Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.059745

ARTICLE

UniTrans: Unified Parameter-Efficient Transfer Learning and Multimodal Alignment for Large Multimodal Foundation Model

Jiakang Sun[1,2], **Ke Chen**[1,2], **Xinyang He**[1,2], **Xu Liu**[1,2], **Ke Li**[1,2] **and Cheng Peng**[1,2,*]

[1]Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, 610213, China
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, 101499, China

*Corresponding Author: Cheng Peng. Email: pengchengcasit@163.com

Received: 16 October 2024; Accepted: 08 January 2025; Published: 26 March 2025

ABSTRACT: With the advancements in parameter-efficient transfer learning techniques, it has become feasible to leverage large pre-trained language models for downstream tasks under low-cost and low-resource conditions. However, applying this technique to multimodal knowledge transfer introduces a significant challenge: ensuring alignment across modalities while minimizing the number of additional parameters required for downstream task adaptation. This paper introduces UniTrans, a framework aimed at facilitating efficient knowledge transfer across multiple modalities. UniTrans leverages Vector-based Cross-modal Random Matrix Adaptation to enable fine-tuning with minimal parameter overhead. To further enhance modality alignment, we introduce two key components: the Multimodal Consistency Alignment Module and the Query-Augmentation Side Network, specifically optimized for scenarios with extremely limited trainable parameters. Extensive evaluations on various cross-modal downstream tasks demonstrate that our approach surpasses state-of-the-art methods while using just 5% of their trainable parameters. Additionally, it achieves superior performance compared to fully fine-tuned models on certain benchmarks.

KEYWORDS: Parameter-efficient transfer learning; multimodal alignment; image captioning; image-text retrieval; visual question answering

1 Introduction

The current paradigm in artificial intelligence has shifted from developing domain-specific models to pretraining large models on extensive datasets, followed by fine-tuning for downstream tasks [1]. This shift has led to the development of several prominent large pre-trained models, such as LLaMA [2], SAM [3] and BLIP [4]. However, as the number of parameters in foundational pre-trained models continues to grow (such as the 175B parameters in GPT-3 [5]), the computational and storage resources required for full-parameter fine-tuning have increased significantly. Consequently, it has become crucial to identify methods that strike an effective balance between cost efficiency and fine-tuning performance.

Transfer learning has effectively addressed the challenge of applying knowledge from one task to another related task, enhancing learning efficiency and generalization ability [6,7]. In particular, parameter-efficient transfer learning has attracted particular attention, as it enables knowledge transfer by adding only a small number of additional training parameters, thereby drastically reducing computational and storage requirements [8–10]. Current mainstream efficient-parameter transfer methods can be primarily categorized into prompt tuning, adapter tuning, and selective tuning. These methods either freeze the backbone and fine-tune only the added extra parameters or select a small subset of parameters from the backbone for training.

This approach significantly reduces computational costs while achieving performance comparable to full fine-tuning. This method has demonstrated substantial success in both natural language processing and computer vision. Recently, there has been growing interest in multimodal foundational models. However, many of these studies either target only a single downstream task, neglecting the diverse range of tasks in multimodal models [11,12], or they directly add learnable parameters using unimodal models' parameter-efficient fine-tuning methods ignoring the alignment between different modalities [13]. Alternatively, some introduce excessive redundant parameters [14]. These approaches fail to solve the core problem in parameter-efficient fine-tuning for multimodal models: ***achieving downstream knowledge transfer with minimal additional parameters while ensuring proper alignment between modalities*.**

To address this issue, we design a novel and effective framework, UniTrans, to facilitate cross-modal knowledge transfer. Although efficient parameter transfer learning techniques, which involve adding extra parameters, have been widely applied in natural language processing and computer vision, there is still a lack of sufficient exploration in the multimodal domain. Based on extensive experiments, we propose **vector-based cross-modal random matrix adaptation (VCRA)**, which leverages Low-Rank Adaptation (LoRA) to decompose low-rank matrices into learnable scaling vectors and shared low-rank random matrices. VCRA employs a pair of random matrices for weight sharing, allowing fine-grained information between the image and text modalities to interact, thereby enhancing the visual-language modality representation.

Furthermore, during the fine-tuning process of multimodal base models on downstream tasks, the originally aligned image and text features may become disrupted, causing them to shift within their respective feature domains. To address this, we design the **multimodal consistency alignment module (MCAM)** and the **query-augmentation side network (QASN)**, which serve as regularizers for feature alignment during the fine-tuning process. MCAM, from a contrastive learning perspective, constrains the similarity ranking consistency between image-text pairs by designing a simple yet effective loss function without introducing additional parameters. At the same time, we observe that when the fusion network of a multimodal model has too many layers, query information loss occurs, which impacts the fusion and alignment between modalities. To resolve this, we propose the lightweight QASN, which adaptively supplements query information at various layers of the fusion network through a side-network approach, preventing matching errors between images and text caused by information loss. Finally, we evaluate our method on multiple cross-modal benchmarks, and the results show that our approach requires only 5% of the training parameters compared to state-of-the-art methods, significantly reducing training costs while outperforming traditional methods. Additionally, when compared to full-parameter training methods, our approach achieves comparable or even better performance (Fig. 1).

In summary, our contributions can be summarized as follows:

1) We propose a lightweight and effective framework, UniTrans, for efficient cross-modal parameter knowledge transfer. The framework consists of VCRA and two modules, MCAM and QASN, designed to enhance modality alignment during the fine-tuning process.
2) Based on Low-Rank Adaptation, we design a more suitable adapter for multimodal models, VCRA, which facilitates modality interaction and modality-specific adaptation through shared random matrices and learnable scaling vectors.
3) We design two modules, MCAM and QASN, to constrain modality alignment, further improving the performance of multimodal downstream tasks.
4) We conduct experiments on multiple multimodal benchmarks. The results show that our method reduces the trainable parameters to 5% compared to state-of-the-art method without sacrificing performance, and it significantly outperforms traditional methods. These benchmark test results are of significant importance for future research.

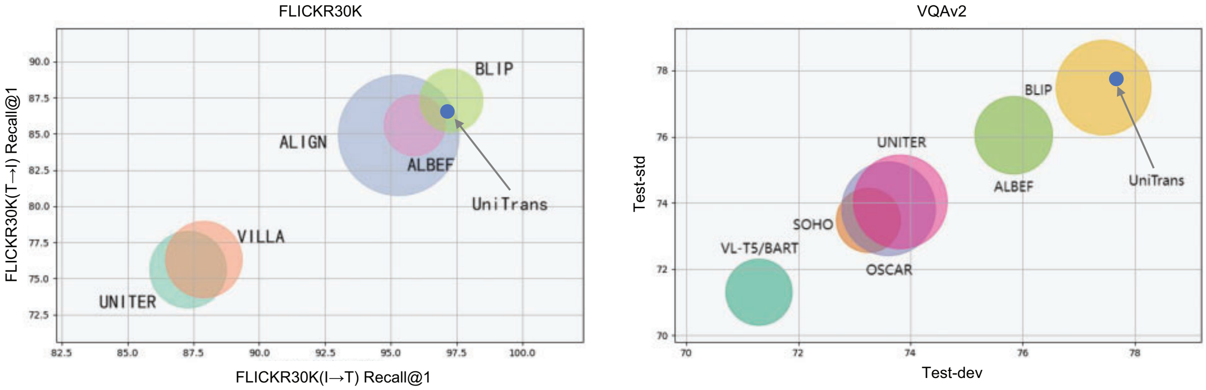

Figure 1: Performance comparison between image-text retrieval (left) and VQA (right) tasks. The size of the bubble represents the number of trainable parameters. Our UniTrans has achieved competitive performance in both tasks

2 Related Work

2.1 Vision-Language Models

In recent years, increasing efforts have focused on applying Vision-Language Models (VLMs) pre-trained with large-scale image-text pairs to downstream tasks [15,16]. Unlike pre-trained large language models [5,17,18], VLMs typically extract multimodal features through separate text and image encoders, and then align these features using fusion mechanisms such as contrastive learning [15], transformer modules [4], Q-Former [19], or MLP [20]. The fused features are applied to multimodal downstream tasks. BLIP integrates both an encoder and a decoder, enabling support for both multimodal alignment and multimodal generation tasks within a single foundational model. This paper explores efficient parameter transfer methods tailored for multimodal models based on the BLIP framework.

2.2 Parameter-Efficient Transfer Learning

Cross-modal alignment refers to establishing correspondences between information from different modalities, enabling machines to recognize and understand the same or related information across various modalities. Current cross-modal alignment methods can be broadly categorized into attention-based alignment [21–23], large cross-modal model-based alignment [24,25], parameter-free interaction-based alignment [26–28], and structure-based alignment [29]. However, these methods are typically applied during the model training process. As the number of model parameters grows, the pre-training and fine-tuning paradigm has become predominant, highlighting the need for inter-modal alignment mechanisms specifically designed for the fine-tuning process. Based on this, we have designed Query-Augmentation Side Network and Multimodal Consistency Alignment Module, which can serve as regularizers for fine-grained cross-modal alignment during the fine-tuning process.

As the number of parameters in foundational pre-trained models continues to grow, the cost of full-parameter fine-tuning for downstream tasks has become increasingly prohibitive, drawing more attention from the engineering community to parameter-efficient transfer learning. This approach facilitates knowledge transfer in downstream tasks by adding a small number of additional parameters and can be broadly categorized into the following types: prompt tuning [10,30,31], adapter tuning [9,8,32,33], and selective tuning [34,35]. While these methods have achieved significant success in the field of natural language processing (NLP), they remain underexplored in the multimodal domain. Existing work either applies these methods to multimodal models without accounting for the alignment between different modalities or focuses solely on a single downstream task. A recent study, UniAdapter [14], pioneered parameter-efficient transfer learning for multimodal models but introduced excessive redundant parameters. In contrast, our proposed method, UniTrans, reduces the number of parameters by an order of magnitude, while delivering comparable or even superior performance across multiple downstream tasks.

2.3 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation is a type of parameter-efficient transfer learning that involves adapter fine-tuning. Unlike methods that add adapters, LoRA approximates weight changes through low-rank matrices during fine-tuning, allowing it to merge seamlessly with pre-trained weights during inference without introducing extra computational overhead. This significantly reduces the computational and storage resources required for tuning, providing an innovative solution for large pre-trained models. Base on this, AdaLoRA [36] uses Singular Value Decomposition (SVD) decomposition during fine-tuning to selectively remove insignificant singular values, dynamically adjusting the rank of the low-rank matrix for more efficient updates. Tied-LoRA [37] further reduces trainable parameters by using weight tying. Dora [38] enhances LoRA's learning ability and training stability by decomposing pre-trained weights into magnitude and direction. FedPara [39] improves fine-tuning efficiency by introducing Hadamard product reparameterization weights into the low-rank matrix, breaking the low-rank limitation.

Although Low-Rank Adaptation (LoRA) and its variants significantly reduce the computational cost of fine-tuning large pre-trained language models, their potential in multimodal fine-tuning remains largely unexplored. Our work investigates the application of LoRA in multimodal parameter-efficient transfer learning. By introducing trainable scaling vectors and cross-modal shared low-rank weight matrices, we achieve efficient knowledge transfer with minimal trainable parameters, while ensuring effective feature alignment between modalities.

3 Methodology

In this section, we first describe the framework of the multimodal foundational model we aim to fine-tune and the Low-Rank Adaptation (LoRA) [8]. We then introduce our parameter-efficient transfer learning method for multimodal models, UniTrans. This includes Vector-based Cross-modal Random Matrix Adaptation (VCRA), modality alignment design, as well as the Query-Augmentation Side Network (QASN) and Multimodal Consistency Alignment Module (MCAM).

3.1 Preliminary

3.1.1 Vision-Language Framework

We use BLIP as the backbone of our frozen pre-trained model. BLIP features a multimodal hybrid encoder-decoder structure (MED), unifying image-text matching and generation tasks. It employs a Vision Transformer (ViT) [40] as the image encoder and BERT [18] as the text encoder, with different components activated depending on the downstream multimodal task. For image-text matching tasks, cross-attention is added to the text encoder to fuse image and text features, and a special token $[Encode]$ is prepended to the input text. For generation tasks, a $[Decoder]$ token is inserted at the beginning of the input text, and the bidirectional self-attention layers are replaced with causal self-attention to generate captions for the given image.

3.1.2 Low-Rank Adaptation (LoRA)

LoRA utilizes low-rank matrices to approximate the weight changes during fine-tuning, effectively reducing the number of required parameters. Formally, for a pre-trained weight matrix $W_0 \in R^{(m \times n)}$, the weight update can be constrained as a low-rank matrix decomposition, as shown in Eq. (2). During fine-tuning, the original weights W_0 remain frozen, and only the low-rank matrices are updated via gradient descent. Due to the low-rank nature, the dimension r is typically small, making the size of the low-rank matrices significantly smaller than that of the original parameter matrix, where $A \in R^{(r \times n)}$ and $B \in R^{(m \times r)}$, and $r \ll \min(m, n)$. While LoRA offers an effective solution for efficient parameter transfer, it has not been further explored in the context of multimodal models.

3.2 Vector-Based Cross-Modal Random Matrix Adaptation (VCRA)

For pre-trained multimodal models, performing full fine-tuning on downstream tasks with small-scale datasets not only wastes computational resources but also risks knowledge forgetting and disrupting the feature alignment space. Therefore, we introduce additional trainable parameters to minimize or limit changes to the original parameters as much as possible. The model parameters updated through backpropagation using the fine-tuning data D:

$$\nabla_W = \nabla_{\Delta W} = \frac{\partial L\left(D; W_0 + \Delta W\right)}{\partial \Delta W} \tag{1}$$

LoRA adapts the weight space of the entire network by fine-tuning a matrix product of two low-rank matrices. However, directly applying LoRA to multimodal models yields unsatisfactory results due to the lack of interaction between modalities. To address this, we decompose the matrix into two low-rank matrices and two scaling vector as the projection, share knowledge across these matrices between modalities, and then apply projections to adapt the weight matrices of each layer for each modality.

Formally, compared to LoRA, VCRA not only decomposes the trainable weights into a pair of low-rank matrices A and B, but also introduces two trainable scaling vectors:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{2}$$

$$h = W_0 x + \Delta W x = W_0 x + \Lambda_b B \Lambda_a A x \tag{3}$$

where trainable scaling vectors represented as diagonal matrices Λ_a and Λ_b. These vectors effectively scale or deactivate specific rows and columns of the random matrices. The random matrices $A \in R^{(r \times d)}$ and $B \in R^{(d \times r)}$ in VCRA are shared across visual, textual and cross-modal modalities, where d and r represent the input and bottleneck dimensions, respectively. This sharing mechanism not only reduces the number of parameters significantly but also enhances cross-modal interaction:

$$VCRA(x^M) = W_0 x^M + s \cdot \Lambda_b^M B \Lambda_a^M A x^M \tag{4}$$

where s represents the learnable scaling factor, $M \in \{V, T, C\}$, V denotes the visual modality, T denotes textual modality and C is cross-modal modality. Although we use shared random matrices for cross-modal information transfer, learning modality-specific knowledge is crucial for improving the transferability of multimodal models. Therefore, we apply modality-specific scaling vectors Λ_a^M and Λ_b^M to the visual, text encoders and feature fusion network with cross-attention layer, ensuring adaptability to each modality.

Compared to Lora, VeRA shares low rank matrices between different modalities and transformer layers and uses scaled vectors to adapt weight updates, greatly reducing the number of trainable parameters. Formally speaking, we use d_{model} to denote the dimension of finetuned layers and N_{tuned} to represent the number of these layers. The number of trainable parameters for VCRA can be expressed as $\Theta = 2 \times d_{\text{model}} \times r + N_{\text{tuned}} \times (d_{\text{model}} + r)$, contrasting with LoRA's $\Theta = 2 \times N_{\text{tuned}} \times d_{\text{model}} \times r$, when we apply petl to the FFN of each layer. Specifically, for the lowest rank (i.e., $r = 1$), trainable parameters of VCRA is about half of LoRA. However, as r insceases, the growth of LoRA trainable parameters is much faster than VCRA.

3.3 Modality Alignment Design

Besides VCRA, UniTrans also includes two additional modules, Query-Augmentation Side Network (QASN) and Multimodal Consistency Alignment Module (MCAM), to further enhance cross-modal alignment and multimodal fusion.

3.3.1 Query-Augmentation Side Network (QASN)

In the process of fine-tuning multimodal models, a factor that hinders modal alignment is that the multimodality fusion network is deep, which can lead to the loss of query information. To solve this problem, we designed a query information augmentation pipeline that runs parallel to the fusion network for adapted feature aggregation and information supplement.

Formally, as shown in Fig. 2, given the multimodality fusion branch network consists of N transformer blocks, the forward process can be expressed as $x^F = b_N(b_{N-1}(\ldots b_1(x^T, x^V), x^V), x^V)$, where b_N represents the N-th transformer block, x^V and x^T represent the image features and the text features, respectively, and x^F represents the fusion features. We apply VCRA to aggregate the fusion information from fusion branch network. Denote $w_i = \Lambda_b B \Lambda_a A$ as the weight matrix accounting for the i-th block, with A and B are the shared random matrices, and Λ_a and Λ_a are the scaling vectors. The query-augmentation side network gradually collects information from each block:

$$h_{i+1} = h_i + x_i^F w_i \tag{5}$$

where h_{i+1} is the output of the h-th layer of QASN. At the same time, we will supplement the query information flowing in QSAN back into the fusion network:

$$x_{i+1}^F = b_i(x_i^F, x^V) + f(h_{i+1}) \tag{6}$$

We did not take the query representation h as the residual of the fusion representation x^F, but instead train an information filter to explicitly model the contribution of supplementary query information. Our filter design is simple and effective, and can be represented as $f(x) = \boldsymbol{\sigma}(\boldsymbol{\alpha} \odot x + \boldsymbol{\beta})$. Here, α and β represent learnable parameter matrices, which are initialized to 1 and 0, respectively. Finally, the filtered information is processed through the activation function $\boldsymbol{\sigma}$. As a result, multimodality fusion branch network with supplementary query information is more conducive to transfer knowledge for downstream tasks.

3.3.2 Multimodal Consistency Alignment Module (MCAM)

In addition to using unsupervised contrastive learning during pre-training, BLIP also employs it during downstream task fine-tuning. As we know, with sufficient training samples, the contrastive loss function effectively brings positive pairs closer and pushes negative pairs farther apart. However, the scale of downstream task fine-tuning datasets is typically limited, which allows individual noisy samples to interfere with the feature representation process, potentially disrupting the previously aligned feature space. To address this issue, we introduce MCAM (Fig. 3), which captures fine-grained relationships between samples through rank consistency, mitigating the impact of noisy samples.

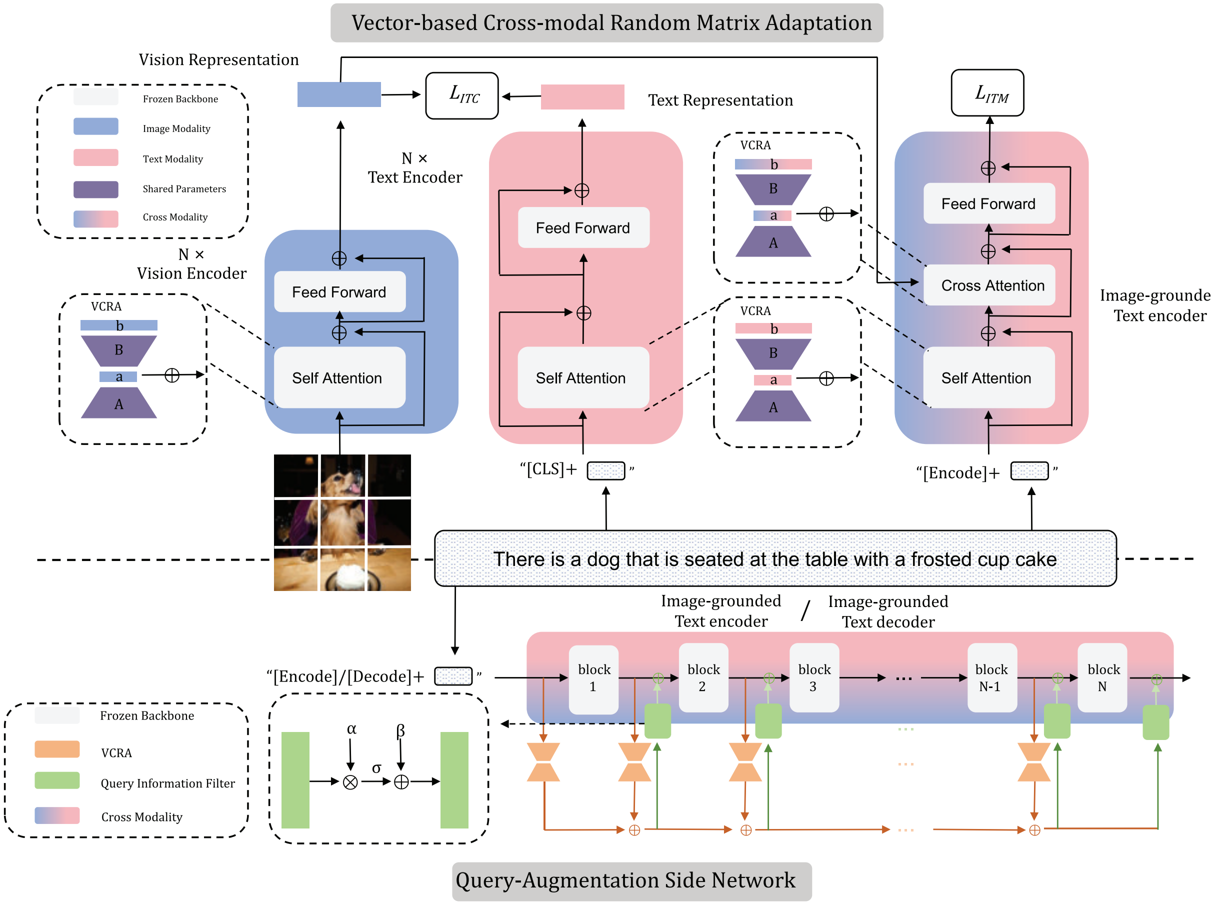

Figure 2: Illustration of Vector-Based Cross-Modal Random Matrix Adaptation (shown above) and Query-Augmentation Side Network (shown below). VCRA achieves modal interaction through sharing low rank matrices and introduces trainable scaling vectors to adapt to weight updates of various modalities and layers. QASN enhances the fusion between modalities by adaptively supplementing query information in the fusion network

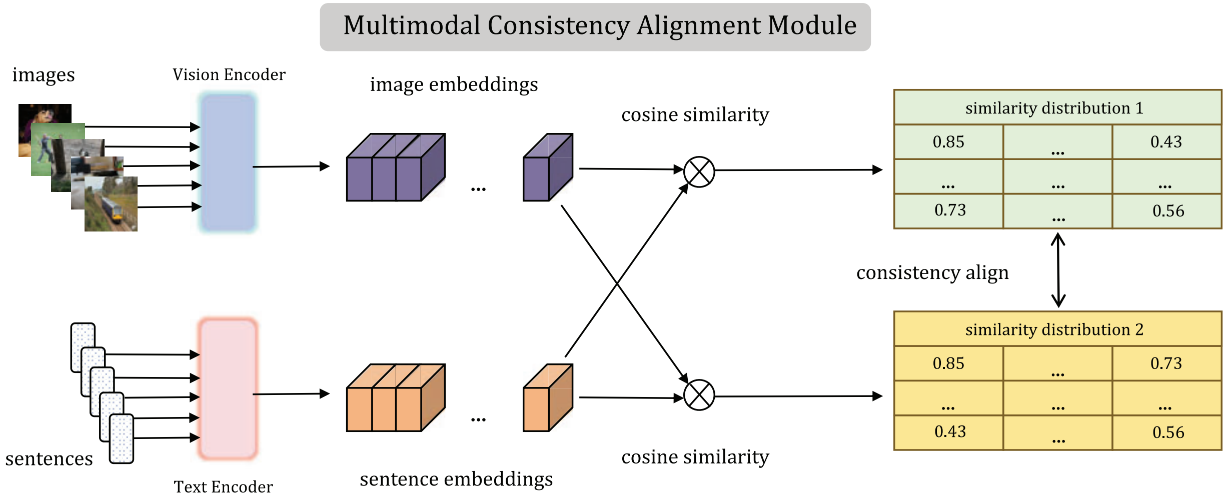

Figure 3: Illustration of Multimodal Consistency Alignment Module, it enhances fine-grained alignment between modalities by matching the similarity ranking consistency between image-text pairs in the contrastive learning process

Using contrastive learning loss to align image and text features lacks modeling the degree of repulsion for negative samples. For example, pushing an picture of a dog ’equally’ away from the descriptions ”cat”, ”tiger”, and ”house” can lead to the loss of fine-grained similarity information between the image and text and affecting feature alignment. We introduce the Multimodal Consistency Alignment Module to solve this problem by matching the consistency of similarity rankings between image-text pairs.

Formally, for a mini-batch of image-text pairs denoted as (x^V, x^T), where x^V represents the image and x^T represents the corresponding text, image and text are processed through the visual encoder $f(\cdot)$ and text encoder $g(\cdot)$, respectively, to obtain modality-specific representations. BLIP uses L_{ITC} as the contrastive loss:

$$\begin{aligned} L_{\mathrm{ITC}} = &-\sum_{i=1}^{N} \log \frac{e^{\phi(f(x_i^V),g(x_i^T))/\tau_1}}{\sum_{j=1}^{N} e^{\phi(f(x_i^V),g(x_j^T))/\tau_1}} \\ &-\sum_{i=1}^{N} \log \frac{e^{\phi(g(x_i^T),f(x_i^V))/\tau_1}}{\sum_{j=1}^{N} e^{\phi(g(x_i^T),f(x_j^V))/\tau_1}} \end{aligned} \tag{7}$$

$$\phi(f(x^V), g(x^T)) = \frac{f(x^V)\top g(x^T)}{\|f(x^V)\| \cdot \|g(x^T)\|} \tag{8}$$

where τ_1 is a temperature hyperparameter. While L_{ITC} is effective at distinguishing between positive and negative image-text pairs, it does not consider differences between highly relevant and moderately relevant pairs. Our proposed multimodal consistency alignment module introduces ranking information to capture fine-grained image-text relationships, enhancing modality representation and strengthening cross-modal alignment.

Specifically, for a given image-text pair within a mini-batch (x_i^V, x_i^T), we can obtain a list $S(x_i^V) = \{\phi(f(x_i^V), g(x_j^T))\}_{j=1}^{N}$ that represents the cosine similarity between the image x_i^V and all texts in the batch, as well as a list $S(x_i^T) = \{\phi(g(x_i^T), f(x_j^V))\}_{j=1}^{N}$ that represents the cosine similarity between the text x_i^T and all images in the batch. We aim for the corresponding elements in $S(x_i^V)$ and $S(x_i^T)$ to have the same ranking positions. This allows us to capture the fine-grained ranking information between the image and various negative text samples, as well as between the text and different negative image samples. We achieve ranking consistency by minimizing the the Jensen-Shannon (JS) divergence of the two top one probability distributions:

$$\begin{aligned} L_{\mathrm{ITR}} &= \sum_{i=1}^{N} \mathrm{JS}(\widetilde{\mathrm{S}}_{\tau_1}(x_i^V) \| \widetilde{\mathrm{S}}_{\tau_1}(x_i^T)) \\ &= \frac{1}{2} \sum_{i=1}^{N} (\widetilde{S}_{\tau_1}(x_i^V) \log \left(\frac{2\widetilde{S}_{\tau_1}(x_i^V)}{\widetilde{S}_{\tau_1}(x_i^V) + \widetilde{S}_{\tau_1}(x_i^T)} \right) \\ &\quad + \widetilde{S}_{\tau_1}(x_i^T) \log \left(\frac{2\widetilde{S}_{\tau_1}(x_i^T)}{\widetilde{S}_{\tau_1}(x_i^V) + \widetilde{S}_{\tau_1}(x_i^T)} \right)) \end{aligned} \tag{9}$$

where $\widetilde{S}_{\tau_1}(x_i^V) = softmax(S(x_i^V)/\tau)$ and $\widetilde{S}_{\tau_1}(x_i^T) = softmax(S(x_i^T)/\tau)$ represent the top-one probability distributions of $S(x_i^V)$ and $S(x_i^T)$, respectively, and τ is a hyperparameter. The final loss function is:

$$L_{\mathrm{all}} = L_{\mathrm{ITC}} + L_{\mathrm{ITR}} \tag{10}$$

4 Experiments

4.1 Implementation Details

We apply BLIP-base as our vision-language backbone for Image Caption, Image-Text Retrieval and VQA downstream tasks. During the fine-tuning process, the parameters of the backbone model are kept frozen. The experiments were implemented using PyTorch on 8 × NVIDIA 3090 GPU. As shown in Table 1, we present the training details for UniTrans. We applied VCRA to the projection of query, key and value in Attention layer, and the scaling vectors of key projection share parameters with value projection. At the start of fine-tuning, we initialize the random matrices A and B with random values drawn from a normal distribution, while vector Λ_a is initialized to 1 and vector Λ_b is initialized to 0. For the query information filters in QASN (components α and β), are initialized to 1 and 0, respectively. For the textual data in video datasets, we performed simple preprocessing steps, such as truncating words that exceed the maximum sentence length.

Table 1: Setting hyperparameters for fine-tuning training of multimodal downstream tasks

Config	Image captioning	Image-text retrieval			Visual question answering	
	COCO (caption)	MSCOCO	Flickr	DiDemo	VQAv2	MSRVTT-QA
learning rate	1e−5	1e−5	1e−4	1e−5	2e−5	2e−5
batch size	128	128	128	32	128	64
epochs	6	5	6	10	10	10
training input	384	384	384	8 × 224	384	8 × 224
inference input	384	384	384	16 × 224	384	16 × 224

4.2 Baselines & Datasets & Evaluation Metrics

We evaluate UniTrans across six benchmarks, covering three cross-modal tasks: Image Captioning, Image-Text Retrieval, and VQA. For the image captioning task, we use the COCO-Caption [41] dataset with the COCO Caption Karpathy split as the test set, employing BLEU@4 and CIDEr as evaluation metrics. BLEU measures the n-gram precision between the generated and reference captions, while CIDEr evaluates the similarity between candidate and reference captions by calculating the cosine similarity of their TF-IDF vectors. For the image-text retrieval task, we use the MSCOCO [41], Flickr30K [42], and Didemo datasets [43], with Recall at K (R@K) as evaluation metrics. R@K aims to calculate the ratio of queries that successfully retrieve the ground truth as one of the first K results. For the VQA task, we use the VQAv2 [44] and MSRVTT-QA [45] datasets, with the evaluation results obtained from the official validation platforms provided by the datasets.

4.3 Results

4.3.1 Performance Comparisons on Cross-modal Tasks

Tables 2 and 3 show the performances of UniTrans for image-text retrieval task on Flickr30K and MSCOCO. As shown, UniTrans achieves performance comparable to UniAdapter with only 0.2M parameters on both Flickr30K and MSCOCO, even outperforming it on certain metrics. Additionally, our method's performance is very close to that of fully fine-tuning the BLIP backbone while exceeding previous fully fine-tuned methods.

Table 2: Results on flickr

Method	# Tunable	Text retrieval			Image retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
Full fine-tuning							
UNITER [46]	330M	87.3	98.0	99.2	75.6	94.1	96.8
UNIMO [47]	330M	89.4	98.9	99.8	78.0	94.2	97.1
ALIGN [48]	820M	95.3	99.8	**100.0**	84.9	97.4	98.6
ALBEF [49]	210M	95.9	99.8	**100.0**	85.6	97.5	**98.9**
BLIP [4]	223M	**97.3**	99.9	**100.0**	**87.3**	**97.6**	**98.9**
Frozen backbone							
LoRA (r = 32)	10.6M	96.2	99.7	99.8	85.8	97.1	98.4
UniAdapter (r = 128)	4.8M	97.1	**100.0**	**100.0**	86.5	97.4	98.8
UniAdapter (r = 512)	19.0M	97.1	99.9	**100.0**	86.4	97.4	**98.9**
UniTrans (ours, r = 64)	**0.2M**	97.2	**100.0**	**100.0**	86.4	97.4	98.8

Note: Bold represents optimal performance.

Table 3: Results on MSCOCO

Method	# Tunable	Text retrieval			Image retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
Full fine-tuning							
Unicoder-VL [24]	–	62.3	87.1	92.8	46.7	76.0	85.3
OSCAR [50]	330M	70.0	91.1	95.5	54.0	80.8	88.5
ALIGN [48]	820M	77.0	93.5	96.9	59.9	83.3	89.8
ALBEF [49]	210M	77.6	94.3	97.2	60.7	84.3	90.5
BLIP [4]	223M	**81.9**	**95.4**	**97.8**	**64.3**	**85.7**	**91.5**
Frozen backbone							
LoRA (r = 32)	10.6M	80.0	94.1	97.2	62.1	84.4	90.6
UniAdapter (r = 128)	4.8M	79.8	94.2	97.5	62.3	84.5	90.8
UniAdapter (r = 512)	19.0M	80.1	94.6	97.4	62.6	84.6	90.9
UniTrans (ours, r = 64)	**0.2M**	79.7	94.2	97.4	62.2	84.4	90.5

Note: Bold represents optimal performance.

Unlike image-text retrieval, the VQA task requires multimodal generation capabilities. Therefore, in addition to the encoder, a decoder is necessary for text generation. Due to the structural differences between the encoder and decoder, we did not share the random matrices parameters, resulting in a slight increase in trainable parameters. However, the total parameter count remains significantly lower than the baseline. According to Table 4, our method outperforms all fine-tuning methods, demonstrating that our approach is well-suited for multimodal generation tasks.

Table 4: Results on VQAv2

Method	# Tunable	VQAv2	
		Test-dev	Test-std
Full fine-tuning			
VL-T5/BART [51]	165M	–	71.30
SOHO [52]	155M	73.25	73.47
OSCAR [50]	330M	73.61	73.82
UNITER [46]	330M	73.82	74.03
ALBEF [49]	266M	75.84	76.04
BLIP [4]	337M	77.44	77.48
Frozen backbone			
LoRA (r = 32)	–	–	–
UniAdapter (r = 128)	4.8M	73.72	73.71
UniAdapter (r = 512)	19.0M	75.44	75.56
UniTrans (ours, r = 64)	**0.3M**	**77.68**	**77.77**

Note: Bold represents optimal performance.

Table 5 shows the performance of our method on the Image Captioning task using the COCO caption dataset. The results indicate that UniTrans outperforms the baseline and is only slightly below the fully fine-tuned BLIP method. As shown in Table 6, our UniTrans also outperforms the baseline on video datasets.

Table 5: Results on the Image Captioning dataset COCO Caption, where B@4 represents BLEU@4 and C denotes CIDEr

Method	# Pre-train	COCO caption karpathy test	
		B@4	C
Full fine-tuning			
Enc-Dec [53]	15M	–	110.9
VinVL [54]	5.7M	38.2	129.3
LEMON [55]	200M	**40.3**	**133.3**
BLIP [4]	14M	38.6	129.7
BLIP [4]	129M	39.7	**133.3**
Frozen backbone			
LoRA (r = 64)	129M	38.88	131.5
UniAdapter (r = 128)	129M	39.0	132.1
UnTrans (ours, r = 64)	129M	39.2	132.3

Note: Bold represents optimal performance.

Table 6: Results on the video-text retrieval dataset DiDemo and the video visual question answering dataset MSRVTT-QA

Method	# Tunable	DiDemo			Method	# Tunable	MSRVTT-QA
		R@1	R@5	R@10			Test acc
Full fine-tuning							
CLIPBERT [56]	135M	20.4	48.0	60.8	CLIPBERT [56]	135M	37.4
Frozen in Time [57]	180M	34.6	65.0	74.7	CoMVT [58]	–	39.5
ALPRO [59]	245M	35.9	67.5	78.8	ALPRO [59]	245M	42.1
VIOLET [60]	306M	32.6	62.8	74.7	Just-Ask [61]	200M	41.5
All-in-one [62]	110M	32.7	61.4	73.5	VIOLET [60]	306M	43.9
CLIP4Clip [63]	124M	42.8	68.5	79.2	MERLOT [64]	233M	43.1
Frozen backbone							
LoRA (r = 32)	10.6M	50.9	75.3	82.4	–	–	–
UniAdapter (r = 128)	4.8M	49.0	75.5	83.3	UniAdapter (r = 128)	4.8M	44.2
UniAdapter (r = 512)	19.0M	52.1	**77.3**	85.2	UniAdapter (r = 512)	19.0M	44.7
UniTrans (ours, r = 64)	**0.2M**	**52.4**	**77.3**	**85.3**	UniTrans (ours, r = 64)	**0.3M**	**44.8**

Note: Bold represents optimal performance.

From the comparative experimental results on multimodal benchmarks, it can be observed that Unitrans achieves competitive performance while reducing the trainable parameters of LoRA to just 5%. This is attributable to two main factors. First, we leverage VeRA to share a low-rank random matrix across modalities and employ scaling vectors to adapt weight updates. This shared mechanism reduces the number of large random matrices involved in training, effectively decreasing parameter size while enhancing inter-modal information exchange. Second, QASN adaptively supplements query information within the fusion network, and MCAM regularizes fine-grained image-text alignment during fine-tuning through contrastive learning loss, strengthening multimodal alignment. In comparison, UniAdapter enables modality interaction during fine-tuning by employing a shared MLP in the adapter. However, it lacks information sharing across layers, and the MLP itself involves a considerable number of trainable parameters, particularly as the rank r increases. Based on this, we extend the concept of UniAdapter to Low-Rank Adaptation and further optimize it by introducing QASN and MCAM to facilitate multimodal alignment during fine-tuning. As a result, we achieve performance comparable to UniAdapter while significantly reducing the number of trainable parameters.

4.3.2 Training Efficiency and Storage Cost

As shown in Table 7, we present the training time and GPU memory costs for the three tasks: Image Captioning, Image-Text Retrieval, and Visual Question Answering. We consider the training time and storage cost of fully fine-tuning BLIP as one unit. From the table, it can be observed that our UniTrans outperforms both fully fine-tuned BLIP and UniAdapter in terms of training time and GPU memory cost.

Table 7: Comparison of training time and GPU memory usage

Method	#Tunable	Image captioning		#Tunable	Image-text retrieval		#Tunable	VQA	
		Time	Memory		Time	Memory		Time	Memory
Full fine-tuning	213M	1.00	1.00	223M	1.00	1.00	337M	1.00	1.00
UniAdapter (r = 512)	19.0M	0.83	0.81	19.0M	0.88	0.86	19.0M	0.93	0.80

(Continued)

Table 7 (continued)

Method	#Tunable	Image captioning		#Tunable	Image-text retrieval		#Tunable	VQA	
		Time	Memory		Time	Memory		Time	Memory
UniAdapter (r = 128)	4.8M	0.81	0.77	4.8M	0.86	0.82	4.8M	0.92	**0.72**
UniTrans (r = 128)	0.36M	0.81	0.77	0.36M	0.86	0.81	0.46M	0.89	**0.71**
UniTrans (r = 64)	**0.26M**	**0.80**	**0.75**	**0.26M**	**0.83**	**0.79**	**0.35M**	**0.89**	**0.68**

Note: Bold represents optimal performance.

4.3.3 The Impact of Rank on UniTrans

We also explored the impact of rank size on UniTrans. As shown in Fig. 4a,b, experiments conducted on Flickr30K indicate that as the rank increases, performance of cross-modal downstream fine-tuning improves. However, when the rank reaches a certain threshold, the performance gains slow down or even slightly decrease. This may be due to the rank increase making the random matrix too large, introducing redundant parameters and leading to overfitting. As shown in Fig. 4c, as the rank R increases, the growth in parameters for our UniTrans is significantly slower compared to LoRA and UniAdapter, demonstrating the superiority of our method.

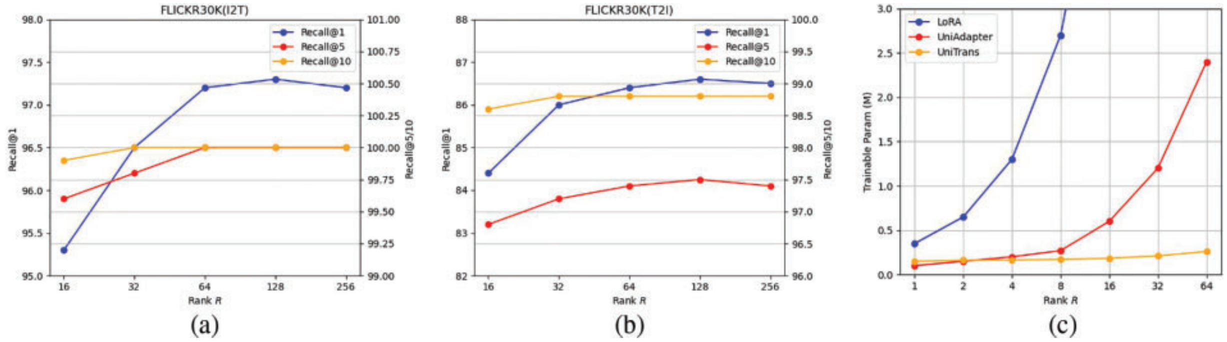

Figure 4: how rank affects UniTrans. (a) and (b) show the impact of different ranks on UniTrans performance on the Flickr30K dataset. (c) demonstrates the scalability of parameters compared to other PETL methods

4.3.4 Visualization Results

To intuitively understand the performance of UniTrans in multimodal downstream tasks, we present detailed visualization results in Figs. 5–7. Fig. 5 shows a comparison between the captions generated by UniTrans on the COCO Caption dataset and the ground-truth. It can be seen that UniTrans is capable of accurately generating image captions, with key words closely matching the ground-truth. Fig. 6 demonstrates the results of the Image-Text Retrieval task on the Flickr30K dataset, where UniTrans accurately retrieves the corresponding image given a textual caption. We also present the results of the VQA task on the VQAv2 dataset, where UniTrans provides accurate answers. However, the responses are somewhat brief, which may be attributed to the relatively smaller parameter size of the multimodal base model. In the future, we plan to conduct experiments with larger-scale models.

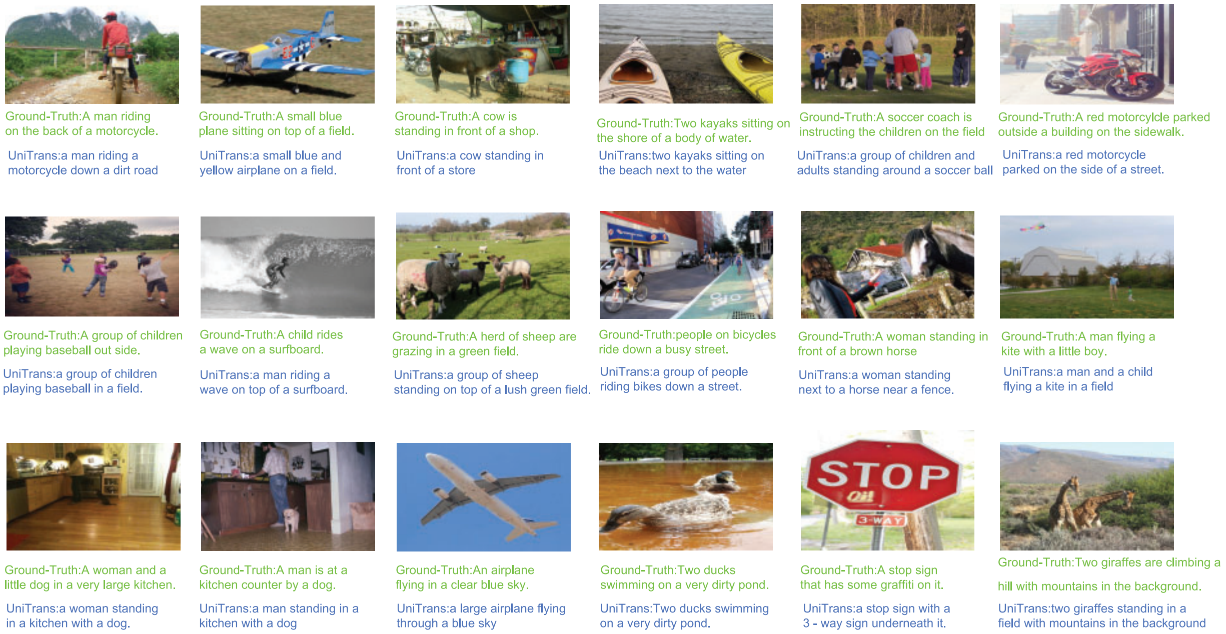
Ground-Truth:A man riding on the back of a motorcycle.
UniTrans:a man riding a motorcycle down a dirt road
Ground-Truth:A small blue plane sitting on top of a field.
UniTrans:a small blue and yellow airplane on a field.
Ground-Truth:A cow is standing in front of a shop.
UniTrans:a cow standing in front of a store
Ground-Truth:Two kayaks sitting on the shore of a body of water.
UniTrans:two kayaks sitting on the beach next to the water
Ground-Truth:A soccer coach is instructing the children on the field
UniTrans:a group of children and adults standing around a soccer ball
Ground-Truth:A red motorcylcle parked outside a building on the sidewalk.
UniTrans:a red motorcycle parked on the side of a street.
Ground-Truth:A group of children playing baseball out side.
UniTrans:a group of children playing baseball in a field.
Ground-Truth:A child rides a wave on a surfboard.
UniTrans:a man riding a wave on top of a surfboard.
Ground-Truth:A herd of sheep are grazing in a green field.
UniTrans:a group of sheep standing on top of a lush green field.
Ground-Truth:people on bicycles ride down a busy street.
UniTrans:a group of people riding bikes down a street.
Ground-Truth:A woman standing in front of a brown horse
UniTrans:a woman standing next to a horse near a fence.
Ground-Truth:A man flying a kite with a little boy.
UniTrans:a man and a child flying a kite in a field
Ground-Truth:A woman and a little dog in a very large kitchen.
UniTrans:a woman standing in a kitchen with a dog.
Ground-Truth:A man is at a kitchen counter by a dog.
UniTrans:a man standing in a kitchen with a dog
Ground-Truth:An airplane flying in a clear blue sky.
UniTrans:a large airplane flying through a blue sky
Ground-Truth:Two ducks swimming on a very dirty pond.
UniTrans:Two ducks swimming on a very dirty pond.
Ground-Truth:A stop sign that has some graffiti on it.
UniTrans:a stop sign with a 3 - way sign underneath it.
Ground-Truth:Two giraffes are climbing a hill with mountains in the background.
UniTrans:two giraffes standing in a field with mountains in the background

Figure 5: Visualization results on COCO Caption

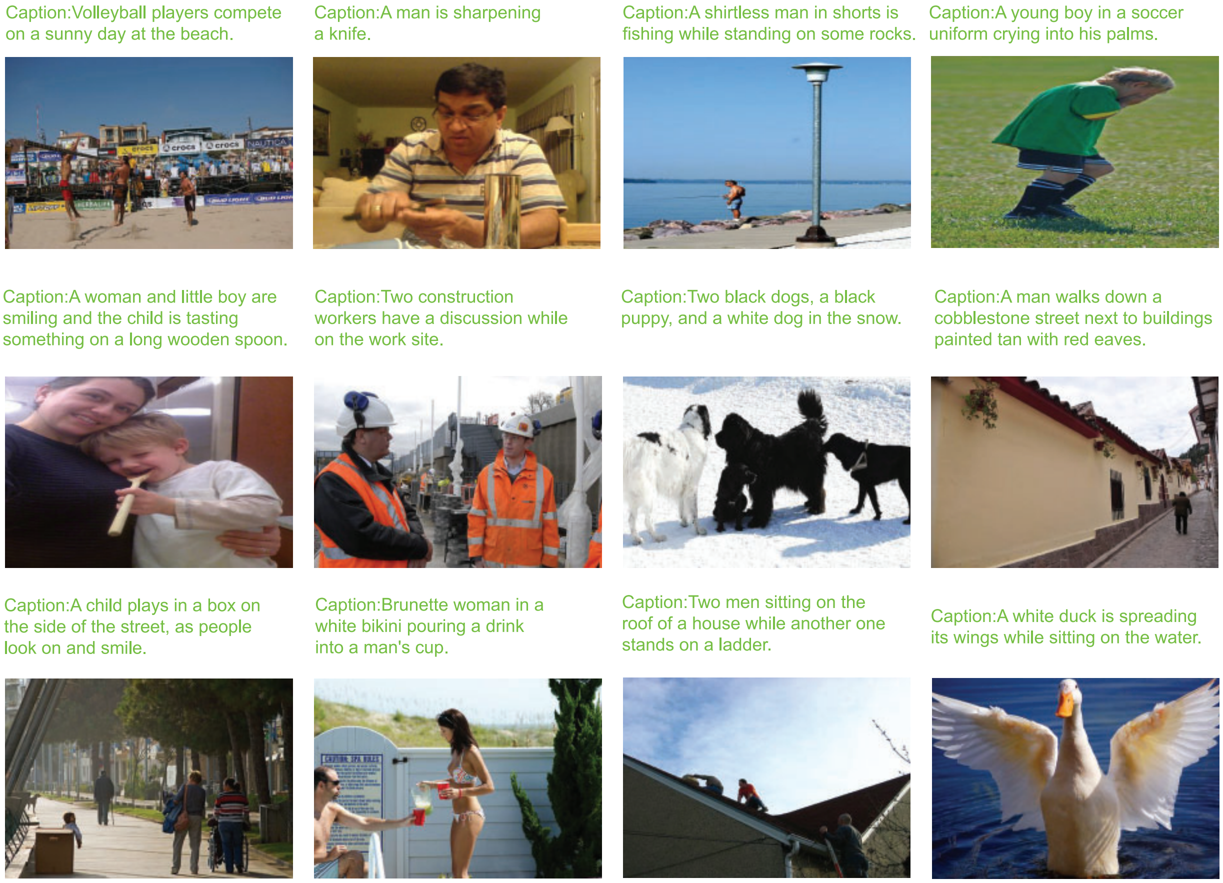
Caption:Volleyball players compete on a sunny day at the beach.
Caption:A man is sharpening a knife.
Caption:A shirtless man in shorts is fishing while standing on some rocks.
Caption:A young boy in a soccer uniform crying into his palms.
Caption:A woman and little boy are smiling and the child is tasting something on a long wooden spoon.
Caption:Two construction workers have a discussion while on the work site.
Caption:Two black dogs, a black puppy, and a white dog in the snow.
Caption:A man walks down a cobblestone street next to buildings painted tan with red eaves.
Caption:A child plays in a box on the side of the street, as people look on and smile.
Caption:Brunette woman in a white bikini pouring a drink into a man's cup.
Caption:Two men sitting on the roof of a house while another one stands on a ladder.
Caption:A white duck is spreading its wings while sitting on the water.

Figure 6: Visualization results on Flickr30K

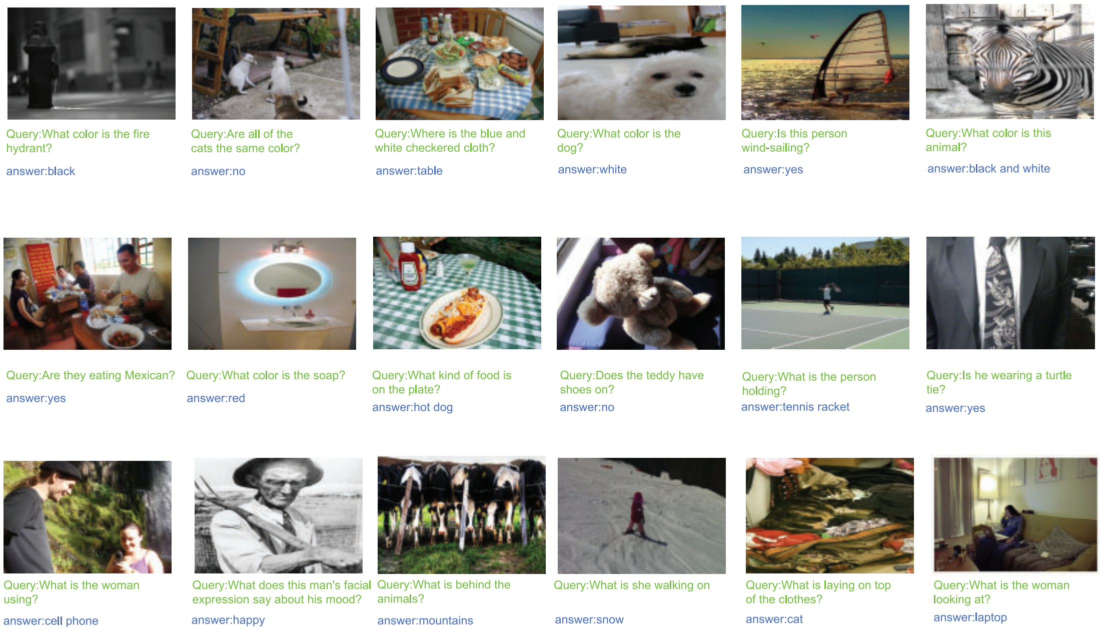
Query:What color is the fire hydrant?
answer:black
Query:Are all of the cats the same color?
answer:no
Query:Where is the blue and white checkered cloth?
answer:table
Query:What color is the dog?
answer:white
Query:Is this person wind-sailing?
answer:yes
Query:What color is this animal?
answer:black and white
Query:Are they eating Mexican?
answer:yes
Query:What color is the soap?
answer:red
Query:What kind of food is on the plate?
answer:hot dog
Query:Does the teddy have shoes on?
answer:no
Query:What is the person holding?
answer:tennis racket
Query:Is he wearing a turtle tie?
answer:yes
Query:What is the woman using?
answer:cell phone
Query:What does this man's facial expression say about his mood?
answer:happy
Query:What is behind the animals?
answer:mountains
Query:What is she walking on
answer:snow
Query:What is laying on top of the clothes?
answer:cat
Query:What is the woman looking at?
answer:laptop

Figure 7: Visualization results on VQAv2

4.3.5 Ablation Study

To assess the effectiveness of each module in UniTrans, we conducted ablation experiments on image-text retrieval using the Flickr30K benchmark and the DiDemo benchmark. We compared UniTrans with three baselines: a frozen BLIP, fully fine-tuned BLIP, and LoRA. As shown in Tables 8 and 9, when the random matrices are not shared ('+scaling vectors' means only scaling vector is used, and random matrices are shared in each transformer block of single modality), the number of trainable parameters is significantly reduced, but the performance deteriorates. However, our proposed VCRA demonstrates clear improvements, highlighting its contribution to modality alignment. The results also indicate that the QASN and the MCAM, both designed to enhance modality alignment, further boost performance. Notably, the latter achieves performance gains without introducing additional parameters.

Table 8: Results of ablation on the Flickr

Method	# Tunable	Text retrieval			Image retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
Frozen	0M	86.9	98.0	99.1	78.1	94.0	97.7
Full fine-tuning	223M	97.3	99.9	100.0	87.3	97.6	98.9
LoRA (r = 32)	10.6M	96.2	99.7	99.8	85.8	97.1	98.4
+scaling vectors	0.28M	92.0	98.6	99.5	82.1	95.9	98.0
+shared random matrices	0.18M	95.3	99.6	99.9	84.4	96.7	98.4
+QASN	0.21M	96.1	99.7	100.0	85.2	96.8	98.6
+MCAM (UniTrans)	0.21M	96.5	99.8	100.0	86.0	97.2	98.8

Table 9: Results of ablation on the DiDemo

Method	# Tunable	DiDemo		
		R@1	R@5	R@10
Linear probe	0.4M	39.7	64.6	74.9
Full fine-tuning	223M	51.3	79.1	85.7
LoRA (r = 32)	10.6M	50.9	75.3	82.4
+scaling vectors	0.28M	48.4	74.9	81.6
+shared random matrices	0.18M	49.1	76.2	83.3
+QASN	0.21M	50.6	76.5	84.0
+MCAM (UniTrans)	0.21M	51.8	77.1	85.2

5 Conclusion & Future Work

In this paper, we propose a novel paradigm for efficient cross-modal knowledge transfer, UniTrans, to enhance modality interaction and alignment during the fine-tuning process of multimodal models. The concepts of VCRA, along with the modality alignment modules QASN and MCAM, are simple and lightweight, allowing them to be extended to different multimodal base models without altering their inherent structure, thereby effectively adapting to various fine-grained visual-language tasks. Extensive evaluations on multiple downstream benchmarks demonstrate that our method achieves superior performance with fewer than 1M parameters.

Moreover, UniTrans has its limitations, which provide directions for our future work. (1) UniTrans has only been validated on the multimodal model architecture with fusion networks represented by BLIP, and has not been tested on architectures such as the Q-Former (represented by BLIP2) or the MLP-based architecture (represented by LLava). In the future, we will explore a wider range of multimodal model architectures. (2) Our method has been compared to classic cross-modal tasks such as Image Captioning, Image-Text Retrieval, and VQA, but has not yet been explored for other cross-modal tasks, such as text-to-image generation. Moving forward, we will extend our approach to a broader set of tasks. (3) The fine-tuning datasets we used are relatively small in scale. In the future, we plan to conduct experiments on larger-scale datasets.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Jiakang Sun, Ke Chen; data collection: Xinyang He; analysis and interpretation of results: Jiakang Sun, Xu Liu, Ke Li; draft manuscript preparation: Jiakang Sun, Cheng Peng. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in the The website link below: MSCOCO: https://cocodataset.org/#download, accessed on 10 July 2024. Flickr30K: https://shannon.cs.illinois.edu/DenotationGraph/data/index.html, accessed on 10 July 2024. MSRVTT: https://www.mediafire.com/folder/h14iarbs62e7p/shared, accessed on 10 July 2024. DiDemo: https://github.com/jpthu17/EMCL, accessed on 10 July 2024. VQAv2: https://visualqa.org/download.html, accessed on 10 July 2024. COCO Caption: https://cocodataset.org/#download, accessed on 10 July 2024.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Rajič F, Ke L, Tai YW, Tang CK, Danelljan M, Yu F. Segment anything meets point tracking. arXiv:230701197. 2023. doi:10.48550/arXiv.2307.01197.
2. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv:230213971. 2023. doi:10.48550/arXiv.2302.13971.
3. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023; Paris, France. p. 4015–26. doi:10.1109/ICCV51070.2023.00371.
4. Li J, Li D, Xiong C, Hoi S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning; 2022; Baltimore, MD, USA: PMLR. p. 12888–900.
5. Brown TB. Language models are few-shot learners. arXiv:2005.14165. 2020. doi:10.48550/arXiv.2005.14165.
6. Kheddar H, Himeur Y, Al-Maadeed S, Amira A, Bensaali F. Deep transfer learning for automatic speech recognition: towards better generalization. Knowl Based Syst. 2023;277:110851. doi:10.1016/j.knosys.2023.110851.
7. Iman M, Arabnia HR, Rasheed K. A review of deep transfer learning and recent advancements. Technologies. 2023;11(2):40. doi:10.3390/technologies11020040.
8. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv:2106.09685. 2021. doi:10.48550/arXiv.2106.09685.
9. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning; 2019; Long Beach, CA, USA: PMLR. p. 2790–9.
10. Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. arXiv:2101.00190. doi:10.48550/arXiv.2101.00190.
11. Cho S, Shin H, Hong S, Arnab A, Seo PH, Kim S. Cost aggregation for open-vocabulary semantic segmentation. In: CAT-seg: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024; Seattle WA, USA. p. 4113–23. doi:10.1109/CVPR52733.2024.00394.
12. Rao Y, Zhao W, Chen G, Tang Y, Zhu Z, Huang G, et al. DenseCLIP: language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022; New Orleans, LA, USA. p. 18082–91. doi:10.1109/CVPR52688.2022.01755.
13. Sung YL, Cho J, Bansal M. Vl-adapter: parameter-efficient transfer learning for vision-and-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022; New Orleans, LA, USA. p. 5227–37.
14. Lu H, Huo Y, Yang G, Lu Z, Zhan W, Tomizuka M, et al. UniAdapter: unified parameter-efficient transfer learning for cross-modal modeling. arXiv:2302.06605. 2023. doi:10.48550/arXiv.2302.06605.
15. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning; 2021; PMLR. p. 8748–63.
16. Cen J, Zhou Z, Fang J, Shen W, Xie L, Jiang D, et al. Segment anything in 3D with NeRFs. Adv Neural Inf Process Syst. 2023;36:25971–90.
17. Kheddar H. Transformers and large language models for efficient intrusion detection systems: a comprehensive survey. arXiv:2408.07583. 2024. doi:10.48550/arXiv.2408.07583.
18. Devlin J. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018. doi:10.48550/arXiv.1810.04805.
19. Li J, Li D, Savarese S, Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning; 2023; Honolulu, HI, USA: PMLR. p. 19730–42.
20. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Adv Neural Inf Process Syst. 2024;36:34892–916.

21. Chen H, Ding G, Liu X, Lin Z, Liu J, Han J. IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; Seattle, WA, USA. p. 12655–63. doi:10.1109/CVPR42600.2020.01267.
22. Qu L, Liu M, Wu J, Gao Z, Nie L. Dynamic modality interaction modeling for image-text retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021; Canada. p. 1104–13. doi:10.1145/3404835.3462829.
23. Yang Z, He X, Gao J, Deng L, Smola A. Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, NV, USA. p. 21–9. doi:10.1109/CVPR.2016.10.
24. Li G, Duan N, Fang Y, Gong M, Jiang D. Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training. Proc AAAI Conf Artif Intell. 2020;34:11336–44. doi:10.1609/aaai.v34i07.6795.
25. Tang Y, Yu J, Gai K, Zhuang J, Xiong G, Hu Y, et al. Context-I2W: mapping images to context-dependent words for accurate zero-shot composed image retrieval. Proc AAAI Conf Artif Intell. 2024;38:5180–8. doi:10.1609/aaai.v38i6.28324.
26. Cao M, Yang T, Weng J, Zhang C, Wang J, Zou Y. LocVTP: video-text pre-training for temporal localization. In: European Conference on Computer Vision; 2022; Tel Aviv, Israel: Springer. p. 38–56. doi:10.1007/978-3-031-19809-0_3.
27. Zhuang J, Yu J, Ding Y, Qu X, Hu Y. Towards fast and accurate image-text retrieval with self-supervised fine-grained alignment. IEEE Trans Multimed. 2023;26:1361–72. doi:10.1109/TMM.2023.3280734.
28. Wang X, Zhang R, Shen C, Kong T, Li L. Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 3024–33. doi:10.1109/CVPR46437.2021.00304.
29. Tang Y, Yu J, Gai K, Wang Y, Hu Y, Xiong G, et al. Align before Search: aligning ads image to text for accurate cross-modal sponsored search. arXiv:2309.16141. 2023. doi:10.48550/arXiv.2309.16141.
30. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. arXiv:2104.08691. 2021. doi:10.48550/arXiv.2104.08691.
31. Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT understands, too. AI Open. 2024;5:208–215. doi:10.1016/j.aiopen.2023.08.012.
32. Pfeiffer J, Kamath A, Rücklé A, Cho K, Gurevych I. Adapterfusion: non-destructive task composition for transfer learning. arXiv:2005.00247. 2020. doi:10.48550/arXiv.2005.00247.
33. Rücklé A, Geigle G, Glockner M, Beck T, Pfeiffer J, Reimers N, et al. AdapterDrop: on the efficiency of adapters in transformers. arXiv:2010.11918. 2020. doi:10.48550/arXiv.2010.11918.
34. Zaken EB, Ravfogel S, Goldberg Y. BitFit: simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv:2106.10199. 2021. doi:10.48550/arXiv.2106.10199.
35. Guo D, Rush AM, Kim Y. Parameter-efficient transfer learning with diff pruning. arXiv:2012.07463. 2020. doi:10.48550/arXiv.2012.07463.
36. Zhang Q, Chen M, Bukharin A, Karampatziakis N, He P, Cheng Y, et al. AdaLoRA: adaptive budget allocation for parameter-efficient fine-tuning. arXiv:2303.10512. 2023. doi:10.48550/arXiv.2303.10512.
37. Renduchintala A, Konuk T, Kuchaiev O. Tied-LoRA: enhacing parameter efficiency of lora with weight tying. arXiv:2311.09578. 2023. doi:10.48550/arXiv.2311.09578.
38. Liu SY, Wang CY, Yin H, Molchanov P, Wang YCF, Cheng KT, et al. DoRA: weight-decomposed low-rank adaptation. arXiv:2402.09353. 2024. doi:10.48550/arXiv.2402.09353.
39. Hyeon-Woo N, Ye-Bin M, Oh TH. Fedpara: low-rank hadamard product for communication-efficient federated learning. arXiv:2108.06098. 2021. doi:10.48550/arXiv.2108.06098.
40. Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020. doi:10.48550/arXiv.2010.11929.
41. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D et al. Microsoft COCO: common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference; 2014 Sep 6–12; Zurich, Switzerland: Springer. p. 13–55. doi:10.1007/978-3-319-10602-1_48.

42. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision; 2015; Santiago, Chile. p. 2641–9. doi:10.1109/ICCV.2015.303.
43. Anne Hendricks L, Wang O, Shechtman E, Sivic J, Darrell T, Russell B. Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision; 2017; Venice, Italy. p. 5803–12. doi:10.1109/ICCV.2017.618.
44. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. 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; 2017; Honolulu, HI, USA. p. 6904–13. doi:10.1007/s11263-018-1116-0.
45. Xu D, Zhao Z, Xiao J, Wu F, Zhang H, He X, et al. Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM International Conference on Multimedia; 2017; Mountain View, CA, USA. p. 1645–53. doi:10.1145/3123266.3123427.
46. Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, et al. Uniter: universal image-text representation learning. In: European Conference on Computer Vision; 2020; Glasgow, UK: Springer. p. 104–20. doi:10.1007/978-3-030-58577-8_7.
47. Li W, Gao C, Niu G, Xiao X, Liu H, Liu J, et al. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. arXiv:2012.15409. 2020. doi:10.18653/v1/2021.acl-long.202.
48. Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning; 2021; PMLR. p. 4904–16.
49. Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi SCH. Align before fuse: vision and language representation learning with momentum distillation. Adv Neural Inf Process Syst. 2021;34:9694–705.
50. Li X, Yin X, Li C, Zhang P, Hu X, Zhang L et al. Oscar: object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference; 2020 Aug 23–28; Zurich, Switzerland: Springer. p. 121–37. doi:10.1007/978-3-030-58577-8_8.
51. Cho J, Lei J, Tan H, Bansal M. Unifying vision-and-language tasks via text generation. In: International Conference on Machine Learning; 2021; PMLR. p. 1931–42.
52. Huang Z, Zeng Z, Huang Y, Liu B, Fu D, Fu J. 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; 2021. p. 12976–85. doi:10.1109/CVPR46437.2021.01278.
53. Changpinyo S, Sharma P, Ding N, Soricut R. Conceptual 12m: pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 3558–68. doi:10.1109/CVPR46437.2021.00356.
54. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, et al. VinVL: making visual representations matter in vision-language models. arXiv:2101.00529. 2021. doi:10.48550/arXiv.2101.00529.
55. Hu X, Gan Z, Wang J, Yang Z, Liu Z, Lu Y, et al. Scaling up vision-language pre-training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022; New Orleans, LA, USA. p. 17980–9. doi:10.1109/CVPR52688.2022.01745.
56. Lei J, Li L, Zhou L, Gan Z, Berg TL, Bansal M, et al. 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; 2021. p. 7331–41. doi:10.1109/CVPR46437.2021.00725.
57. Bain M, Nagrani A, Varol G, Zisserman A. 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; 2021; Montreal, QC, Canada. p. 1728–38. doi:10.1109/ICCV48922.2021.00175.
58. Seo PH, Nagrani A, Schmid C. Look before you speak: visually contextualized utterances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 16877–87. doi:10.1109/CVPR46437.2021.01660.
59. Li D, Li J, Li H, Niebles JC, Hoi SC. Align and prompt: video-and-language pre-training with entity prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; New Orleans, LA, USA; 2022. p. 4953–63. doi:10.1109/CVPR52688.2022.00490.

60. Fu TJ, Li L, Gan Z, Lin K, Wang WY, Wang L, et al. VIOLET: end-to-end video-language transformers with masked visual-token modeling. arXiv:2111.12681. 2021. doi:10.48550/arXiv.2111.12681.
61. Yang A, Miech A, Sivic J, Laptev I, Schmid C. Just ask: learning to answer questions from millions of narrated videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021; Montreal, QC, Canada. p. 1686–97. doi:10.1109/ICCV48922.2021.00171.
62. Wang J, Ge Y, Yan R, Ge Y, Lin KQ, Tsutsui S, et al. All in one: exploring unified video-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; Vancouver, BC, Canada. p. 6598–608. doi:10.1109/CVPR52729.2023.00638.
63. Luo H, Ji L, Zhong M, Chen Y, Lei W, Duan N, et al. Clip4Clip: an empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing. 2022;508:293–304. doi:10.1016/j.neucom.2022.07.028.
64. Zellers R, Lu X, Hessel J, Yu Y, Park JS, Cao J, et al. MERLOT: multimodal neural script knowledge models. Adv Neural Inf Process Syst. 2021;34:23634–51.