A Parameter-Efficient and Fine-Grained Prompt Learning for Vision-Language Models

Yongbin Guo[1,2] **Shuzhen Li**[2] **Zhulin Liu**[*,1] **Tong Zhang**[1,2,3] **C. L. Philip Chen**[1,2,3]
[1]Guangdong Provincial Key Laboratory of Computational AI Models and Cognitive Intelligence, School of Computer Science & Engineering, South China University of Technology, Guangzhou, China
[2]Pazhou Lab, Guangzhou, China
[3]Engineering Research Center of the Ministry of Education on Health Intelligent Perception and Paralleled Digital-Human, Guangzhou, China
{tomrongxu,melody13szli}@gmail.com, {liuzhl,tony,philipchen}@scut.edu.cn

Abstract

Current vision-language models (VLMs) understand complex vision-text tasks by extracting overall semantic information from large-scale cross-modal associations. However, extracting from large-scale cross-modal associations often smooths out semantic details and requires large computations, limiting multi-modal fine-grained understanding performance and efficiency. To address this issue, this paper proposes a detail-oriented prompt learning (DoPL) method for vision-language models to implement fine-grained multi-modal semantic alignment with merely 0.25M trainable parameters. According to the low-entropy information concentration theory, DoPL explores shared interest tokens from text-vision correlations and transforms them into alignment weights to enhance text prompt and vision prompt via detail-oriented prompt generation. It effectively guides the current frozen layer to extract fine-grained text-vision alignment cues. Furthermore, DoPL constructs detail-oriented prompt generation for each frozen layer to implement layer-by-layer localization of fine-grained semantic alignment, achieving precise understanding in complex vision-text tasks. DoPL performs well in parameter-efficient fine-grained semantic alignment with only 0.12% tunable parameters for vision-language models. The state-of-the-art results over the previous parameter-efficient fine-tuning methods and full fine-tuning approaches on six benchmarks demonstrate the effectiveness and efficiency of DoPL in complex multi-modal tasks.

1 Introduction

Vision-language models (VLMs) are large-scale pre-trained models in the multimodal domain that simultaneously process and understand both visual and textual information(Radford et al., 2021; Li et al., 2022b). These models excel in cross-modal tasks such as image captioning, visual question answering, and image-text retrieval.

How to effectively capture cross-modal associations is a fundamental challenge in vision-language models(Zhang et al., 2021; Li et al., 2022b). Dual-stream and fused-stream VLMs typically focus on global semantic relationships between modalities. For example, Dual-stream VLMs like CLIP(Radford et al., 2021) and ALIGN(Jia et al., 2021) captured cross-modal association with global feature similarity between modalities. Fused-stream VLMs like BLIP(Li et al., 2022b) and X-VLM(Zeng et al., 2021) improved alignment through cross-modal attention. However, these global perspective modeling approaches often overlook subtle semantic details and struggle to accurately capture image-text relationships(Yao et al., 2021), as shown in Figure 1(b) and Figure 1(c). Furthermore, these models with massive parameters rely on full fine-tuning to capture cross-modal relationships for downstream tasks, which leads to high computational costs. Therefore, exploring parameter-efficient transfer learning (PETL) methods is essential for more efficient model adaptation.

The main challenge in PETL for VLMs is effectively enabling cross-modal interactions while maintaining parameter efficiency. PETL methods(Liu et al., 2023; Gao et al., 2024) for VLMs typically address the resource-intensive nature of full fine-tuning by freezing the backbone and introducing a small number of trainable parameters for task-specific adaptation. They also optimize cross-modal interactions within these constraints. For instance, prompt-based methods like MaPLe(Khattak et al., 2023) utilized multi-modal prompts learning to enhance alignment and adapt to downstream tasks. Adapter-based methods like MMA(Yang et al., 2024) leveraged shared multi-modal projection layers and high-layer adapters to strengthen vision-text connections. However, these methods sacrifice the ability to capture fine-grained seman-

* Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31346–31359

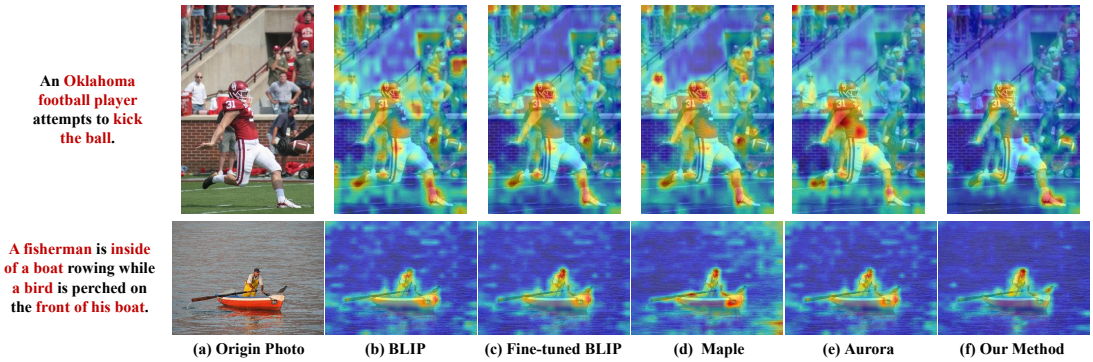


Figure 1: Comparison of cross-modal attention map on FLICKR30K shows that our DoPL effectively captures detailed cross-modal semantic relationships.

tic details in favor of achieving parameter-efficient adaptation, as shown in Figure 1(d) and 1(e). This sacrifice impacts their performance in tasks requiring precise alignment, such as visual question answering. Therefore, this work focuses on fine-grained semantic alignment and parameter-efficient adaptation in VLMs.

To this end, we propose a novel detail-oriented prompt learning method for VLMs that simultaneously achieves fine-grained semantic alignment and parameter efficiency. Guided by the low-entropy information concentration theory, DoPL leverages the detail-oriented prompt generation module to identify shared interest tokens from text-vision prompt correlations and transform them into alignment weights to enhance text-vision prompt pairs. Then, DoPL utilizes the enhanced pairs to generate detail-oriented prompts for each frozen layer. Subsequently, It progressively guides each frozen layer with prompts to extract fine-grained text-vision alignment cues, enabling layer-by-layer localization of fine-grained semantic alignment and achieving precise understanding in vision-text complex tasks. As a result, DoPL achieves both parameter-efficient adaptation and stronger fine-grained cross-modal semantic alignment.

To conclude, our contributions are as follows:

- DoPL is an innovative parameter-efficient and fine-grained prompt learning for vision-language models. It simultaneously achieves fine-grained multi-modal semantic alignment and parameter efficiency with only 0.12% additional parameters.
- This paper proposes a non-parametric detail-oriented prompt generation (DPG) to improve prompts' ability in fine-grained multi-modal semantic alignment, enabling the precise understanding performance of VLMs in complex multi-modal tasks.

2 Related Work

2.1 Vision-Language Models

Vision-language models (VLMs) integrate visual and textual information to learn cross-modal relationships(Radford et al., 2021; Li et al., 2022b). VLMs can be broadly classified into single-stream, dual-stream, and fused-stream models.

Single-stream models fuse visual and textual data early and pass them through a unified encoder to learn cross-modal relationships. Examples include VL-BERT(Su et al., 2019), OSCAR(Li et al., 2020), and SimVLM(Wang et al., 2021). These models simplify integration but may struggle to distinguish modality-specific features. Dual-stream models process visual and textual information separately through independent encoders and preserve modality-specific features before cross-modal interaction. CLIP(Radford et al., 2021), ALIGN(Jia et al., 2021), and Florence(Yuan et al., 2021) are examples. Although they excel in large-scale pre-training, they often overlook fine-grained details and limit their performance in complex tasks.

Fused-stream models combine the strengths of both architectures by using independent encoders for each modality and cross-modal attention for integration. Models like BLIP(Li et al., 2022b), X-VLM(Zeng et al., 2021), and ALBEF(Li et al., 2021) provide better alignment but at higher computational costs due to the complexity of cross-modal integration, particularly during full fine-tuning.

Our work focuses on dual-stream and fused-stream models and significantly advances the enhancement of fine-grained alignment between modalities while maintaining a low computational overhead.

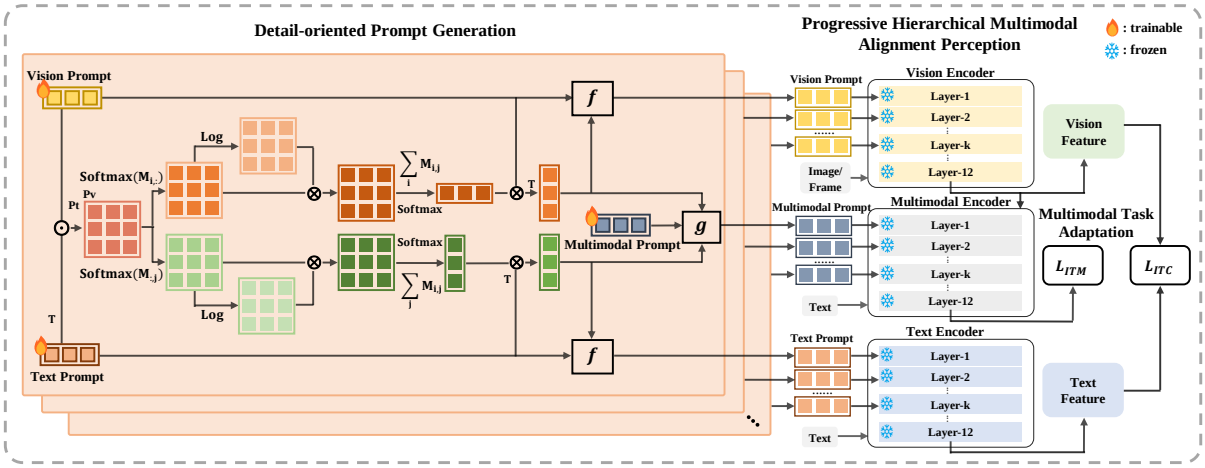


Figure 2: Overview of our detail-oriented prompt learning method: Only the prompts are trained while the backbone remains frozen. The non-parametric DPG enhances prompts based on the low-entropy information concentration theory, fostering fine-grained semantic alignment.

2.2 Parameter-Efficient Transfer Learning for VLMs

Parameter-Efficient Transfer Learning (PETL) methods address the challenges of full fine-tuning in pre-trained models by reducing trainable parameters while maintaining performance(Zhou et al., 2022b; Khattak et al., 2023). For VLMs, PETL requires balancing parameter efficiency and cross-modal interactions. Current PETL in VLMs can be categorized into prompt-based, adapter-based, and reparameterization methods.

Prompt-based methods insert task-specific prompts into vision/language branches while freezing the original model parameters. CoOp(Zhou et al., 2022b) optimized continuous text prompts for few-shot learning. CoCoOp(Zhou et al., 2022a) refined text prompts based on image instances. VPT(Yu et al., 2024) employed visual prompts for vision-language tasks. MaPLe(Khattak et al., 2023) and Bloom(Wang et al., 2024) employed multi-modal prompts with cross-modal interactions.

Adapter-based methods introduce lightweight adapter layers and fine-tune only adapter parameters. Clip-Adapter(Gao et al., 2024) added adapter layers in text and image encoders. UniAdapter(Lu et al., 2023) used shared down-projection layers and residual learning for cross-modal knowledge sharing. MMA(Yang et al., 2024) employed shared projection layers and high-layer adapters to strengthen visual-textual connections.

Reparameterization methods improve transfer learning efficiency by modifying model components. Aurora(Wang et al., 2023a) reduced trainable parameters through mode approximation.

Although existing PETL methods reduce train costs, they typically focus on coarse alignment and overlook fine-grained semantic details. Our method achieves parameter efficiency and fine-grained semantic alignment to address both challenges simultaneously.

3 Method

This section introduces the detail-oriented prompt learning method for VLMs. DoPL guides each freeze layer to extract text-vision alignment cues via the non-parametric detail-oriented prompt generation and implements layer-by-layer localization of fine-grained semantic alignment. DoPL primarily includes detail-oriented prompt generation, progressive hierarchical multimodal alignment perception, and multimodal task adaption. The architecture of our framework is illustrated in Figure 2, and a detailed description is provided below.

3.1 Detail-oriented Prompt Generation

This section focuses on acquiring detail-oriented prompts to improve the model's ability to capture subtle semantic relationships and achieve fine-grained cross-modal alignment. We first construct modality-specific prompts to capture modality-specific information. Then, the non-parametric detail-oriented prompt generation module is used to refine these prompts and enhance their semantic sensitivity according to the low-entropy information concentration theory.

Modality-specific Prompts Construction: To ensure consistent multimodal understanding and retain both shared and modality-specific information, we utilize a multi-modal prompt method for constructing a shared knowledge foundation. It

involves two stages: first, a shared source prompt P^s is constructed as a foundation for all modalities, and second, modality-specific prompts are generated from P^s by using tailored mapping layers for each modality. This approach retains core information while integrating modality-specific features, achieving consistency and diversity.

Formally, the modality-specific prompts are generated as follows:

$$P^\alpha = \phi_\alpha(P^s\mathcal{W}_\alpha + \beta_\alpha), \quad \alpha = \{t, v, m\}, \tag{1}$$

where $\phi_\alpha(\cdot)$ is the corresponding linear function, $\{\mathcal{W}_\alpha,\ \beta_\alpha\}$ is the weight and bias of the corresponding modality-specific mapping layers. P^t, P^v and P^m are text, vision and multimodal prompts, respectively.

Detail-oriented Prompts Construction: Existing multi-modal prompt methods(Khattak et al., 2023; Wang et al., 2024) often use unidirectional or bidirectional prompt transformation structures to enable cross-modal interaction. However, they introduce significant parameter overhead and primarily capture coarse-grained alignment due to their reliance on high-level interactions.

To address these limitations, we propose the low-entropy information concentration theory, which establishes a principled connection between entropy minimization and semantic alignment strength. In the context of soft prompts, consider a similarity matrix $R_{i,j}$ quantifying the relationship between text prompt tokens x_i and visual prompt tokens y_j. High entropy in $R_{i,j}$ indicates that a text prompt token associates with multiple visually unrelated prompt tokens y_j, reflecting ambiguous or weak semantic alignment (e.g., "dog" activating regions containing both a cat and a car). Conversely, low entropy implies that x_i predominantly aligns with a specific subset of y_j, demonstrating concentrated semantic correspondence. This principle is grounded in mutual information theory: minimizing conditional entropy $H(Y|X)$ maximizes the shared information $I(X;Y)$, thereby enhancing semantic coherence between modalities. The detailed mathematical derivation and proof of this theory are provided in Appendix [A].

Guided by this theoretical foundation, we propose the non-parametric DPG module to leverage low entropy values from text-vision correlations, enhancing the prompts' ability to capture fine-grained semantic relationships while maintaining parameter efficiency.

In order to effectively capture the shared interest tokens between text and vision prompts, it is essential to explore their cross-modal interrelationships. Therefore, DPG constructs a cross-modal correlation matrix $R_{i,j}$ using $(P^t)^l$ and $(P^v)^l$ in the l-th layer ($l = 1, 2, ..., k$). The matrix $R_{i,j}$ is acquired as follows:

$$R_{i,j} = softmax(((P^t)^l)^T) \cdot softmax((P^v)^l), \tag{2}$$

where $\cdot$ represents the matrix multiplication. $R_{i,j}$ quantifies cross-modal relationships by calculating the similarity between the corresponding representations, providing a comprehensive measure.

To extract shared interest tokens from the constructed $R_{i,j}$, DPG uses information entropy. Entropy can quantify the alignment strength between the information from different modalities. Therefore, DPG first normalizes $R_{i,j}$, ensuring the values are on a comparable scale and ready for entropy calculation. The matrix $R_{i,j}$ encodes the relationships between a single vision prompt token and all text prompt tokens in the columns, and DPG applies column-wise normalization for text-to-vision relationships:

$$N_{i,j}^{t\to v} = \frac{exp(R_{i,j})}{\sum_{k=1} exp(R_{k,j})}, \tag{3}$$

The matrix $R_{i,j}$ encodes the relationships between a single text prompt token and all vision prompt tokens in the rows, and DPG applies row-wise normalization for vision-to-text relationships:

$$N_{i,j}^{v\to t} = \frac{exp(R_{i,j})}{\sum_{k=1} exp(R_{i,k})}, \tag{4}$$

where i represents text prompt token, j represents vision prompt token and k iterates over rows and columns for normalization.

To quantify the cross-modal informational content between tokens, DPG then computes entropy matrices through element-wise multiplication of the normalized matrices and their logarithmic counterparts. The entropy matrices are acquired as follows:

$$h_{i,j}^{t\to v} = N_{i,j}^{t\to v} \odot log(N_{i,j}^{t\to v} + \epsilon), \tag{5}$$

$$h_{i,j}^{v\to t} = N_{i,j}^{v\to t} \odot log(N_{i,j}^{v\to t} + \epsilon), \tag{6}$$

these matrices quantify the alignment strength between text and vision prompt tokens at a granular level, reflecting the degree of correspondence between the modalities.

According to the low-entropy information concentration theory, low entropy values indicate concentrated relationships and reflect strong alignment, while high entropy values suggest weaker and dispersed alignments. DPG computes the summed entropy matrices by performing row and column summation operations on the entropy matrices to capture the relationship between a token and all tokens in the other modality. Specially, Row-wise summation of $h_{i,j}^{t\to v}$ generates the text-to-vision entropy matrix while column-wise summation of $h_{i,j}^{v\to t}$ generates the vision-to-text entropy matrix. The formulas are given as follows:

$$H_i^{t\to v} = -\sum_j h_{i,j}^{t\to v}, \tag{7}$$

$$H_j^{v\to t} = -\sum_i h_{i,j}^{v\to t}, \tag{8}$$

where i iterates overs rows for $H_i^{t\to v}$ and j iterates over columns for $H_j^{v\to t}$. $H_i^{t\to v}$ measures the relationships of each text prompt token with all vision tokens. $H_j^{v\to t}$ measures the relationships of each vision prompt token with all text prompt tokens.

To assess the token's capacity to capture semantic alignment details, DPG converts $H_i^{t\to v}$ and $H_j^{v\to t}$ into weights, emphasizing low-entropy regions with stronger alignment to enhance fine-grained alignment extraction. The weights are:

$$W_i^{t\to v} = \frac{exp(-H_i^{t\to v})}{\sum_{k=1} exp(-H_k^{t\to v})}, \tag{9}$$

$$W_j^{v\to t} = \frac{exp(-H_j^{v\to t})}{\sum_{k=1} exp(-H_k^{v\to t})}, \tag{10}$$

the matrices dynamically adjust the focus and accentuate regions with stronger alignment to refine cross-modal information flow. Therefore, DPG combines the weight matrices with their respective prompts. This combination enables precise extraction of text-to-vision and vision-to-text alignment information by highlighting tokens with a stronger ability to capture fine-grained alignment details. The formulas are as follows:

$$A^{t\to v} = (W_i^{t\to v})^T \cdot (P^t)^l, \tag{11}$$

$$A^{v\to t} = W_j^{v\to t} \cdot (P^v)^l, \tag{12}$$

Subsequently, DPG employs residual learning to integrate the alignment information with the prompts, enhancing their sensitivity to semantic details. The text, vision, and multimodal prompts at l-th layer are enhanced as follows:

$$(P^t)^l = (P^t)^l + A^{t\to v}, \tag{13}$$

$$(P^v)^l = (P^v)^l + A^{v\to t}, \tag{14}$$

$$(P^m)^l = (P^m)^l + A^{t\to v} + A^{v\to t}, \tag{15}$$

these enhanced prompts possess stronger fine-grained semantic alignment extraction capabilities, guiding the inserted frozen layer to perform detailed alignment and capture subtle cross-modal relationships.

3.2 Progressive Hierarchical Multimodal Alignment Perception

This section introduces a layer-wise progressive perception process that enhances cross-modal alignment precision using detail-oriented prompts at each frozen layer. Our model is based on BLIP(Li et al., 2022b). The text encoder processes textual input through a 12-layer structure and extracts semantic features from the text. Similarly, the vision encoder processes visual input through a 12-layer structure and learns vision features. The multimodal encoder processes text input and visual features to learn joint representations with 12 layers.

In the l-th layer($l = 1, 2, ..., k$), the detail-oriented prompts P^t, P^v and P^m are inserted into the text, vision, and multimodal encoders, respectively. These prompts are concatenated with the key and value matrices in the attention mechanism:

$$K = concat(P^\alpha, K), \tag{16}$$

$$V = concat(P^\alpha, V), \tag{17}$$

where $\alpha = \{t, v, m\}$. These prompts guide the current frozen layer to extract text-vision alignment cues, improving cross-modal interactions while preserving modality-specific features. It also sets the stage for the layer-wise progressive perception process. In the first k layers, the inserted detail-oriented prompts guide the frozen layers to progressively localize fine-grained semantic alignment, with each layer refining and propagating alignment information through interactions with the corresponding encoder layers.

As DoPL guides the frozen layers through this step-by-step localization of fine-grained semantic alignment, it achieves alignment precision and computational efficiency, ultimately enhancing the model's ability to achieve precise understanding in complex vision-text tasks.

Method	# Tunable	MSCOCO TR			MSCOCO IR			FLICKR30K TR			FLICKR30K IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Methods with full fine-tuning:													
UNITER(Chen et al., 2020)	330M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98	99.2	75.6	94.1	96.8
VILLA(Gan et al., 2020)	330M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR(Li et al., 2020)	330M	73.5	92.2	96	57.5	82.8	89.8	-	-	-	-	-	-
ALIGN(Jia et al., 2021)	820M	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100	84.9	97.4	98.6
ALBEF(Li et al., 2021)	210M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100	85.6	97.5	98.9
BLIP(Li et al., 2022b)	223M	**81.9**	**95.4**	**97.8**	**64.3**	**85.7**	**91.5**	**97.3**	99.9	**100**	**87.3**	**97.6**	**98.9**
Methods with forzen backbone:													
LoRA(r=32)(Hu et al., 2022)	10.6M	76.7	91.2	96.0	60.08	82.21	89.5	96.3	99.7	99.8	84.8	96.6	98.4
Maple(len=8)(Khattak et al., 2023)	5.4M	76.87	91.32	96.1	60.1	82.32	89.7	96.6	99.8	99.9	85.1	97.1	98.62
Bloom(len=8)(Wang et al., 2024)	3.54M	77.28	93.7	96.96	60.85	83.87	90.07	96.8	99.8	100	85.3	97.12	98.62
DoRA(r=32)(Liu et al., 2024)	10.8M	77.8	93.5	96.6	61.1	83.9	90.1	96.7	99.8	100	85.1	97.2	98.5
UniAdapter*(r=512)(Lu et al., 2023)	19.5M	79.6	94.5	97.26	62.53	84.95	90.97	96.7	99.7	100	86.16	97.34	98.82
Aurora*(r=64)(Wang et al., 2023a)	0.6M	78	93.4	96.66	61.45	83.95	90.39	96.7	99.8	100	85.76	97.24	98.72
DoPL(ours, len=8)	0.25M	79.9	94.94	**97.8**	62.6	85.12	90.97	**97.3**	**100**	**100**	86.2	**97.6**	98.86

Table 1: Results on image-text retrieval datasets MSCOCO and FLICKR30K, # Tunable: the number of tunable parameters, the second best result is marked by underline. '*' donate the performance obtained by (Ji et al., 2024).

3.3 Multimodal Task Adaptation

To adapt the model for various multimodal tasks, DoPL follows BLIP(Li et al., 2022b) and employs task-specific loss functions. For image-text and video-text retrieval tasks, DoPL uses Image-Text Matching Loss(Li et al., 2021) to assess image-text pair matching and Image-Text Contrastive Loss(Radford et al., 2021; Jia et al., 2021) to enhance semantic consistency through contrastive learning. For generative tasks like VQA and video question answering, DoPL uses Language Modeling Loss(Li et al., 2022b) to aid in generating accurate answers.

4 Experiment

4.1 Experimental Settings

Datasets and Baselines. We evaluate our method on six public multimodal downstream tasks: image-text retrieval datasets (MSCOCO(Lin et al., 2014) and FLICKR30K(Plummer et al., 2015)), video-text retrieval datasets (MSR-VTT(Xu et al., 2016) and DiDeMo(Anne Hendricks et al., 2017)), visual question answering dataset (VQA(Goyal et al., 2017)) and video question answering dataset (MSRVTT-QA(Xu et al., 2017)). And we compare our method with two types of approaches: full fine-tuning methods for each task and frozen backbone methods, including Maple(Khattak et al., 2023), Bloom(Wang et al., 2024), UniAdapter(Lu et al., 2023) and Aurora(Wang et al., 2023a). Maple and Bloom are re-implemented on the BLIP backbone in our experiments.

Implementation Details Our method's vision-language backbone is based on BLIP for the six downstream tasks. During the fine-tuning process, the backbone model's parameters are kept frozen. We use the AdamW optimizer with a weight decay of 0.05, set the prompt length uniformly to 8 for all modalities, and set the inserted layers k to 9. All experiments are implemented in PyTorch on NVIDIA A100 GPUs(80GB).

4.2 Comparisons with State-Of-The-Art

Image-text Retrieval. Table 1 presents the performance of our framework on image-text retrieval tasks with MSCOCO and FLICKR30K. DoPL outperforms frozen backbone models like UniAdapter(19.5M) and Aurora(0.6M) and competes with larger full fine-tuning methods such as BLIP(223M) and ALBEF(221M) with only 0.25M parameters. This demonstrates DoPL's sensitivity to semantic details and remarkable ability to achieve fine-grained semantic alignment with minimal computational cost.

Visual Question Answering. Table 2 presents the performance of our multi-modal fine-grained alignment framework on VQA and VideoQA tasks using VQAv2 and MSRVTT-QA. Our DoPL outperforms frozen backbone methods and achieves strong performance in generative tasks with fewer parameters(0.25M) while also competing with full fine-tuning methods like BLIP(337M) and All-in-one(110M). This demonstrates our DoPL's ability to capture fine-grained cross-modal relationships, enhancing alignment precision and efficiency.

Method	# Tunable	VQAv2		Method	# Tunable	MSRVTT-QA
		test-dev	test-std			test acc
Methods with full fine-tuning:						
VL-T5/BART(Cho et al., 2021)	165M	-	71.30	ClipBERT(Lei et al., 2021)	135M	37.4
SOHO(Huang et al., 2021)	155M	73.25	73.47	ALPRO(Li et al., 2022a)	245M	42.1
OSCAR(Li et al., 2020)	330M	73.61	73.82	Just-Ask(Yang et al., 2021)	200M	41.5
UNITER(Chen et al., 2020)	330M	73.82	74.03	VIOLET(Fu et al., 2021)	306M	43.9
ALBEF(Li et al., 2021)	266M	75.84	76.04	MERLOT(Zellers et al., 2021)	233M	43.1
BLIP(Li et al., 2022b)	337M	77.44	77.48	All-in-one(Wang et al., 2023b)	110M	44.3
Methods with frozen backbone:						
LoRA(r=32)(Hu et al., 2022)	10.6M	75.1	75.2	LoRA(r=32)(Hu et al., 2022)	10.6M	44.0
Maple(len=8)(Khattak et al., 2023)	5.4M	75.28	75.67	Maple(len=8)(Khattak et al., 2023)	5.4M	44.3
Bloom(len=8)(Wang et al., 2024)	3.54M	75.92	76.12	Bloom(len=8)(Wang et al., 2024)	3.54M	44.4
DoRA(r=32)(Liu et al., 2024)	10.8M	75.89	76.17	DoRA(r=32)(Liu et al., 2024)	10.8M	44.4
UniAdapter(r=512)(Lu et al., 2023)	19.5M	75.44	75.56	UniAdapter(r=512)(Lu et al., 2023)	19.5M	44.7
Aurora(r=64)(Wang et al., 2023a)	0.6M	77.69	77.87	Aurora(r=64)(Wang et al., 2023a)	0.6M	**44.8**
DoPL(ours, len=8)	**0.25M**	**77.71**	**77.9**	DoPL(ours, len=8)	0.25M	**44.8**

Table 2: Results on visual question answering datasets VQAv2 and MSRVTT-QA.

Method	Input	# Tunable	MSR-VTT				Didemo			
			R@1	R@5	R@10	MdR	R@1	R@5	R@10	MdR
Methods with full fine-tuning:										
ClipBERT(Lei et al., 2021)	16x448	135M	22.0	46.8	59.9	6.0	20.4	48.0	60.8	6.0
Frozen in Time(Bain et al., 2021)	32x224	180M	31.0	59.5	70.5	3.0	34.6	65.0	74.7	3.0
ALPRO(Li et al., 2022a)	8x224	245M	33.9	60.7	73.2	3.0	35.9	67.5	78.8	3.0
VIOLET(Fu et al., 2021)	5x224	306M	34.5	63.0	73.4	-	32.6	62.8	74.7	-
All-in-one(Wang et al., 2023b)	9x224	110M	37.9	68.1	77.1	-	32.7	61.4	73.5	3.0
CLIP-Hhiker(Bain et al., 2022)	120x224	124M	47.7	74.1	82.9	-	-	-	-	-
Methods with frozen backbone:										
CLIP-Prompt(Ju et al., 2022)	16x224	6.4M	36.7	64.6	-	-	-	-	-	-
LoRA(r=32)(Hu et al., 2022)	16x224	10.6M	39.9	66.9	76.9	3.0	33.8	61.2	71.32	2.0
Maple(len=8)(Khattak et al., 2023)	16x224	5.4M	40.2	67.2	77.2	3.0	32.6	60.2	71.32	2.0
Bloom(len=8)(Wang et al., 2024)	16x224	3.54M	44.5	69.8	79.7	2.0	35.49	61.7	72.38	2.0
DoRA(r=32)(Liu et al., 2024)	16x224	10.8M	45.3	70.4	80.1	2.9	35.8	63.4	73.21	2.0
UniAdapter*(r=512)(Lu et al., 2023)	16x224	19.5M	42.4	68.4	77.4	2.0	32.8	60.02	71.19	2.0
Aurora*(r=64)(Wang et al., 2023a)	16x224	0.6M	45.1	69.7	79.4	2.0	35.59	63.61	73.08	2.0
DoPL(ours, len=8)	16x224	**0.25M**	**47.8**	**74.4**	**84.0**	2.0	**39.98**	**65.5**	**74.87**	2.0

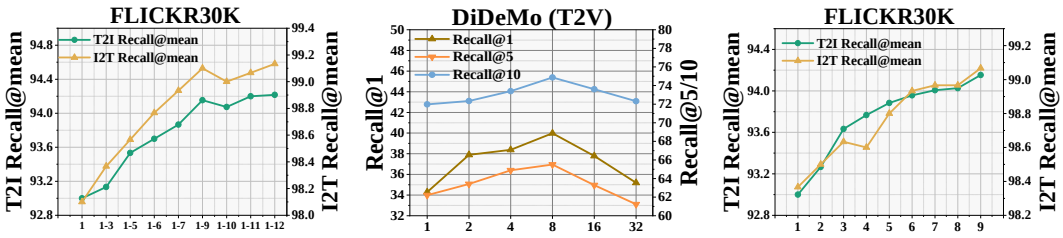
Table 3: Results on video-text retrieval datasets MSR-VTT and Didemo (Text->Video), '*' donate the performance obtained by (Ji et al., 2024).

Video-text Retrieval. Table 3 presents the performance of our framework on video-text retrieval using the MSRVTT and DiDemo datasets. Our DoPL outperforms all frozen backbone methods using significantly fewer parameters(0.25M). It also demonstrates strong performance compared to full fine-tuning methods like VIOLET(306M) and CLIP-Hhiker(124M). The DPG module enhances the semantic sensitivity of the prompts, enabling our model to capture fine-grained cross-modal relationships in video-text tasks. This highlights the model's superior ability in fine-grained semantic alignment and its efficiency in handling complex multimodal data.

4.3 Analysis Experiment

Variants of Inserting Multi-modal Prompts. We conducted ablation experiments to identify which modality's prompt most contributes to cross-modal learning, as shown in Table 4. The results demonstrate that inserting prompts into the multi-modal encoder improves performance over individual text or image prompts, highlighting the importance of multimodal alignment. Furthermore, the best performance is achieved by inserting prompts into all modality encoders, which we adopt as the default configuration.

Impact of Prompts Depth. We evaluate the impact of adding prompts to different encoder layers.



(a) Prompt Layers Depth **(b) Prompt Length** **(c) DPG Layers Depth**

Figure 3: Experiment on prompt depth (**a**), prompt length (**b**) and DPG layers depth (**c**). The results are divided on two y-axes for clear demonstration, where T2I Recall@mean/Recall@1 is on the left axis and I2T Recall@mean/Recall@5/10 is on the right one.

V	T	M	# Tunable	R@1	R@5	R@10	MedR
✓			0.08M	36.49	62.3	71.58	2.0
	✓		0.08M	36.69	63.2	71.54	2.0
✓	✓		0.16M	38.49	63.51	73.78	2.0
		✓	0.08M	38.19	63.42	73.68	2.0
✓	✓	✓	0.25M	**39.98**	**65.5**	**74.87**	2.0

Table 4: Insering modality-specific prompt (Visual V, Textual T and Multimodal M) on Didemo(Text->Video)

As shown in Figure 3 (a), performance improves with the number of layers but plateaus after nine layers. These results demonstrate the effectiveness of detail-oriented prompt insertion for cross-modal alignment, and we adopt nine layers as the default configuration to balance performance and efficiency.

Impact of Prompts Length. We also evaluate the impact of prompt length on performance. As shown in Figure 3 (b), performance improves with increasing prompt length up to 8, but further increases in length result in diminished returns and even a performance drop. We select a prompt length of 8 to balance performance and efficiency as the default setting.

DPG Module Visualization Validation. To verify the role of the DPG module's low-entropy regions in enhancing cross-modal alignment, we compared the attention distribution with and without the DPG module using Grad-CAM in Figure 4. The results show that the DPG module enables the model to capture finer image details, improving text-vision alignment accuracy.

Layer-wise Progressive Alignment Validation. We analyze the impact of progressively inserting the DPG module across layers and find consistent performance improvement with more layers integrated, as shown in Figure 3 (c). The experimental results demonstrate that the DPG module enhances the prompts and facilitates the model in progressively locating cross-modal alignment information across layers, thereby improving overall performance.

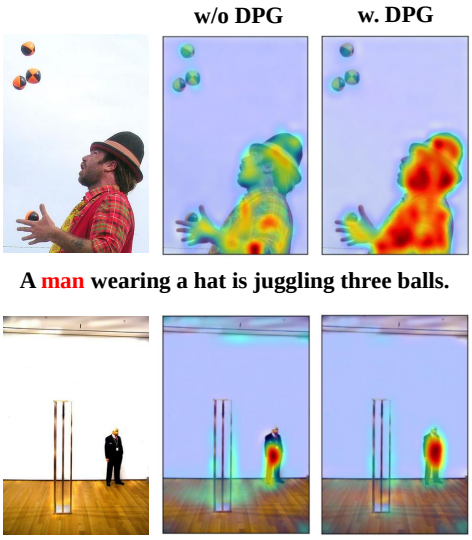


Figure 4: Cross-attention map visualizations on Flickr30K highlight the DPG's ability to pinpoint visually relevant regions for specific text words.

Generalization Capability Validation. To assess the generalization capability of our method, we applied it to various vision-language models (VLMs), including CLIP and ALBEF, as well as larger backbones like BLIP-Large and BEiT 3-Large. We conducted cross-model experiments on the FLICKR30K dataset. The results in Table 5 indicate that our approach attains performance levels comparable to, or surpassing, those of several full fine-tuning VLMs. This indicates that our method exhibits strong transferability and can enhance performance in other VLMs with fewer parameters.

Training Efficiency and Storage Cost. The proposed DoPL method achieves remarkable efficiency through its parameter-efficient design and non-parametric cross-modal alignment strategy. As shown in Table 6, we compare training time effi-

model	#Tunable	FLICKR30K TR			FLICKR30K IR		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP(Radford et al., 2021) (Full finetuning)	151.3M	69.6	90.6	95	67	89.1	93.6
DoPL(len=8)	0.16M	69.8	90.7	95	66.9	89	93.6
ALBEF(Li et al., 2021) (Full finetuning)	419.3M	95.9	99.8	100	85.6	97.5	98.9
DoPL(len=8)	0.18M	95.7	99.9	100	85	97.4	98.9
BLIP-Large(Li et al., 2022b) (Full finetuning)	591M	97.4	99.8	99.9	87.6	97.7	99
DoPL(len=8)	0.5M	97.4	100	100	87.1	97.7	99.1
BEIT 3-Large(Wang et al., 2022) (Full finetuning)	746M	97.1	100	100	87.5	97.9	99.1
DoPL(len=8)	0.5M	96.8	99.9	100	87.2	97.8	99

Table 5: Transferring our method to various backbones and evaluating on Flickr30K.

Method	# Param	MSCOCO		FLICKR30k		MSRVTT		Didemo		VQA		VideoQA	
		Time	Memory	Time	Memory	Time	Memory	Time	Memory	Time	Memory	Time	Memory
Full fine-tuning	223M/337M	1.28	1.4	1.25	1.39	-	-	-	-	2.1	1.92	-	-
UniAdapter	19.5M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Adapter	28.9M	1.05	1.06	1.02	1.04	1.08	1.03	1.04	1.02	1.02	1.03	1.05	1.06
DoRA	10.8M	0.90	0.92	0.92	0.93	0.95	0.94	0.93	0.93	0.91	0.90	0.92	0.91
LoRA	10.6M	0.87	0.90	0.90	0.91	0.94	0.91	0.92	0.91	0.90	0.89	0.89	0.89
Maple	5.4M	0.84	0.87	0.87	0.89	0.90	0.89	0.88	0.89	0.87	0.87	0.86	0.87
Aurora	0.6M	0.95	0.97	0.97	0.96	0.97	0.96	0.96	0.95	0.97	0.98	0.95	0.92
DoPL	**0.25M**	**0.81**	**0.85**	**0.84**	**0.86**	**0.85**	**0.88**	**0.85**	**0.87**	**0.84**	**0.86**	**0.85**	**0.86**

Table 6: Comparison on the training time and GPU memory (Relative GPU hours and memory cost are normalized by UniAdapter (set as 1.00) and the backbone is BLIP).

ciency and DoPL achieves 0.25M tunable parameters (0.12% of full fine-tuning) while maintaining optimal training speed and memory consumption across diverse tasks. For example, on computation-intensive VQA tasks, it reduces training time to 0.84× and memory usage to 0.86× relative to UniAdapter (set as baseline 1.00), outperforming state-of-the-art methods like Aurora (0.6M) and Maple (5.4M). This efficiency is enabled by a dual mechanism: first, through an ultra-low parameter footprint achieved by freezing the backbone network and appending lightweight prompt vectors to attention layers; second, via non-parametric dynamic alignment that concatenates prompts to Key/Value matrices, preserving hardware-friendly parallelism in multi-head attention while enabling layer-wise fine-grained alignment under the low-entropy information concentration theory.

In summary, DoPL optimally balances parameter efficiency and training resource consumption, offering a scalable solution for complex multimodal tasks without compromising accuracy. These advantages are further validated across MSCOCO, FLICKR30k, MSRVTT, Didemo, VQA, and VideoQA benchmarks in Table 6, where DoPL consistently achieves the lowest GPU hours and memory cost among all parameter-efficient methods.

5 Conclusion

This paper proposes a novel detail-oriented prompt learning method for VLMs, achieving both parameter efficiency and fine-grained semantic alignment. DoPL implements layer-by-layer localization of fine-grained semantic alignment with merely 0.25M trainable parameters by incorporating the detail-oriented prompt generation module. This method enhances fine-grained alignment between vision and text information while maintaining computational efficiency. Extensive evaluations on six cross-modal benchmarks demonstrate that DoPL consistently outperforms previous PETL methods and surpasses full fine-tuning methods.

Limitations

While our method effectively enhances the model's ability to capture fine-grained semantic alignment, it experiences slower convergence on larger datasets than full fine-tuning like other PETL methods. Future work could focus on optimizing convergence speed through more efficient optimization strategies while maintaining strong alignment performance.

Acknowledgments

This work was funded in part by the National Natural Science Foundation of China grant under number 62222603, in part by the STI2030-

Major Projects grant from the Ministry of Science and Technology of the People's Republic of China under number 2021ZD0200700, in part by the Key-Area Research and Development Program of Guangdong Province under number 2023B0303030001, in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2019ZT08X214), and in part by the Science and Technology Program of Guangzhou under number 2024A04J6310 supported by "the Fundamental Research Funds for the Central Universities".

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2022. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*.

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

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*.

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

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. 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.

Yuheng Ji, Yue Liu, Zhicheng Zhang, Zhao Zhang, Yuting Zhao, Gang Zhou, Xingwei Zhang, Xinwang Liu, and Xiaolong Zheng. 2024. Advlora: Adversarial low-rank adaptation of vision-language models. *arXiv preprint arXiv:2404.13425*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022a. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963.

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation.

Advances in neural information processing systems, 34:9694–9705.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo, and Fei Tan. 2023. Deeply coupled cross-modal prompt learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7957–7970.

Haoyu Lu, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. 2023. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. In *The Twelfth International Conference on Learning Representations*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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

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

Haixin Wang, Xinlong Yang, Jianlong Chang, Dian Jin, Jinan Sun, Shikun Zhang, Xiao Luo, and Qi Tian. 2023a. Parameter-efficient tuning of large-scale multimodal foundation model. *Advances in Neural Information Processing Systems*, 36:15752–15774.

Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023b. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.

Qiang Wang, Ke Yan, and Shouhong Ding. 2024. Bilateral adaptive cross-modal fusion prompt learning for clip. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9001–9009.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. 2024. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Bruce XB Yu, Jianlong Chang, Haixin Wang, Lingbo Liu, Shijie Wang, Zhiyu Wang, Junfan Lin, Lingxi Xie, Haojie Li, Zhouchen Lin, et al. 2024. Visual tuning. *ACM Computing Surveys*, 56(12):1–38.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34:23634–23651.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

A Theoretical Proof of Low-entropy Information Concentration Theory

To rigorously justify the proposed DoPL method, we provide the necessary theoretical proof establishing the relationship between low-entropy regions and optimal cross-modal alignment. Specifically, we prove that minimizing the conditional entropy of token-level associations between vision and text modalities is mathematically equivalent to maximizing their mutual information, thereby enforcing fine-grained semantic alignment. This theorem not only supports the low-entropy information concentration theory introduced in our method but also guarantees that the derived alignment weights (e.g., $W_i^{t\to v} \propto \exp(-H_i^{t\to v})$) can effectively enhance discriminative cross-modal interactions.

In the following, we formally present and prove the theorem, demonstrating that strong alignment corresponds to low entropy and high mutual information, which serves as the foundation for DPG's ability to localize and amplify fine-grained semantic correspondences.

A.1 Preliminaries

Given the text prompt $M^t \in \mathbb{R}^{l\times d}$ and the vision prompt $M^v \in \mathbb{R}^{l\times d}$, perform matrix multiplication to obtain the cross-modal association matrix $R_{i,j} = (M^t)^T \cdot M^v \in \mathbb{R}^{l\times l}$ (where $R_{i,j}$ represents the association strength between text token x_i and visual token y_j), and define the normalized association matrices $N_{i,j}^{t\to v} = \frac{\exp(R_{i,j})}{\sum_{k=1}^{n}\exp(R_{k,j})}$ and $N_{i,j}^{v\to t} = \frac{\exp(R_{i,j})}{\sum_{k=1}^{m}\exp(R_{i,k})}$ (where the former corresponds to the conditional probability $P(x_i|y_j)$, and the latter corresponds to the conditional probability $P(y_j|x_i)$), as well as the joint probability distribution $P(x_i, y_j) = \frac{\exp(R_{i,j})}{\sum_{i=1}^{n}\sum_{j=1}^{m}\exp(R_{i,j})}$.

Based on the aforementioned joint probability distribution, define the marginal distributions $P(x_i) = \sum_{j=1}^{m} P(x_i, y_j)$ and $P(y_j) = \sum_{i=1}^{n} P(x_i, y_j)$ (where the former is the marginal distribution of the text token x_i, and the latter is the marginal distribution of the image token y_j).

For the text token x_i and the image token y_j, define their respective conditional entropies as:

$$H_{x_i}^{t\to v} = -\sum_{j=1}^{m} N_{i,j}^{t\to v} \log N_{i,j}^{t\to v} = H(Y|x_i), \tag{1}$$

$$H_{y_j}^{v\to t} = -\sum_{i=1}^{n} N_{i,j}^{v\to t} \log N_{i,j}^{v\to t} = H(X|y_j), \tag{2}$$

To ensure the equivalence between the minimization of the cross-modal conditional entropy and the maximization of mutual information, the following theorem is proposed.

A.2 Theorem

Theorem: When the conditional entropy from text-to-image or image-to-text reaches its minimum value, the mutual information $I(X;Y)$ between the random variable of the text modality X and the random variable of the visual modality Y reaches its maximum value. Moreover, the joint probability distribution $P(x_i, y_j)$ of the text and visual modalities exhibits a strictly diagonal trend (i.e., semantic alignment), and the following properties hold:

- **Global Minimization of Conditional Entropy:** When both the text-to-visual conditional entropy $H(Y|X)$ and visual-to-text conditional entropy $H(X|Y)$ reach their global minima, the reduction of any x_i's conditional entropy $H(Y|x_i)$ directly corresponds to the increase of its local mutual information $I(x_i;Y)$ with the visual modality. Similarly, the decrease of any y_j's conditional entropy $H(X|y_j)$ is equivalent to the enhancement of its local mutual information $I(X;y_j)$with the textual modality.

- **Synergistic Optimization of Joint and Marginal Distributions:** The optimal cross-modal association strength $R_{i,j}$ satisfies $R_{i,j} = \log P(x_i) + \log P(y_j) + C$, where the constant ($C$) controls the deviation between the joint distribution and the independent distributions. When ($C > 1$), the joint distribution $P(x_i, y_j)$ tends to become diagonal (i.e., strict alignment between x_i and y_j), and the extremum point of the mutual information $I(X;Y)$ corresponds to the global minimization of the conditional entropy.

A.3 Proof

Obtain the conditional entropy based on the marginal distributions:

$$\begin{aligned} H(Y|X) &= \sum_{i=1}^{n} P(x_i) H(Y|x_i) \\ &= \sum_{i=1}^{n} P(x_i) H_{x_i}^{t \to v}, \end{aligned} \tag{3}$$

$$\begin{aligned} H(X|Y) &= \sum_{j=1}^{m} P(y_j) H(X|y_j) \\ &= \sum_{j=1}^{m} P(y_j) H_{y_j}^{v \to t}, \end{aligned} \tag{4}$$

The expression for mutual information can be obtained based on both the marginal distributions and the joint probability distribution, which is used to measure cross-modal dependency relationships:

$$I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i) P(y_j)}, \tag{5}$$

Meanwhile, the relationship between conditional entropy and mutual information can be derived from the fundamental equation of information theory:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y), \end{aligned} \tag{6}$$

To further investigate the relationship between the maximum mutual information and the minimum conditional entropy, we solve for the maximum mutual information using Lagrange functions. The detailed derivation is given below.

First, given the joint probability distribution:

$$P(x_i, y_j) = \frac{\exp(R_{i,j})}{\sum_{k=1}^{n} \sum_{l=1}^{m} \exp(R_{k,l})}, \tag{7}$$

where $R_{i,j}$ are the free parameters to be optimized.

Substitute the simplified joint probability distribution into the definition of mutual information and expand it:

$$\begin{aligned} & I(X;Y) \\ &= \sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i) P(y_j)} \\ &= \sum_{i,j} \frac{\exp(R_{i,j})}{\sum_{k=1}^{n} \sum_{l=1}^{m} \exp(R_{k,l})} \\ &\quad \Big[R_{i,j} - \log \left(\sum_{k=1}^{n} \sum_{l=1}^{m} \exp(R_{k,l}) \right) \\ &\quad - \log \left(\sum_{l=1}^{m} \frac{\exp(R_{i,l})}{\sum_{k=1}^{n} \sum_{l=1}^{m} \exp(R_{k,l})} \right) \\ &\quad - \log \left(\sum_{k=1}^{n} \frac{\exp(R_{k,j})}{\sum_{k=1}^{n} \sum_{l=1}^{m} \exp(R_{k,l})} \right) \Big], \end{aligned} \tag{8}$$

Further simplify it to:

$$\begin{aligned} & I(X;Y) \\ &= \frac{1}{\sum_{k=1}^{n} \sum_{l=1}^{m} \exp(R_{k,l})} \sum_{i,j} \exp(R_{i,j}) \\ &\quad \Bigg[R_{i,j} - \log(\sum_{k=1}^{n} \sum_{l=1}^{m} \exp(R_{k,l})) \\ &\quad - \log P(x_i) - \log P(y_j) \Bigg], \end{aligned} \tag{9}$$

To solve for the maximum mutual information, introduce the constraint condition $\sum_{i,j} \exp(R_{i,j}) = Z$ and construct the Lagrange function:

$$\begin{aligned} \mathcal{L} = I(X;Y) - \lambda \left(\sum_{i,j} \exp(R_{i,j}) - Z \right) \\ (Z = \sum_{i,j} \exp(R_{i,j})), \end{aligned} \tag{10}$$

Take the derivative of the free parameter $R_{i,j}$ to be optimized, decomposing it into the derivative of the mutual information term and the derivative of the constraint term, where the derivative of the mutual information term is:

$$\begin{aligned} \frac{\partial I}{\partial R_{i,j}} = \frac{\exp(R_{i,j})}{Z} \Bigg[1 + R_{i,j} - \log Z \\ - \log P(x_i) - \log P(y_j) \Bigg], \end{aligned} \tag{11}$$

The derivative of the constraint term is:

$$\frac{\partial}{\partial R_{i,j}}\left[-\lambda\left(\sum_{k,l}\exp(R_{k,l})-Z\right)\right] = -\lambda\exp(R_{i,j}), \tag{12}$$

Combining Equations 11 and 12, we obtain the partial derivative of Equation 10 with respect to $R_{i,j}$ and set it to zero:

$$\frac{\partial I}{\partial R_{i,j}} = \frac{\exp(R_{i,j})}{Z}\Big[1 + R_{i,j} - \log Z - \log P(x_i) - \log P(y_j)\Big] - \lambda\exp(R_{i,j}) = 0, \tag{13}$$

Further simplifying the extremum condition, multiply both sides of Equation 13 by $Z/\exp(R_{i,j})$ to obtain:

$$1 + R_{i,j} - \log Z - \log P(x_i) - \log P(y_j) - \lambda Z = 0, \tag{14}$$

By setting $\lambda Z = C$ (constant), we obtain:

$$R_{i,j} = \log P(x_i) + \log P(y_j) + (C-1), \tag{15}$$

This indicates that the optimal association strength $R_{i,j}$ is related to the logarithm of marginal probabilities. Substituting Equation 15 into Equation 7 yields:

$$P(x_i, y_j) \propto \exp(R_{i,j}) = P(x_i)P(y_j)e^{C-1}, \tag{16}$$

When $C = 1$, we obtain the independent distribution $P(x_i, y_j) = P(x_i)P(y_j)$, where the mutual information $I(X;Y) = 0$(minimum).

When $C > 1$, the joint distribution deviates from independence, and the mutual information increases. The maximum mutual information occurs when$P(x_i, y_j)$ exhibits a diagonal distribution (i.e., strict alignment between x_i and y_j), at which point the conditional entropy $H(Y|x_i)$ (or $H(X|y_j)$) is minimized, verifying that low-entropy regions correspond to strong alignment.

For Property 1, according to Equation (6) (the relationship between conditional entropy and mutual information), when $H(X)$ and $H(Y)$ are fixed, a decrease in conditional entropy $H(X|Y)$ or $H(Y|X)$ corresponds to an increase in mutual information $I(X;Y)$. Furthermore, Property 1 (Global Minimization of Conditional Entropy) can be derived from Equations 7-8. Specifically:

- When $H_{x_i}^{t\to v}$ decreases, $H(Y|X)$ decreases and $I(X;Y)$ increases, indicating that the text token x_i carries more information about Y (the image).
- When $H_{y_j}^{v\to t}$ decreases, $H(X|Y)$ decreases and $I(X;Y)$ increases, indicating that the image token y_j carries more information about X (the text).

For Property 2, Equation 15 shows that the optimal association strength $R_{i,j}$ is related to the sum of log marginal probabilities. Equation 16 further reveals the relationship between joint and marginal probabilities, leading to Property 2 (Synergistic Optimization of Joint and Marginal Distributions). Specifically:

- The optimal cross-modal association strength $R_{i,j}$ satisfies $R_{i,j} = \log P(x_i) + \log P(y_j) + C$, where the constant C controls the deviation between joint and independent distributions.
- When $C > 1$, the joint distribution $P(x_i, y_j)$ tends toward diagonality (strict alignment between x_i and y_j), and the extremum of $I(X;Y)$ corresponds to global minimization of conditional entropy.

□

Therefore, according to the aforementioned theorem and properties, low entropy values $H_{x_i}^{t\to v}$ directly correspond to high mutual information $I(x_i;Y)$ (similarly, low entropy values $H_{y_j}^{v\to t}$ directly correspond to high mutual information $I(X;y_j)$), proving that low-entropy regions in cross-modal associations indicate stronger semantic alignment. This theory provides the mathematical foundation for the weight allocation in the DPG module: the weight allocation $W_i^{t\to v} \propto \exp(-H_i^{t\to v})$ or $W_j^{v\to t} \propto \exp(-H_j^{v\to t})$ is mathematically equivalent to reinforcing high mutual information regions. By minimizing $H_i^{t\to v}$ and $H_j^{v\to t}$, the model can focus on fine-grained alignment signals, thereby improving multimodal task performance.