$$\boldsymbol{R} = (\boldsymbol{I} + \boldsymbol{Q})(\boldsymbol{I} - \boldsymbol{Q})^{-1}, \tag{2}$$

where $\boldsymbol{Q}$ is skew-symmetric. Here, for $\boldsymbol{R}$ in CLIP's image encoder, we remove the orthogonality constraint, defined as:

$$\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{I}, \tag{3}$$

where $\boldsymbol{I}$ is an identity matrix. Considering the relatively large dimension d of the pre-trained matrix, for better efficiency, we introduce a block-diagonal structure by parameterizing $\boldsymbol{R}$ with b blocks, formally as:

$$\boldsymbol{R} = \mathrm{diag}(\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots \boldsymbol{R}_i, \cdots, \boldsymbol{R}_b) = \begin{bmatrix} \boldsymbol{R}_1 & & \\ & \ddots & \\ & & \boldsymbol{R}_b \end{bmatrix}, \tag{4}$$

where $\boldsymbol{R}_i \in \mathbb{R}^{d/b \times d/b}$. Specifically, denote $\mathcal{R}^V = \{\boldsymbol{R}_{v1}, \cdots, \boldsymbol{R}_{v\ell}, \cdots, \boldsymbol{R}_{vL}\}$ and $\mathcal{R}^E = \{\boldsymbol{R}_{e1}, \cdots, \boldsymbol{R}_{e\ell}, \cdots, \boldsymbol{R}_{eL}\}$ as the sets of block-diagonal matrices in CLIP's image encoder and text encoder, respectively, where L is its number of Transformer layers, $\boldsymbol{R}_{v\ell} \in \mathbb{R}^{d_v \times d_v}$, and $\boldsymbol{R}_{e\ell} \in \mathbb{R}^{d_e \times d_e}$. For simplicity, we set $d_v = d_e = d$. Overall, we develop a H-CLIP framework, and for an input feature map $\mathbf{M}_\ell$ in the ℓ^{th} Transformer layer of CLIP, the right branch produces the adjusted feature map via H-CLIP, $\tilde{\mathbf{M}}_\ell$, formally via:

$$\tilde{\mathbf{M}}_\ell = \begin{cases} \mathcal{F}_\ell(\mathbf{M}_\ell; \boldsymbol{R}_\ell \mathbf{W}_\ell), & \text{if } \boldsymbol{R}_\ell \in \mathcal{R}^V \\ \mathcal{F}_\ell(\mathbf{M}_\ell; \boldsymbol{R}_\ell \mathbf{W}_\ell), \quad \text{s.t. } \boldsymbol{R}_\ell^\top \boldsymbol{R}_\ell = \boldsymbol{R}_\ell \boldsymbol{R}_\ell^\top = \boldsymbol{I} & \text{otherwise}, \end{cases} \tag{5}$$

where $\mathbf{W}_\ell$ is a pre-trained weight matrix in ℓ^{th} layer of CLIP's encoder, and $\mathcal{F}_\ell$ represents ℓ^{th} layer of CLIP's encoder. During the fine-tuning phase, H-CLIP is fine-tuned in conjunction with the original parameter space of CLIP, which is loaded from the pre-trained checkpoint and remains frozen.

4.3 Dual Cross Relation Communication

Although in POF, we relax the orthogonal constraint for CLIP's image encoder to learn segmentation knowledge, each layer of the image encoder still incorporates a limited number of parameters, which largely restricts the flexibility of the projection adjustment due to the limitation of Hidden Markov

Chain along layers [24, 46, 36]. To address this limitation, one might consider fully fine-tuning instead of using a small number of parameters. However, this approach can cause a misalignment between image and text features in CLIP, resulting in sub-optimal performance [52]. Based on the above analysis, we introduce Dual Cross-Relation Communication (DCRC), which facilitates interaction among different layers and modalities (i.e., text and image). DCRC explicitly enhances the flexibility of fine-tuned projection adjustments and prevents misalignment issues.

DCRC introduces cross-layer and cross-modality communication among different block-diagonal matrices, achieved through two relation projections. To do this, we first treat all blocks in ℓ^{th} layer as an individual slice in this 3-order tensor $\mathcal{T}_\ell$, which is derived as follows:

$$\mathcal{T}_\ell = [\boldsymbol{R}_{v\ell 1}, \boldsymbol{R}_{e\ell 1}, \cdots, \boldsymbol{R}_{v\ell i}, \boldsymbol{R}_{e\ell i}, \cdots, \boldsymbol{R}_{v\ell b}, \boldsymbol{R}_{e\ell b}] \in \mathbb{R}^{q\times q\times(b+b)}, \tag{6}$$

Where $q = d/b$. Then, we treat the tensor $\mathcal{T}_\ell$ as an individual slice within a 4-order tensor $\mathcal{T}$, defined as follows:

$$\mathcal{T} = [\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_\ell, \cdots, \mathcal{T}_L] \in \mathbb{R}^{q\times q\times(b+b)\times L}. \tag{7}$$

Initially, according to the characteristics of gradient propagation in deep learning theory, i.e., chain rule, each frontal slice $\boldsymbol{R}_{\cdot\ell i} \in \{\mathbb{R}^{q\times q}\}^{(b+b)\times L}$ is updated sequentially in CLIP's encoder. As a result, updating the $\mathcal{T}$ lacks cross-frontal-slice communication, limiting the flexibility of adjusting fine-tuned projection. To address this, we introduce two special tensor products, i.e., 3**-order T-product** and **Higher-order T-product**.

Definition 4.1(3-order T-product) For $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and $\mathcal{B} \in \mathbb{R}^{n_2\times l\times n_3}$, the 3-order T-product $\mathcal{C} \in \mathbb{R}^{n_1\times l\times\times n_3} = \mathcal{A} * \mathcal{B}$ is defined as:

$$\mathcal{C} = \mathcal{A} * \mathcal{B} = \text{fold}(\text{circ}(\mathcal{A}) \cdot \text{unfold}(\mathcal{B})), \tag{8}$$

where "$\cdot$" represents standard matrix product.

Definition 4.2(Higher-order T-product) For $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3\cdots\times n_p}$ and $\mathcal{B} \in \mathbb{R}^{n_2\times l\times n_3\times\cdots\times n_p}$, the High-order T-product $\mathcal{C} \in \mathbb{R}^{n_1\times l\times n_3\cdots\times n_p} = \mathcal{A} * \mathcal{B}$ is defined as:

$$\mathcal{C} = \mathcal{A} * \mathcal{B} = \text{fold}(\text{circ}(\mathcal{A}) * \text{unfold}(\mathcal{B})). \tag{9}$$

If $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$, according to the **3-order T-product**, there is an invertible transform $S_3(\cdot) : \mathbb{R}^{n_1\times n_2\times n_3} \to \mathbb{R}^{n_1\times n_2\times n_3}$ in third dimension and it transform the Eq. (8) as:

$$\mathcal{C} = S_3^{-1}(S_3(\mathcal{A}) \odot S_3(\mathcal{B})) = S_3^{-1}(\bar{\mathcal{A}} \odot \bar{\mathcal{B}}) = S_3^{-1}(\bar{\mathcal{C}}), \tag{10}$$

where $\bar{\mathcal{C}} = \bar{\mathcal{A}} \odot \bar{\mathcal{B}}$ denotes the frontal-slice-wise product (Definition 2.1 refers to [19]) $\bar{\mathcal{C}}(;,;,i) = \bar{\mathcal{A}}(;,;,i) \cdot \bar{\mathcal{B}}(;,;,i), i = 1, 2, \cdots, n_3$ and $S_3^{-1}(\cdot)$ is the inverse transform of $S_3(\cdot)$. According to the definition of the frontal-slice-wise product, the invertible transform $S_3(\cdot)$ is formulated as:

$$\bar{\mathcal{A}} = S_3(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{S}_3, \tag{11}$$

where "$\times_3$" denotes the mode-3 product and $\mathbf{S}_3 \in \mathbb{R}^{n_3\times n_3}$ is an arbitrary invertible matrix. Similarly, the inverse transform of Eq. (11) is derived as:

$$\mathcal{A} = S_3^{-1}(\bar{\mathbf{A}}) = \bar{\mathcal{A}} \times_3 \mathbf{S}_3^{-1}. \tag{12}$$

Similarly, if $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_p}$, according to the **Higer-order T-product**, there are invertible transform $S_i(\cdot) : \mathbb{R}^{n_1\times n_2\times\cdots\times n_p} \to \mathbb{R}^{n_1\times n_2\times\cdots\times n_p}, i = 3, 4, \cdots, p$ in i^{th} dimension and they transform the Eq. (9) as:

$$\mathcal{C} = \tilde{S}^{-1}(\tilde{S}(\mathcal{A}) \odot \tilde{S}(\mathcal{B})) = \tilde{S}^{-1}(\bar{\mathcal{A}} \odot \bar{\mathcal{B}}) = \tilde{S}^{-1}(\bar{\mathcal{C}}), \tag{13}$$

where $\tilde{S}(\mathcal{A}) = S_p(S_{p-1}(\cdots S_3(\mathcal{A})\cdots))$, $\bar{\mathcal{C}} = \bar{\mathcal{A}} \odot \bar{\mathcal{B}}$ denotes the frontal-slice-wise product $\bar{\mathcal{C}}(;,;,i) = \bar{\mathcal{A}}(;,;,i) \cdot \bar{\mathcal{B}}(;,;,i), i = 1, 2, \cdots, n_3 n_4 \cdots n_p$ and $\tilde{S}^{-1}(\cdot)$ is the inverse transform of $\tilde{S}(\cdot)$. Similarly, the inverse transform $\tilde{S}(\cdot)$ is formulated as:

$$\bar{\mathcal{A}} = \tilde{S}(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{S}_3 \times_4 \mathbf{S}_4 \cdots \times_p \mathbf{S}_p, \tag{14}$$

and its inverse transform is derived as:

$$\mathcal{A} = \tilde{S}^{-1}(\bar{\mathcal{A}}) = \bar{\mathcal{A}} \times_3 \mathbf{S}_3^{-1} \times_4 \mathbf{S}_4^{-1} \cdots \times_p \mathbf{S}_p^{-1}. \tag{15}$$

Model	VLM	Additional Backbone	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
			Traditional Fine-Tuning					
ZS3Net [1]	-	ResNet-101	-	-	-	19.4	38.3	-
LSeg [25]	CLIP ViT-B/32	ResNet-101	-	-	-	-	47.4	-
ZegFormer [8]	CLIP ViT-B/16	ResNet-101	4.9	9.1	16.9	42.8	86.2	62.7
ZSseg [51]	CLIP ViT-B/16	ResNet-101	7.0	-	20.5	47.7	88.4	-
OpenSeg [15]	ALIGN	ResNet-101	4.4	7.9	17.5	40.1	-	63.8
OVSeg [30]	CLIP ViT-B/16	ResNet-101c	7.1	11.0	24.8	53.3	92.6	-
ZegCLIP [59]	CLIP ViT-B/16	-	-	-	-	41.2	93.6	-
CAT-Seg [7]	CLIP ViT-B/16	-	12.0	19.0	31.8	57.5	94.6	77.3
			Parameter-efficient Fine-Tuning					
SAN [50]	CLIP ViT-B/16	-	10.1	12.6	27.5	53.8	94.0	-
Ours	CLIP ViT-B/16	-	**12.4**	**19.3**	**32.4**	**57.9**	**95.2**	**78.2**
			Traditional Fine-Tuning					
LSeg [25]	CLIP ViT-B/32	ViT-L/16	-	-	-	-	52.3	-
OpenSeg [15]	ALIGN	Eff-B7	8.1	11.5	26.4	44.8	-	70.2
OVSeg [30]	CLIP ViT-L/14	Swin-B	9.0	12.4	29.6	55.7	94.5	-
SAN [50]	CLIP ViT-L/14	-	12.4	15.7	32.1	57.7	94.6	-
ODISE [49]	CLIP ViT-L/14	Stable Diffusion	11.1	14.5	29.9	57.3	-	-
CAT-Seg [7]	CLIP ViT-L/14	-	16.0	23.8	37.9	63.3	97.0	82.5
			Parameter-efficient Fine-Tuning					
SAN [50]	CLIP ViT-L/14	-	12.4	15.7	32.1	57.7	94.6	-
Ours	CLIP ViT-L/14	-	**16.5**	**24.2**	**38.4**	**64.1**	**97.7**	**83.2**

Table 1: **Comparison with state-of-the-art methods on standard benchmarks.** The best-performing results are presented in bold, while the second-best results are underlined. “VLM”: visual language model.

Derivation. please refer to supplementary material. ■

According to Eqs. (29), (30) and (31), we adopt its idea and design arbitrary invertible relation matrix $\mathbf{S}_3 \in \mathbb{R}^{(b+b)\times(b+b)}$ and $\mathbf{S}_4 \in \mathbb{R}^{L\times L}$ to capture the cross-modality and cross-layer information in $\mathcal{T}$. Then the updated tensor $\mathcal{T}_w$ is formulated as:

$$\mathcal{T}_w = \mathcal{T} \times_3 \mathbf{S}_3 \times_4 \mathbf{S}_4 \in \mathbb{R}^{q\times q\times(b+b)\times L}, \tag{16}$$

where the relation matrix $\mathbf{S}_3$ and $\mathbf{S}_4$ are learnable. To better capture the nonlinear interactions inside the whole parameter space, we further adopt k layers deep neural network (DNN) $f_3(\cdot)$ and $f_4(\cdot)$ to replace the transform $\times_3\mathbf{S}_3$ and $\times_4\mathbf{S}_4$, respectively, and the DNN $f_3(\cdot)$ is formulated as:

$$f_3(\mathcal{T}) = \sigma(\sigma(\cdots\sigma(\sigma(\mathcal{A}\times_3 \mathbf{W}_1)\times_3 \mathbf{W}_2)\cdots)\times \mathbf{W}_{k-1})\times \mathbf{W}_k, \tag{17}$$

where $\sigma(\cdot)$ is a nonlinear scalar function and matrices $\{\mathbf{W}_j \in \mathbb{R}^{(b+b)}\}_{j=1}^{k}$. The DNN $f_4(\cdot)$ is similar. Finally, the $\mathcal{T}$ is updated by $\mathcal{T} = \mathcal{T} + \boldsymbol{\alpha}\mathcal{T}_w$, where $\boldsymbol{\alpha} \in \mathbb{R}^{(b+b)\times L}$ is a learnable parameter.

5 Experiments

5.1 Experimental Setup

Datasets. Following previous studies [7, 48], we utilizes the COCO-Stuff dataset [2] as our training set. This dataset comprises approximately 118,000 densely annotated images across 171 distinct semantic categories. During inference, we carry out comparisons with state-of-the-art methods across several semantic segmentation datasets, including ADE20K [55], PASCAL VOC [12], and PASCAL-Context [35]. **ADE20K [55]** is a classical semantic segmentation dataset comprising around 20,000 training images and 2,000 validation images. Besides, it includes two different test sets: `A-150` and `A-847`. The test set `A-150` has 150 common categories, while the test set `A-847` has 847 categories. **PASCAL VOC [12]** is a small dataset for semantic segmentation, which includes 1464 training images and 1449 validation images. The dataset contains 20 different foreground categories. We name it as `PAS-20`. In line with [7], we also report a score on `PAS-20`^{b}, which involves “background” as the 21st category. **PASCAL-Context [35]** is upgraded from the original PASCAL VOC dataset. It includes two different test sets: `PC-59` and `PC-459` for evaluation. The test set `PC-59` has 59 categories, while the test set `PC-459` has 459 categories.

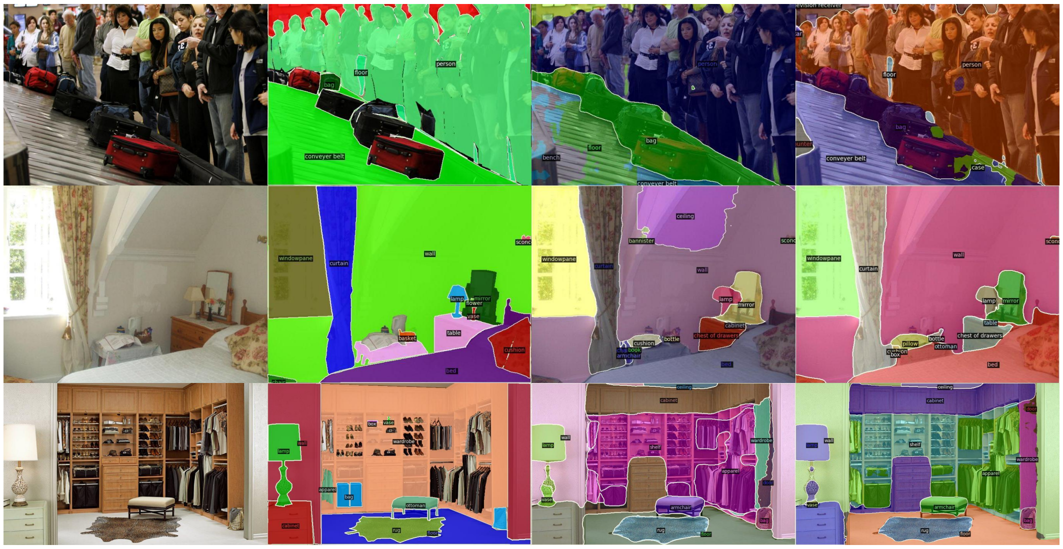

Image Ground truth CAT-Seg [7] Ours

Figure 2: Comparison of qualitative reults on ADE20K [55] with 150 categories. we compare Our method with CAT-Seg [7].

Evaluation metric. Following prior works [7, 48], we adopt mean Intersection over Union (mIoU) to evaluate the semantic segmentation performance on the three benchmarks.

Implementation Details. We implement our method using the Transformer-based CLIP model. Following the protocol established in [7], we evaluate our results on two versions of the CLIP model: ViT-B/16 and ViT-L/14. For training, we use the Adam optimizer [22] with an initial learning rate of 5×10^{-6} for CLIP, and a weight decay of 10^{-4}. Training is conducted with one image per mini-batch. We set $q = 128$ for balancing efficiency and performance. The function $f_3(\cdot)$ and $f_4(\cdot)$ are implemented using two 2-layer MLPs. We act the cost-based approach provided in [7] as our decoder. All models are trained over 80,000 iterations on 4 NVIDIA RTX 3090 GPUs.

5.2 Main Results

Comparing to SOTAs. Here, we compare our proposed H-CLIP with several state-of-the-art methods, as shown in Table 1, using six test sets across three benchmarks. Overall, we achieve the best results. Most existing open-vocabulary semantic segmentation methods employ traditional fine-tuning approaches, i.e., full or partial fine-tuning (tuning certain layers of CLIP). While these methods offer sufficient flexibility for learning new knowledge, they often result in a significant performance drop on unseen classes, as observed with OVSeg [30]. Among these methods, CAT-Seg [7] achieves performance comparable to ours. However, its fine-tuning scheme is manually controlled through different layer combinations, necessitating a careful design to balance generalization and flexibility, while ours does not suffer from such an issue. Then, compared to SAN [50], another parameter-efficient fine-tuning method that introduces only a limited number of tunable parameters, our approach significantly outperforms it, achieving improvements of 6.6% on the PC-459 dataset and 3.9% on the PC-59 dataset with ViT-B/16 as the base model. These results demonstrate the effectiveness of our method in preserving generalization while learning segmentation knowledge.

Methods	OVSeg [30]	CAT-Seg [7]	SAN [50]	Ours
Param. (M)	147.2	25.6	8.4	**5.6**

Table 2: **Efficiency comparison** in terms of learnable parameters.

Qualitative results. Here, we visualize our method's representative example segmentation results against prevailing methods, e.g., CAT-Seg [7] in the PC-459 dataset. As shown in Figs. 2 - 4, we observe that our approach is able to generalize on diverse scenarios and produce more accurate results.

Image | Ground truth | CAT-Seg [7] | Ours

Figure 3: Comparison of qualitative reults on VOC2010 [12] with 59 categories.

Image | Ground truth | CAT-Seg [7] | Ours

Figure 4: Comparison of qualitative reults on ADE20K [55] with 847 categories.

Method	POF	DCRC	Param. (M)	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
Freeze	✗	✗	0	4.4	6.6	24.8	49.4	92.5	71.9
LoRA [16]	✗	✗	7.5	11.4	17.6	28.6	55.1	94.2	76.7
H-CLIP	✓	✗	5.62	12.3	19.0	31.6	56.4	94.6	76.3
	✗	✓	0.01	7.6	10.9	26.8	53.6	92.7	74.5
	✓	✓	5.63	**12.4**	**19.3**	**32.4**	**57.9**	**95.2**	**78.2**

Table 3: **Ablation study on the components of H-CLIP**. "LoRA": a mainstream parameter-efficient tuning method with a comparable number of parameters for comparison. "POF": Partial Orthogonal Fine-tuning. "DCRC": Dual Cross Relation Communication. The base model is ViT-B/16.

Efficiency comparison. We compare the efficiency of our method with other approaches, including OVSeg [30], CAT-Seg [7], and SAN [50], all of which utilize CLIP ViT models. The comparison,

	Block dimension q	Param. (M)	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
(a)	256×256	22.52	12.4	19.2	**32.7**	57.6	**95.4**	77.9
	128×128	5.63	**12.4**	**19.3**	32.4	**57.9**	95.2	**78.2**
	64×64	1.41	11.7	18.4	31.7	56.9	95.0	76.4
	Orthogonal Constraint	Param. (M)	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
(b)	w/o	7.51	11.9	18.5	32.2	57.5	**95.3**	76.9
	with	3.76	12.2	19.1	31.4	57.1	94.3	76.8
	POF	5.63	**12.4**	**19.3**	**32.4**	**57.9**	95.2	**78.2**

Table 4: **Ablation study on different designs in POF**. We show the impact of (a) different block dimensions q and (b) orthogonal constraints. The base model is ViT-B/16.

summarized in Table 2, shows that our method employs the fewest trainable parameters while balancing the generalization of the pre-trained model and the flexibility for learning new knowledge. Additionally, since we introduce a lightweight architecture for calculating relations, specifically two relation matrices, the inference overhead is negligible during the inference phase.

5.3 Ablative Studies

Ablation of Main Components. Here, we conduct an ablation study to demonstrate the benefits of each component of our proposed H-CLIP: partial orthogonal fine-tuning (POF) and dual cross-relation communication (DCRC). We use the ViT-B/16 [9] version of CLIP as the baseline, shown in row 1 of Table 3. Additionally, we implement a mainstream parameter-efficient fine-tuning (PEFT) method, LoRA [16], for comparison with a similar number of learnable parameters, as shown in row 2. Note that LoRA can improve performance compared to the baseline, demonstrating that PEFT is a viable approach for this task. Then, comparing row 5 to row 2, we observe significant performance gains, indicating that our results are driven by our targeted solution rather than merely the number of parameters. Moreover, row 3 shows that using only POF preserves generalization on unseen classes, particularly in the A-847 dataset. Meanwhile, solely adapting DCRC shows limited improvement, as it only enhances communication among frozen weight matrices. Finally, integrating DCRC with POF yields clear performance gains, e.g., a 12.6% improvement on the PC-459 dataset.

Different Design of POF. Table 4 presents experiments introducing different designs into POF. The design of POF is related to (1) block dimension, i.e., q, and (2) how orthogonality constraints are applied. In (a), the results show that larger In (a), the results show that larger q generally performs better than smaller q. However, we find a good trade-off between performance and parameter efficiency, with $q = 128$ working well across datasets and tasks. Therefore, we maintain this setting in other experiments. In (b), we show that both blindly applying orthogonality constraints to the learnable matrices of all layers and not using any constraints at all can degrade performance on most test sets, demonstrating the value of our analysis with the hyperspherical energy principle.

6 Conclusion

In this paper, we propose a H-CLIP framework to address three issues: 1) high computational cost, 2) misalignment between the two inherent modalities of CLIP, and 3) degraded generalization ability on unseen categories when equipping CLIP with pixel-level prediction ability for open-vocabulary semantic segmentation. Specifically, we propose a symmetrical parameter-efficient fine-tuning (PEFT) strategy conducted in hyperspherical space for both of the two CLIP modalities. Specifically, the PEFT strategy is achieved by a series of efficient block-diagonal learnable transformation matrices and a dual cross-relation communication module among all learnable matrices to mitigate misalignment between different modalities. Furthermore, we apply an additional constraint to PEFT on the CLIP text encoder according to the hyperspherical energy principle, i.e., minimizing hyperspherical energy during fine-tuning preserves the intrinsic structure of the original parameter space, to prevent the destruction of the generalization ability offered by the CLIP text encoder. Extensive experiments demonstrate that the proposed H-CLIP framework generalized improves segmentation performance across several benchmarks while introducing approximately 4% of CLIP's total parameters. We hope our approach will provide a new direction and inspire future research in this field.

References

[1] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[3] Arthur Cayley. Sur quelques propriétés des déterminants gauches. 1846.

[4] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything? – sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more, 2023.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[7] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation, 2024.

[8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

[11] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[13] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[15] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[17] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[19] Eric Kernfeld, Misha Kilmer, and Shuchin Aeron. Tensor–tensor products with invertible linear transforms. *Linear Algebra and its Applications*, 485:545–570, 2015.

[20] Misha E Kilmer and Carla D Martin. Factorization strategies for third-order tensors. *Linear Algebra and its Applications*, 435(3):641–658, 2011.

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[24] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421, 2021.

[25] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[28] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[30] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023.

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[32] Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. *Advances in neural information processing systems*, 31, 2018.

[33] Canyi Lu, Xi Peng, and Yunchao Wei. Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5996–6004, 2019.

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[35] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

[36] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Lingxi Xie, Qi Tian, and Wei Shen. Parameter efficient fine-tuning via cross block orchestration for segment anything model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[37] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Xiaokang Yang, and Wei Shen. Sam-parser: Fine-tuning sam efficiently by parameter space reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4515–4523, 2024.

[38] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[40] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[41] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[44] JJ Tomson. On the structure of the atom: an investigation of the stability and periods of osciletion of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory atomic structure. *Philos. Mag. Series 6*, 7(39):237, 1904.

[45] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.

[46] Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan Kuruoglu, and Yefeng Zheng. Information theoretic counterfactual learning from missing-not-at-random feedback. *Advances in Neural Information Processing Systems*, 33:1854–1864, 2020.

[47] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.

[48] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation, 2023.

[49] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.

[50] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023.

[51] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022.

[52] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, 2023.

[53] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022.

[54] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[56] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

[59] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[60] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[61] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2605–2615, 2023.

Appendix of H-CLIP

A Derivation of the Definition

In this section, we provide derivations of definitions in the main paper.

Definition 4.1(3-order T-product) For $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $\mathcal{B} \in \mathbb{R}^{n_2 \times l \times n_3}$, the 3-order T-product $\mathcal{C} \in \mathbb{R}^{n_1 \times l \times n_3} = \mathcal{A} * \mathcal{B}$ is defined as:

$$\mathcal{C} = \mathcal{A} * \mathcal{B} = \text{fold}(\text{circ}(\mathcal{A}) \cdot \text{unfold}(\mathcal{B})), \tag{18}$$

where "$\cdot$" represents standard matrix product.

Definition 4.2(Higher-order T-product) For $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3 \cdots \times n_p}$ and $\mathcal{B} \in \mathbb{R}^{n_2 \times l \times n_3 \times \cdots \times n_p}$, the High-order T-product $\mathcal{C} \in \mathbb{R}^{n_1 \times l \times n_3 \cdots \times n_p} = \mathcal{A} * \mathcal{B}$ is defined as:

$$\mathcal{C} = \mathcal{A} * \mathcal{B} = \text{fold}(\text{circ}(\mathcal{A}) * \text{unfold}(\mathcal{B})). \tag{19}$$

Derivation. According to [20], if $\mathcal{A}$ is $n_1 \times n_2 \times n_3$, $\mathcal{A}$ can be block diagonalized by using Discrete Fourier Transformer (DFT) matrix $\mathbf{F}_{n_3} \in \mathbb{R}^{n_3 \times n_3}$ as:

$$(\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_1}) \cdot \text{circ}(\text{unfold}(\mathcal{A})) \cdot (\mathbf{F}^*_{n_3} \otimes \mathbf{I}_{n_2}) = \mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & & \\ & \ddots & \\ & & \mathbf{D}_{n_3} \end{bmatrix} \in \mathbb{R}^{n_1 n_3 \times n_2 n_3}, \tag{20}$$

where "$\otimes$" denotes the Kernecker product, "$\mathbf{F}^*_{n_3}$" denotes the conjugate transpose of $\mathbf{F}_{n_3}$, "$\cdot$" means standard matrix product and $\mathbf{D}$ is a block diagonal matrix. In fact, the i-th block matrix $\mathbf{D}_i$ of $\mathbf{D}$ can be computed by applying DFT of $\mathcal{A}$ along 3-rd dimension. The **3-order T-product** in Eq. (18) can be computed as:

$$(\mathbf{F}^*_{n_3} \otimes \mathbf{I}_{n_1}) \cdot ((\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_1}) \cdot \text{circ}(\text{unfold}(\mathcal{A})) \cdot (\mathbf{F}^*_{n_3} \otimes \mathbf{I}_{n_2})) \cdot (\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_2}) \cdot \text{unfold}(\mathcal{B}). \tag{21}$$

It is readily shown that $(\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_2})$unfold can be computed by applying DFT of $\mathcal{B}$ along 3-rd dimension: the result called $\bar{\mathbf{B}}$. Thus, Eq. (21) remains to multiply each block matrix $\mathbf{D}_i$ of $\mathbf{D}$ with each block matrix $\mathbf{B}_i$ of $\bar{\mathbf{B}}$, then take an inverse DFT along the 3-rd dimension of the result. Hence, the **3-order T-product** in Eq. (18) can be re-formulated as:

$$\mathcal{C} = \text{DFT}_3^{-1}(\text{DFT}_3(\mathcal{A}) \odot \text{DFT}_3(\mathcal{B})) = \text{DFT}_3^{-1}(\bar{\mathcal{A}} \odot \bar{\mathcal{B}}) = \text{DFT}_3^{-1}(\bar{\mathcal{C}}), \tag{22}$$

where $\text{DFT}_3(\cdot)$ is DFT along 3-rd dimension and $\text{DFT}_3^{-1}(\cdot)$ is the inverse DFT along 3-rd dimension. In mathematics, the DFT of $\mathcal{A}$ along 3-rd dimension is formulated as:

$$\bar{\mathcal{A}} = \text{DFT}_3(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{F}_{n_3}. \tag{23}$$

Similarly, the inverse DFT of $\bar{\mathcal{A}}$ along 3-rd dimension is derived as:

$$\mathcal{A} = \text{DFT}_3^{-1}(\bar{\mathcal{A}}) = \bar{\mathcal{A}} \times_3 \mathbf{F}^{-1}_{n_3}. \tag{24}$$

By the detailed theoretical analysis in [33], the DFT has been extended to a general invertible transform S with an invertible transform matrix $\mathbf{S}$. In mathematics, the invertible transform of $\mathcal{A}$ along 3-rd dimension is formulated as:

$$\bar{\mathcal{A}} = \text{S}_3(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{S}_{n_3}. \tag{25}$$

Similarly, the inverse transform of $\bar{\mathcal{A}}$ along 3-rd dimension is derived as:

$$\mathcal{A} = \text{S}_3^{-1}(\bar{\mathcal{A}}) = \bar{\mathcal{A}} \times_3 \mathbf{S}^{-1}_{n_3}. \tag{26}$$

Similarly, if $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_p}$, $\mathcal{A}$ can be block diagonalized by using a sequence of DFT matrices $\mathbf{F}_{n_i} \in \mathbb{R}^{n_i \times n_i}, i = 3, 4, \cdot, p$ as:

$$(\mathbf{F}_{n_p} \otimes \mathbf{F}_{n_{p-1}} \otimes \cdots \otimes \mathbf{F}_{n_3} \otimes \mathbf{I}_{n_1}) \cdot \tilde{\mathcal{A}} \cdot (\mathbf{F}^*_{n_p} \otimes \mathbf{F}^*_{n_{p-1}} \otimes \cdots \otimes \mathbf{F}^*_{n_3} \otimes \mathbf{I}_{n_2}) = \mathbf{D}, \tag{27}$$

where $\tilde{\mathcal{A}} = \text{circ}(\text{unfold}(\mathcal{A})) \in \mathbb{R}^{n_1 n_3 n_4 \cdots n_p \times n_2 n_3 \cdots n_p}$. Since the matrix $\mathbf{D}$ is block diagonal with $n_3 n_4 \cdots n_p$ blocks each of size $n_1 \times n_2$, the **Higher-order T-product** in Eq. (19) can be computed as:

$$(\tilde{\mathbf{F}}^* \otimes \mathbf{I}_{n_1}) \cdot ((\tilde{\mathbf{F}} \otimes \mathbf{I}_{n_1}) \cdot \tilde{\mathcal{A}} \cdot (\tilde{\mathbf{F}}^* \otimes \mathbf{I}_{n_2})) \cdot (\tilde{\mathbf{F}}_{n_3} \otimes \mathbf{I}_{n_2}) \cdot \tilde{\mathcal{B}}, \tag{28}$$

where $\tilde{\mathbf{F}} = \mathbf{F}_{n_p} \otimes \mathbf{F}_{n_{p-1}} \otimes \cdots \otimes \mathbf{F}_{n_3}$. Using the DEF, it is straightforward to show that the block diagonal matrix $\mathbf{D}$ in Eq. (27) can be obtained by repeated DFTs of $\mathcal{A}$ along each dimension expect for 1-st and 2-nd dimension. Similarly, by using a sequence invertible transform $S_j(\cdot), i = 3, 4, \cdot, p$ with invertible transform matrix $\mathbf{S}_i$, the **Higher-order T-product** in Eq. (19) can be re-formulated as:

$$\mathcal{C} = \tilde{S}^{-1}(\tilde{S}(\mathcal{A}) \odot \tilde{S}(\mathcal{B})) = \tilde{S}^{-1}(\bar{\mathcal{A}} \odot \bar{\mathcal{B}}) = \tilde{S}^{-1}(\bar{\mathcal{C}}), \tag{29}$$

where $\tilde{S}(\mathcal{A}) = S_p(S_{p-1}(\cdots S_3(\mathcal{A}) \cdots))$, $\bar{\mathcal{C}} = \bar{\mathcal{A}} \odot \bar{\mathcal{B}}$ denotes the frontal-slice-wise product $\bar{\mathcal{C}}(;,;,i) = \bar{\mathcal{A}}(;,;,i) \cdot \bar{\mathcal{B}}(;,;,i), i = 1, 2, \cdots, n_3 n_4 \cdots n_p$ and $\tilde{S}^{-1}(\cdot)$ is the inverse transform of $\tilde{S}(\cdot)$. The inverse transform $\tilde{S}(\cdot)$ is formulated as:

$$\bar{\mathcal{A}} = \tilde{S}(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{S}_3 \times_4 \mathbf{S}_4 \cdots \times_p \mathbf{S}_p, \tag{30}$$

and its inverse transform is derived as:

$$\mathcal{A} = \tilde{S}^{-1}(\bar{\mathcal{A}}) = \bar{\mathcal{A}} \times_3 \mathbf{S}_3^{-1} \times_4 \mathbf{S}_4^{-1} \cdots \times_p \mathbf{S}_p^{-1}. \tag{31}$$

■

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract clearly states the following claims about the paper's contributions and scope. We aim to address a significant challenge in open-vocabulary semantic segmentation: fine-tuning CLIP to achieve per-pixel predictions without compromising its generalization capabilities. To this end, we propose a novel H-CLIP, which introduces a partial orthogonal fine-tuning strategy that prevents a drop in hyperspherical energy, thereby preserving generalization. Subsequently, H-CLIP employs dual cross-relation communication to increase projection flexibility, facilitating the acquisition of segmentation knowledge.

2. **Limitations**

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

   Answer: [NA]

   Justification: Those are not discussed in the paper.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

   Justification: All assumptions are proved clearly in the main paper or the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: The detailed experimental settings and information are provided in Sec. 5, and this information is sufficient to reproduce the main experimental results.

   Guidelines:

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

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

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

   Answer: [No]

   Justification: The code will be made public after acceptance.

Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: All the experimental settings are clarified in Sec. 5. (and Appendix).

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [No]

   Justification: error bars are not reported because it would be too computationally expensive.

   Guidelines:

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

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: The computing requirements are provided in Sec. 5.

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We ensure our research adheres to the guidelines.

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: there is no societal impact of the work performed.

    Guidelines:

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

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: the paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: All the assets used are properly credited and are the license and terms of use explicitly mentioned and properly respected.

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

13. **New Assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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