Dynamic Adapter with Semantics Disentangling for Cross-lingual Cross-modal Retrieval

Rui Cai[1,2], **Zhiyu Dong**[1,2], **Jianfeng Dong**[1,2*], **Xun Wang**[1,2]

[1] the College of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou, China
[2] Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology, Hangzhou, China
cairuics@gmail.com, dongzhiyuuu@163.com, {djf, wx}@zjgsu.edu.cn

Abstract

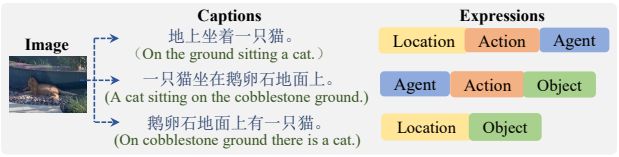
Existing cross-modal retrieval methods typically rely on large-scale vision-language pair data. This makes it challenging to efficiently develop a cross-modal retrieval model for under-resourced languages of interest. Therefore, Cross-lingual Cross-modal Retrieval (CCR), which aims to align vision and the low-resource language (the *target language*) without using any human-labeled target-language data, has gained increasing attention. As a general parameter-efficient way, a common solution is to utilize adapter modules to transfer the vision-language alignment ability of Vision-Language Pretraining (VLP) models from a source language to a target language. However, these adapters are usually *static* once learned, making it difficult to adapt to target-language captions with *varied* expressions. To alleviate it, we propose ***D**ynamic **A**dapter with **S**emantics **D**isentangling* (DASD), whose parameters are dynamically generated conditioned on the characteristics of the input captions. Considering that the semantics and expression styles of the input caption largely influence how to encode it, we propose a semantic disentangling module to extract the semantic-related and semantic-agnostic features from the input, ensuring that generated adapters are well-suited to the characteristics of input caption. Extensive experiments on two image-text datasets and one video-text dataset demonstrate the effectiveness of our model for cross-lingual cross-modal retrieval, as well as its good compatibility with various VLP models.

Code — https://github.com/HuiGuanLab/DASD
Extended version — https://arxiv.org/abs/2412.13510

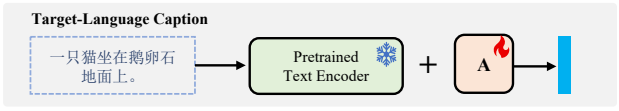
Introduction

With the rapid emergence of images and videos on the Internet, there is a huge demand from users around the world for retrieving visual content of interest by natural language queries (*a.k.a.* cross-modal retrieval) (Li et al. 2021; Zhang et al. 2023; Chang et al. 2023). Recent neural-based cross-modal retrieval models (Bogolin et al. 2022; Lu et al. 2022; Sun et al. 2023) tend to require a large amount of human-labeled text-image/video pair data for training which are available for only a handful of the world's languages. As a result, building a cross-modal retrieval system for users with different language backgrounds is extremely challenging, especially for low-resource languages (*e.g.,* Czech). With this regard, *cross-lingual* cross-modal retrieval (CCR) leverages visual-text pair data in the rich-resource language (the *source language*) to construct a retrieval model for a new language of interest (the *target language*), avoiding substantial manual annotations costs on the target language.

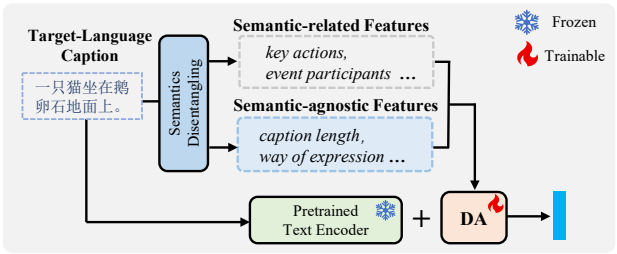
The perennial problem with building target-language retrieval models lies in the paucity of training data, since existing human-labeled resources for low-resource languages are rather limited, and it is extremely expensive and time-consuming to manually annotate images/videos with de-



(a) Captions with the same semantics but different expressions.



(b) A typical adapter for target-language encoding.



(c) Our proposed DASD framework for CCR.

Figure 1: Illustration of the variety of textual expressions and the difference between the traditional adapter and our DASD: (a) Captions of the same image are differently expressed in Chinese-specific ways. (b) Traditional adapters whose parameters are fixed once learned. (c) Our method extracts semantic-related and semantic-agnostic features from captions and thereby produces dynamic adapters (DA).

*The corresponding author.

scriptions in multiple languages. Due to limitations in data and computing resources, existing Vision-Language Pretraining (VLP) models, such as CLIP (Radford et al. 2021) and CCLM (Zeng et al. 2023), could support only one or a few languages, yet there are more than 6,900 languages worldwide (Zhou et al. 2021). Moreover, these VLP models cannot be flexibly extended to new languages, since additional training on target languages will cause performance degeneration of VLP models on the original languages due to the limited model capacity.

A straightforward and cheap solution is converting source-language labeled data into the target language utilizing Machine Translation (MT) tools (*e.g.,* Google Translate[1]). With access to these MT-generated resources, existing works tend to transfer the vision-language alignment ability of VLP models to target languages through cross-lingual alignment (Wang et al. 2024a,b; Pfeiffer et al. 2020; Zhang, Hu, and Jin 2022). Among them, a prior work (Wang et al. 2022) tries to finetune the pre-trained layers in VLP models with cross-lingual alignment objectives, inevitably leading to a certain degree of knowledge forgetting. To alleviate this problem, some adapter-based methods (Pfeiffer et al. 2020; Zhang, Hu, and Jin 2022) have recently been proposed to perform cross-lingual transfer in a parameter-efficient way. These methods freeze VLP models and store cross-lingual knowledge in the light-weight adapters, whose parameters keep static for different inputs. However, during the cross-lingual transfer, the language gaps (Ahmad et al. 2019), such as unique expressions in target languages, could bring complexity and increase the difficulty of extracting the accurate semantics of captions. As illustrated in Figure 1(a), target-language (Chinese) captions describing the same event are expressed in quite different ways. As a result, existing *static* adapters, shown in Figure 1(b), struggle to adapt to target-language captions with *varied* expressions.

To tackle the aforementioned challenge, we propose *Dynamic Adapter with Semantics Disentangling* (DASD), a novel paradigm that adaptively encodes target-language captions by making language adapters conditioned on each input caption rather than keeping them fixed after once learning. In order to obtain adapters that exactly match the input caption, we perform semantics disentangling to capture its distinct and complementary aspects. To be specific, we assume that each caption is entangled by two independent characteristics: semantic-related and semantic-agnostic features. Particularly, the former presents consistent semantic features shared by different modalities, while the latter reflects the characteristics with respect to the mean of expression yet unrelated to semantics, such as word order, sentence length, and other low-level information (shown in Figure 1(c)). Both semantic-related and semantic-agnostic features of input captions are learned by semantics disentangling, which explicitly decouples the two different features through semantic consistency learning and adversarial training. In this way, the disentangled features capture sufficient information needed for characterizing the input caption, which are then fed to the dynamic parameter generation module. Cross-lingual alignments are finally performed with dynamic parameters inserted to adapters, thus allowing the model to encode captions while explicitly accounting for the overall characteristics observed in the textual inputs.

To the best of our knowledge, this is the first work that leverages disentangled semantics to generate dynamic adapters to improve the target-language text encoding in CCR. Our main contributions are summarized as follows:

- We identify the problem of performing accurate text encoding against challenges caused by the diversity in written expression of target-language training samples in CCR, and provide an effective solution based on data-dependent semantics disentanglement.
- We propose a novel parameter-efficient diagram for CCR which dynamically generates parameters of input-aware adapters, enabling the CCR models to encode target-language sentences adaptively.
- We achieve a new state-of-the-art performance on two image-text retrieval datasets and one video-text retrieval dataset. Besides, our model shows good compatibility with various VLP models.

Related Work

Cross-lingual Cross-modal Retrieval

Cross-lingual cross-modal retrieval (CCR) is a method for achieving visual and target language (V-T) alignment without using any manually-annotated visual-text data pairs. This approach can be seen as a specific case of transfer learning to new domains (Fang et al. 2024) with limited resources (Zhang et al. 2020a), helping to mitigate the scarcity of training data for low-resource languages in traditional cross-modal retrieval (Dong et al. 2022a,b; Fang et al. 2023; Zheng et al. 2023). Early works (Aggarwal and Kale 2020; Portaz et al. 2019) on CCR try to transfer the knowledge of English models to low-resource languages by directly fine-tuning the model on MT-generated parallel data. Recently, V+L pretraining models have become popular, aiming to further narrow the gap between different languages and modalities. Among them, M^3P (Ni et al. 2021) learns universal representations that can map objects occurred in different modalities or texts expressed in different languages into a common semantic space. UC^2 (Zhou et al. 2021) translates source-language annotations into the target language automatically and proposes fine-grained pretraining objectives to encourage alignment between image regions and multilingual tokens. Following UC^2, MURAL (Jain et al. 2021) leverages 1.8 billion noisy image-text pairs to pre-train their dual encoder model. After that, CCLM (Zeng et al. 2023) proposes a cross-view language modeling framework, which considers both multi-modal data and multi-lingual data as pairs of two different views of the same object and propose a unified framework to fuse features in different views. Although CCLM successfully outperforms UC2 and MURAL on several benchmarks, it is very expensive to expand CCLM to support new low-resource languages since its pretraining stage require large-amount data and computing resources. Instead of pretraining V+L models from scratch,

[1]https://translate.google.com/

some recent works (Wang et al. 2022, 2024a,b; Cai et al. 2024) try to finetune the upper layers of existing VLP models with machine-translated data, which inevitably leading to a certain degree of knowledge forgetting.

Although these works have achieved improvements on CCR, their methods still require full-model training, which is quite time-consuming and demands significant computational power, making them impractical for researchers with limited hardware resources.

Parameter-Efficient FineTuning for CCR

The pretraining and finetuning paradigms have been proven to be highly effective in different language and vision tasks. Compared to full fine-tuning, Parameter-Efficient FineTuning (PEFT) is more suitable for cases with limited hardware resources, as it freezes the majority of the parameters of the pretrained model while still being able to demonstrate comparable performance in downstream tasks. Various PEFT techniques have been explored, including prompt tuning (Li and Liang 2021; Liu et al. 2024; Wu, Jiang, and Lian 2024; Zhou et al. 2022), Low-Rank Adaptation (LoRA) (Hedegaard et al. 2024; Mao et al. 2024), and adapters (Zhang, Hu, and Jin 2022). In which, the core idea of adapters is to insert light-weight adaptation modules into each layer of the pretrained transformer (Vaswani et al. 2017), and they have been extended across numerous domains. For example, MAD-X (Pfeiffer et al. 2020) extends multilingual pretraining models to support low-resource languages through adapters. Following MAD-X, MAD-G (Ansell et al. 2021) is proposed to generate language adapters based on type characteristics in language representations.

Recently, MLA (Zhang, Hu, and Jin 2022) designs a light-weight language acquisition encoder that supports low-resource languages through language-specific adapters. This approach is somewhat similar to our idea but has some core differences: MLA overlook the diversity of written expression and the noise in translated training samples during the cross-lingual transfer. In contrast, our DASD learns disentangled characteristics of input captions to help the model understand sentences in different styles of expression.

Semantics Disentangling

Recently, learning disentangled representations has been widely applied to a wide spectrum of applications ranging from domain adaption (Cai et al. 2019; Zhang et al. 2024) to text-to-image generation (Yin et al. 2019) and zero-shot learning (Chen et al. 2021; Ye et al. 2021). The core idea behind these work is to factorize input features into semantic-related and semantic-unrelated representations, so that the disentangled semantic-related features could be adapted across domains, modalities or tasks. For example, the prior work (Yin et al. 2019) focusing on text-to-image generation distills semantic commons from the linguistic descriptions, based on which the generated images can keep generation consistency under expression variants. Different from these previous approaches, in this paper, semantic-unrelated representations also play an important role during the transfer across languages, which are utilized for the dynamic adapter generation to improve the semantics extraction of target-language captions.

The Proposed Method

In this paper, we propose a dynamic adapter generation framework with semantics disentangling for CCR. As shown in Figure 2, our framework consists of three key components: 1) a pretrained VLP model as the backbone of our framework whose parameters stay frozen; 2) an input-aware parameter generator which analyzes the characteristics of the target-language input and produces a parameter matrix of adapters accordingly; 3) dynamic adapters inserted to each layer of the frozen VLP model to adaptively empower it with the cross-lingual ability.

Task Definition

We first formally define the setting of CCR, which involves two kinds of languages, namely the source language and the target languages. For the source language S, we have a collection of human-labeled training data $\mathcal{D}^S = \{d_1, d_2, ..., d_n\}$, where each instance d_i consists of a caption S_i^S paired with an image or video V_i. As for the target language T, due to the scarcity of human-labeled data, we assume there are no extra labeled data of text-image/video pairs. The core task of CCR is to obtain a model applicable in the target language T, without using any manually annotated target-language data.

Pretrained VLP Model

Following MLA (Zhang, Hu, and Jin 2022), we choose CLIP (Radford et al. 2021) as the VLP model used in our DASD. It is worth noting that other VLP models can also be applied to our method.

Source-language Text Encoding. Given a sentence S^S in the source language, the corresponding sentence representation $r^S = \Phi^S(S^S; \theta^S)$ is generated through the pretrained text encoder Φ^S, which contains a embedding block and L transformer layers. To preserve the cross-model knowledge of VLP, θ^S keeps fixed during training. Concretely, the input sentence S^S is tokenized and processed into word embeddings $E^S = [e_{0=\texttt{[SOS]}}, ..., e_{M=\texttt{[EOS]}}]$ through the embedding block, where $\texttt{[SOS]}$ and $\texttt{[EOS]}$ are special tokens denoting the boundary of the input sentence. The word embeddings are then fed to the parameter-frozen CLIP's text encoder. The final representation r^S is obtained by performing a linear projection on the last hidden state of the $\texttt{[EOS]}$ token.

Visual Encoding. The Vision Transformer (ViT) (Dosovitskiy et al. 2020; Zhang et al. 2020b) is used as a kind of CLIP image encoder, which takes image patches as input and generates the final feature through a Transformer-based model. For image encoding, given an image V, it is divided into patches $V^{'} = [v_1^{'}, ..., v_N^{'}]$ following ViT. Then, they are linearly projected into patch embeddings $E_p = [e_{\texttt{[CLASS]}}, W_p v_1^{'}, ..., W_p v_N^{'}]$, where $e_{\texttt{[CLASS]}}$ is a special embedding for the whole image and W_p is the linear projection. The hidden states calculation is similar with the text encoder, and the final visual representation r^V is obtained

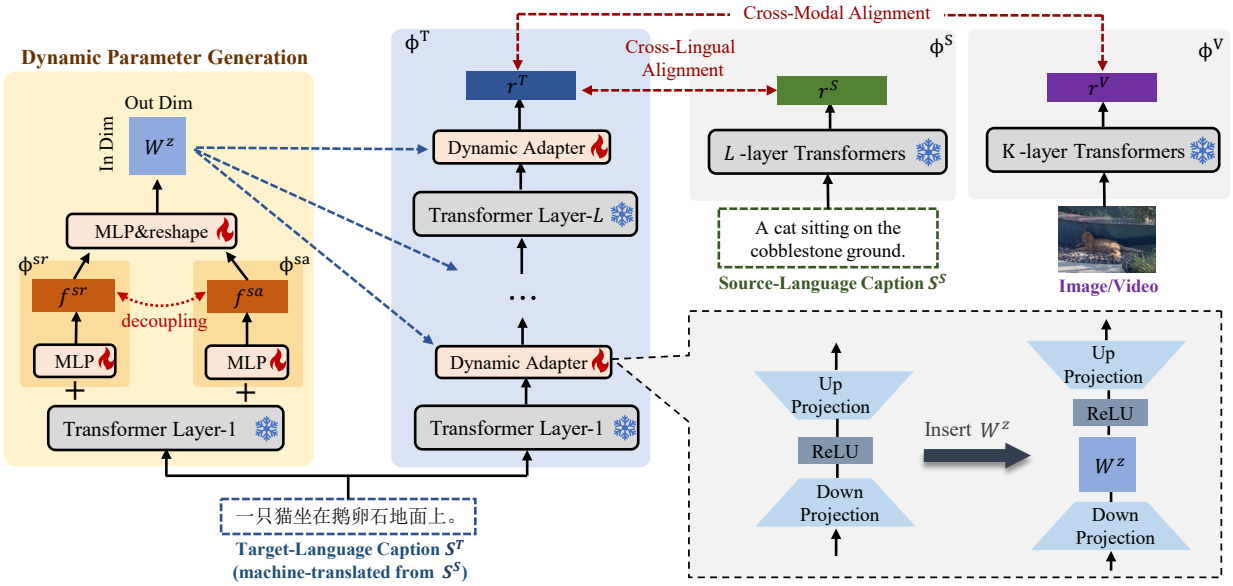


Figure 2: The illustration of our proposed Dynamic Adapter with Semantics Disentangling (DASD). To make dynamic adapters in the target-language branch Φ^T exactly match its input S^T, semantics disentangling is performed to extract semantic-related and semantic-agnostic features (f^{sr} and f^{sa}) from S^T and then generate input-conditional parameters (shown in the leftmost branch). The source-language branch Φ^S and visual branch Φ^V are provided by the frozen VLP model.

by performing a linear projection on the last hidden state of the $e_{[\texttt{CLASS}]}$ token: $r^V = W^b h_0^l$. As for video encoding, following the prior work (Luo et al. 2022), we uniformly sample 12 frames from the video and then perform average pooling over these frame representations to obtain the final video representation.

Input-aware Dynamic Adapters

Given a target-language caption S^T as the input, shown in the leftmost branch in Figure 2, our method generates an input-conditional parameter matrix, which plays a key role in S^T encoding. To generate parameters exactly match S^T, we propose semantics disentangling to extract semantic-related and semantic-agnostic features from S^T.

Semantics Disentangling. In the case of cross-lingual transfer, semantic-agnostic characteristics (such as the word order and the way of expression) in different languages tend to vary greatly, which can hardly be captured by static adapters. As a result, in our framework, we employ a semantics disentangling module to obtain semantic-agnostic features of captions in different target languages. Semantics disentangling performs feature extraction in a parameter-efficient way, whose backbone is only the first frozen layer of Φ^S equipped with two trainable lightweight adapters. Taking S^T as the input, semantic-related and semantic-agnostic features are extracted through semantic consistency learning and adversarial training, respectively.

Semantic-Related Features Extraction. As shown in Figure 2, the extraction of semantic-related features is performed by the module Φ^{sr}, whose input is the target-language sentence S^T which has been tokenized and processed into word embeddings $E^T = [u_{0=[\texttt{SOS}]}, ..., u_{M=[\texttt{EOS}]}]$. These embeddings are then fed to the first layer of parameter-frozen CLIP's text encoder equipped with the semantic-related adapter $\texttt{A}^{\texttt{sr}}$:

$$X_0^{sr} = [W_e^{sr} u_0, W_e^{sr} u_1, ..., W_e^{sr} u_M] + E_{pos} \quad (1)$$

$$H_1^{sr} = \texttt{TransformerLayer}(X_0^{sr}) \quad (2)$$

$$X_1^{sr} = \texttt{A}^{\texttt{sr}}(H_1^{sr}; \theta_1^{sr}) + H_1^{sr} \quad (3)$$

where X_1^{sr} is the hidden state of the pretrained transformer layer and W_e^{sr} is a linear projection to keep dimension consistency with source-language embeddings. The semantic-related adapter $\texttt{A}^{\texttt{sr}}$ in Equation 3 is implemented as a bottleneck MLP with residual connection:

$$\texttt{A}^{\texttt{sr}}(X) = W_{upper}^{sr} \texttt{ReLU}(W_{down}^{sr} X) \quad (4)$$

Similar with Φ^S, the last hidden state of the [EOS] token in the pretrained layer is linearly projected into the semantic-related feature f^{sr}:

$$f^{sr} = W_p^{sr} x_{1,[\texttt{EOS}]}^{sr} \quad (5)$$

Considering the fact that S^T shares the same semantics with its source-language counterpart S^S, we propose semantic consistency learning to explicitly transfer the semantic information gathered by Φ^S from S^S to f^{sr}. As shown in Equation 6, the semantic consistency loss $\mathcal{L}_{sc}$ is defined as the $L1$ distance between r^S and f^{sr}:

$$\mathcal{L}_{sc} = ||r^S - f^{sr}|| \quad (6)$$

Semantic-Agnostic Features Extraction. The semantic-agnostic module Φ^{sa} shares the same backbone with Φ^{sr},

equipped with the semantic-agnostic adapter $\texttt{A}^{\texttt{sa}}$ which works in the similar way with $\texttt{A}^{\texttt{sr}}$. Different from Φ^{sr}, Φ^{sa} produces the semantic-agnostic feature f^{sa} by performing average-pooling over all hidden states in the pretrained transformer layer.

To achieve perfect semantics disentangling, the semantic-agnostic feature f^{sa} extracted from S^T should exclude any semantic information about S^T. Motivated by this, we enforce f^{sa} to be useless to identify its corresponding semantic representations through adversarial training. Specifically, for adversarial training, we construct two feature pairs:(f^{sa},r^S) and (f^{sa},r^{S-}). The former is regarded as the positive sample since r^S is the semantic representation of S^S which shares the same semantics with S^T. The latter is the negative sample and r^{S-} is the semantic representation of a randomly-selected source-language caption. We employ a classifier F to act as the discriminator which is adopted to distinguish the positive and negative samples. The classifier is consisted of a multi-layer feed-forward neural networks, and the discrimination loss in adversarial training is defined as:

$$\mathcal{L}_d = -logF(f^{sa}, r^S) - log(1 - F(f^{sa}, r^{S-}) \quad (7)$$

The parameters of Φ^{sa} are updated to confuse the discriminator by minimizing the loss $\mathcal{L}_{adv} = -\mathcal{L}_d$.

Input-conditional Parameter Generation. After obtaining f^{sr} and f^{sa}, we adopt a multilayer perceptron to extract the global information z of S^T, i.e.,

$$z = MLP(f^{sr} \circ f^{sa}) \quad (8)$$

where $\circ$ means the concatenation operation. Then, the dynamic parameter-matrix W_i^z of layer i are obtained using a single layer linear down-projection:

$$W_i^z = reshape(W_i^{down} z) \quad (9)$$

where $W_i^{down} \in \mathbb{R}^{d_u^2 \times d_z}$, $W_i^z \in \mathbb{R}^{d_u \times d_u}$, and the operation $reshape$ refers to reshaping the vectors produced by the MLP into a matrix form. Down projecting to a dimension $d_u << d_z$ prevents W_i^{down} from being impractically large, keeping our model parameter-efficient. In this way, W^z is dynamically generated conditioned on the target-language input S^T, which are then inserted to adapters in the target-language branch Φ^T.

CCR with Dynamic Adapters

The core of CCR is to align the target-language sentence S^T with its source-language counterpart S^S as well as its visual counterpart V. However, due to the scarcity of human-labeled S^T-V pairs, MT tools are employed to translate S^S into the target language. With access to these paired data from different sources, cross-lingual and cross-modal alignments are performed accordingly.

Target-Language Text Encoding. As shown in Figure 2, the representation of S^T is calculated by the target-language branch Φ^T, which taking word embeddings E^T as the input and extract semantics of S^T at each layer with the help of the dynamic adapter $\texttt{DA}$:

$$X_0^T = [W_e^T u_{0=[\texttt{SOS}]}, ..., W_e^T u_{M=[\texttt{EOS}]}] + E_{pos} \quad (10)$$

$$H_i^T = \texttt{TransformerLayer}(X_{i-1}^T) \quad (11)$$

$$X_i^T = \texttt{DA}(H_i^T; \theta_i^{\texttt{DA}}) + H_i^T \quad (12)$$

where $\theta_i^{\texttt{DA}}$ refers to the parameter of $\texttt{DA}$ in i-th layer and works as follows:

$$\texttt{DA}(X) = W_{upper}^d \texttt{ReLU}(W^z W_{down}^d X) + X \quad (13)$$

Here, $\{W_{upper}^d, W_{down}^d, W^z\} \in \theta^{\texttt{DA}}$. Finally, the last hidden state of the $[\texttt{EOS}]$ token is linearly projected into the semantic representation of S^T: $r^T = W_p x_{[\texttt{EOS}]}^T$. The linear projection W_p is shared with CLIP's text encoder and keeps frozen during the training.

Training Strategy. Considering the the scarcity of target-language resources, following MLA (Zhang, Hu, and Jin 2022), the cross-lingual alignment and the cross-modal alignment are performed independently in our framework. The motivation behind the separate training is to ensure that cross-lingual transfer can always proceed smoothly in case data in a certain modality is missing or of poor quality. The objective in the cross-lingual alignment is minimizing the Mean Square Error (MSE) between the native representation r^S and the non-native representation r^T:

$$\mathcal{L}_{CL} = ||r^S - r^T||^2 \quad (14)$$

As for the cross-modal alignment, it is achieved by performing contrastive learning between target languages and images. The training objective is minimizing the NCE loss (Gutmann and Hyvärinen 2010) defined as follows:

$$\begin{aligned} \mathcal{L}_{CM} = &- log \frac{exp(sim(r^T, r^V))}{\sum_{j=1}^B exp(sim(r_j^T, r^V))} \\ &- \log \frac{exp(sim(r^T, r^V))}{\sum_{j=1}^B exp(sim(r^T, r_j^V))} \end{aligned} \quad (15)$$

where B is the batch size, $sim(\cdot)$ denotes the similarity function (i.e., cosine similarity) and τ is the temperature coefficient. Our model is trained by minimizing the combination of the above losses. Finally, the total loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{CL} + \mathcal{L}_{CM} + \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{sc} \quad (16)$$

where λ_1 and λ_2 are hyper-parameters to balance the importance of disentangling losses.

Experiments

Experimental Settings

Datasets. Evaluations are performed on two image-text retrieval datasets (Multi30K (Elliott et al. 2016) and MSCOCO (Chen et al. 2015)) and a video-text retrieval dataset (MSRVTT (Xu et al. 2016)), referred as ***downstream task datasets (DTD)*** in this paper. Target-language captions are obtained by automatically translating the English captions in DTD with Google Translate. Besides, the web-scraped image-caption dataset CC3M (Sharma et al. 2018) with machine-translated captions is also used for training, from which 300k image-captions pairs are randomly selected and known as CC300K (Zhang, Hu, and Jin 2022).

Retrieval Settings. We conduct experiments under two CCR settings: (1) ***Cross-lingual Finetune***: we first train models using English data in DTD and then further finetune models with target-language data produced by MT tools. Finally, models are tested on DTD target-language datasets. (2) ***Zero-shot***: models are trained on commonly-used datasets (e.g., CC300K) and then directly evaluated on DTD without any DTD finetuning.

Evaluation Metrics. For image-text retrieval, following (Zhang, Hu, and Jin 2022), we report the mean Average Recall (mAR) for image-text retrieval. For video-text retrieval, we follow (Rouditchenko et al. 2023) and use text→video Recall@1 score to evaluate the performance.

Evaluation on Cross-lingual Image-Text Retrieval

Under the ***Cross-lingual Finetune*** setting, image-caption pairs for target languages are obtained in two separate ways: (1) we directly leverage the target-language data in CC300K (following MLA (Zhang, Hu, and Jin 2022)). (2) English captions in Multi30K and MSCOCO are converted into target languages utilizing Google Translate (following CL2CM (Wang et al. 2024a)). In both cases, it can be seen in Table 1 that DASD outperforms all the comparison methods, demonstrating the effectiveness of dynamic adapters. Please note that the SOTA model CL2CM relies on the full-model training, besides, its cross-lingual alignments is carefully designed where token-level alignments are involved to improve the final performance, all of which consumes a large amount of computing budgets. Although DCOT and CL2CM are not open-sourced, they all expand the cross-attention module in NRCCR and therefore are expected to have more trainable parameters than NRCCR. Under the ***Zero-shot*** setting, we observe that the performances of NRCCR, DCOT and CL2CM drop severely due the absence of downstream datasets, which are surpassed by the parameter-efficient model MLA. Among them, DCOT (Wang et al. 2024b) tries to learn noisy correspondence in CCR by quantifying the confidence of the sample pair correlation with optimal transport theory from both the cross-lingual and cross-modal views. Compared with PEFT models, DCOT relies on full-model training, rendering their method much more time and computing consuming. Besides, DCOT only focus on reducing the impact of obvious errors brought by MT, neglecting other factors (e.g., language gaps) which could also hurt the cross-lingual alignment and result in degraded performance. Our performance still achieves the best performance when downstream task data is not available, showing strong zero-shot cross-lingual transfer ability.

Evaluation on Cross-lingual Video-Text Retrieval

For cross-lingual video-text retrieval, experiments are conducted on MSRVTT (Xu et al. 2016) under the same settings with cross-lingual image-text retrieval, where the model searches for the most semantically relevant videos given a text query in a low-resource language. We report the text→video Recall@1 score in Table 2. Under both the ***Zero-shot*** and ***Cross-lingual Finetune*** settings, we simply adopt the same hyperparameter values and training strategy used for the cross-lingual image-text retrieval. As shown in Table 2, for both settings, our model consistently outperforms MLA on all eight target languages, demonstrating a strong cross-lingual ability for text-video retrieval.

Generalizability Analysis

Since the adapter is a lightweight, plug-and-play module, we also investigate whether our proposed DASD is compatible with different VLP models. To this end, we substitute the pretrained CLIP with the recently-proposed M-VLP model CCLM (Zeng et al. 2023), which has been pretrained on the combination of image-caption pairs and parallel corpora. Specifically, since CCLM is a single-stream model and difficult to extend directly, we follow (Wang et al. 2024a) to modify CCLM into a dual-stream model and apply DASD to its text encoder. As reported in Table 3, the cross-lingual text-image retrieval performance of CCLM is further improved when equipped with our proposed dynamic adapters, outperforming the framework built upon CLIP. These results verify that our method is compatible with various VLP models and could achieve a higher performance when equipped with stronger VLP models.

Ablation Studies

To verify the effectiveness of each component in DASD, we conduct ablation studies under the cross-lingual finetune setting on Multi30K and MSCOCO.

The effectiveness of input-conditional parameters. We first investigate the contribution of the dynamic parameters by removing the input-conditional parameter matrix W^z out of DASD, turning it into the traditional adapter with only static parameters. As reported in Table 4, when using static adapters for cross-lingual transfer, we observe a severe performance degradation on all five target languages, demonstrating the importance of the dynamic parameters.

The effectiveness of semantic-related and semantic-agnostic features. We then study the effectiveness of f^{sa} and f^{sr} to dynamic parameter generation. As summarized in Table 5, the inclusion of both kinds of features leads to a certain improvement in mAR on all five target languages. In the case where only one kind of features is employed, we observe that the semantic-related features having a slightly greater positive impact compared to the semantic-agnostic ones. It not only demonstrates the effectiveness of both kinds of features, but also shows the complementary of the semantic-related and semantic-agnostic features.

The impact of semantics disentangling. To examine the necessity of performing semantics disentangling, we investigate the impact brought by different disentangling losses ($\mathcal{L}_{adv}$ and $\mathcal{L}_{sc}$) and report results in Table 6. We observe that with the loss constraints added, the model performance increases on all five languages, validating the effectiveness of the adversarial training and semantic distillation. To our knowledge, no prior work has applied dynamic adapters or semantics disentangling to CCR, and our work fills this gap and thereby gains a clear improvement.

	Method	#TP	Training Data		Multi30K			MSCOCO	
			English	Target Languages	DE	FR	CS	ZH	JA
Cross-lingual Finetune	UC^2 (Zhou et al. 2021)	478M	DTD	CC3M	83.8	77.6	74.2	82.0	71.7
	MURAL (Jain et al. 2021)	300M	DTD	CC300K	76.5	76.7	70.1	-	74.6
	MLA (Zhang, Hu, and Jin 2022)	108M	DTD	CC300K	86.4	87.3	79.5	-	80.4
	DASD (ours)	134M	DTD	CC300K	**87.4**	**88.6**	**83.4**	**88.5**	**84.8**
	NRCCR (Wang et al. 2022)	216M	DTD	MT(DTD)	80.1	80.4	77.9	85.4	84.5
	DCOT (Wang et al. 2024b)	-	DTD	MT(DTD)	82.5	82.6	80.3	86.9	85.9
	CL2CM (Wang et al. 2024a)	-	DTD	MT(DTD)	83.0	83.3	80.9	87.0	86.0
	DASD (ours)	134M	DTD	MT(DTD)	**88.5**	**91.1**	**87.6**	**90.0**	**89.1**
Zero-shot	UC^2 (Zhou et al. 2021)	478M	-	CC3M	62.5	60.4	55.1	-	62.3
	MURAL (Jain et al. 2021)	300M	-	CC300K	62.7	60.8	57.5	-	62.5
	MLA (Zhang, Hu, and Jin 2022)	108M	-	CC300K	80.8	80.9	72.9	78.5	76.7
	DASD (ours)	134M	-	CC300K	**81.9**	**82.1**	**74.3**	**79.6**	**77.5**
	NRCCR (Wang et al. 2022)	216M	-	MT(MSCOCO)	74.8	72.3	68.5	-	-
	DCOT (Wang et al. 2024b)	-	-	MT(MSCOCO)	76.5	74.2	70.7	-	-
	CL2CM (Wang et al. 2024a)	-	-	MT(MSCOCO)	76.9	74.5	71.5	-	-
	DASD (ours)	134M	-	MT(MSCOCO)	**80.1**	**81.3**	**74.9**	-	-

Table 1: Cross-lingual image-text retrieval results on Multi30K and MSCOCO. #TP: *the number of Trainable parameters*, DTD: *Down-stream Task Datasets* (i.e., Multi30K and MSCOCO). Despite being equipped the dynamic parameter generator, the number of trainable parameters in DASD remains comparable with MLA using the static adaptor, maintaining the parameter efficiency of the model while achieving significant improvements under both settings.

	Method	DE	FR	CS	ZH	RU	VI	SW	ES	SUM
CL-FT	MMP (Huang et al. 2021)	21.1	21.8	20.7	20.0	20.5	10.9	14.4	21.9	151.3
	C2KD (Rouditchenko et al. 2023)	24.7	25.4	24.0	23.4	23.1	13.6	20.3	25.5	180.0
	MLA (Zhang, Hu, and Jin 2022)	26.1	26.7	20.5	25.3	18.9	12.9	12.6	27.2	170.2
	DASD (ours)	**28.8**	**30.5**	**26.3**	**28.0**	**25.9**	**14.8**	**22.1**	**29.7**	**206.1**
ZS	MMP (Huang et al. 2021)	19.4	20.7	19.3	18.2	19.1	8.2	8.4	20.4	133.7
	MLA (Zhang, Hu, and Jin 2022)	20.1	22.0	15.7	18.3	14.4	8.2	10.7	20.2	129.6
	DASD (ours)	**23.7**	**23.9**	**21.4**	**22.4**	**21.7**	**11.2**	**15.3**	**23.1**	**162.7**

Table 2: Cross-lingual video-text retrieval results on Multi-MSRVTT, CL-FT: *Cross-lingual Fine-tune*, ZS: *Zero-shot*. Our proposed DASD performs the best over baseline methods on all target languages.

Method	Multi30K			MSCOCO		SUM
	FR	DE	CS	ZH	JA	
CLIP (Radford et al. 2021)	-	-	-	-	-	-
CLIP+ours	91.1	88.5	87.6	90.0	89.1	446.3
CCLM (Zeng et al. 2023)	81.7	83.9	80.2	85.2	82.7	413.7
CCLM+ours	**91.4**	**89.3**	**88.5**	**91.9**	**91.6**	**452.7**

Table 3: The performances of our method using different VLP models as the backbone. Our dynamic adapter could not only expand the monolingual VLP model (CLIP) to multiple target languages, but also exhibits a good compatibility with different VLP models.

Method	Multi30K			MSCOCO		SUM
	FR	DE	CS	ZH	JA	
Traditional Adapter	88.3	87.0	83.1	85.3	85.7	429.4
Dynamic Adapter (ours)	**91.1**	**88.5**	**87.6**	**90.0**	**89.1**	**446.3**

Table 4: Effectiveness of the dynamic adapter for CCR on Multi30K and MSCOCO. Using input-conditional parameters (W^z) brings in substantial performance gain.

Visualization Analysis

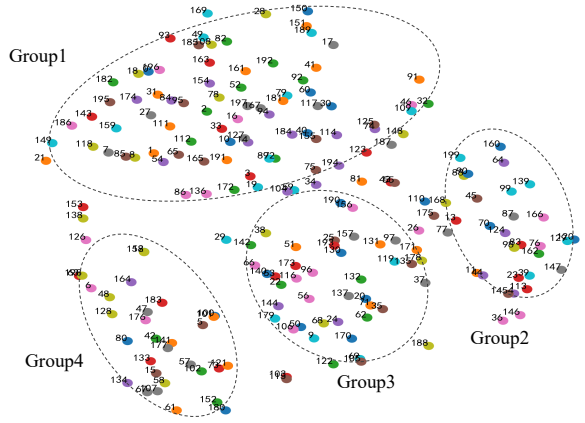
In Figure 3, we use t-SNE to visualize the semantic-agnostic representations of 200 Chinese sentences randomly selected from MSCOCO testset. As illustrated in Figure 3(a), the semantic-agnostic representations produced by DASD have been automatically clustered into 4 groups. Figure 3(b) lists some corresponding sentences of each group, and we observe that sentences in the same group are expressed in a similar way. Concretely, sentences in group 1 start with a quantifier followed by the key object, and sentences in group 2 begin with the location of the key object. Compared to group 1, sentences in group 3 include an additional descriptive modifier before the key object. Group 4 contains long sentences that are divided into two parts. This

Features		Multi30K			MSCOCO		SUM
f^{sr}	f^{sa}	FR	DE	CS	ZH	JA	
✓	✓	**91.1**	**88.5**	**87.6**	**90.0**	**89.1**	**446.3**
✓	×	90.2	87.7	86.8	89.1	88.0	441.8
×	✓	89.9	87.8	86.3	88.7	88.1	440.8

Table 5: Effectiveness of the semantic-related and semantic-agnostics features (f^{sr} and f^{sa}).

loss		Multi30K			MSCOCO		SUM
$\mathcal{L}_{sc}$	$\mathcal{L}_{adv}$	FR	DE	CS	ZH	JA	
✓	✓	**91.1**	**88.5**	**87.6**	**90.0**	**89.1**	**446.3**
✓	×	90.7	87.9	87.3	89.6	88.3	443.8
×	✓	90.6	88.1	86.9	89.5	88.5	443.6
×	×	90.1	87.4	86.5	89.1	88.0	441.1

Table 6: Effectiveness of different disentangling losses.



(a) Visualization of semantic-agnostic features.

Group1: Quantifier + Object + …
174: 一群长颈鹿站在一棵大树边。
(A group of giraffes stand near a big tree.)
33: 一辆卡车上面到处都是涂鸦。
(A truck with graffiti all over it.)
1: 一只猫有站在两只毛绒动物旁边。
(A cat stands next to two stuffed animals.)

Group2: Location + Object
147: 盘子里有一块面包和沙拉。
(Bread and salad on plate.)
124: 院子里摆放着火炉和椅子。
(The yard has a stove and chairs.)
98: 一个金属架子上有牙刷。
(A metal rack holds toothbrushes.)

Group3: Quantifier + Descriptive Modifier + Object + …
193: 一位身穿牛仔裤的男子用脚将滑板翘起。
(A man in jeans lifts his skateboard with his foot.)
24: 一个撑着一把伞的女人走在一条小路上。
(A woman holding an umbrella walks on a path.)
170: 一个戴着领带的男人在街上走着。
(A man wearing a tie is walking on the street.)

Group4: Event + Comma + Supplementary Information
61: 这是一幅画，里面一个男孩和一个女孩坐在海滩上的一把椅子上。
(This is a painting, in which a boy and a girl are sitting on a chair on the beach.)
80: 网球场上一名运动员正在打比赛，旁边有裁判和观众。
(A player is playing a game on the tennis court, with referees and spectators nearby.)
57: 一个金发男孩坐在桌前，手里拿着一根烤肠正在吃早餐。
(A blond boy is sitting at the table, holding a grilled sausage and eating breakfast.)

(b) The corresponding Chinese sentences in each group.

Figure 3: Visualization of the semantic-agnostic features of 200 randomly-selected MSCOCO Chinese sentences.

visualization result confirms that our DASD effectively captures semantic-agnostic characteristics of captions through semantics disentangling.

Conclusion

This paper proposes dynamic adapters with semantics disentangling for CCR. By characterizing target-language captions from two distinct and complementary aspects, our DASD dynamically generates adapters for input captions in varied forms. Extensive experiments show the effectiveness of DASD and its new SOTA performance. Given DASD is simple and effective, we believe it can also be used as a new strong baseline for other cross-lingual transfer tasks.

Acknowledgments

This work was supported by the Pioneer and Leading Goose R&D Program of Zhejiang (No. 2024C01110), the National Natural Science Foundation of China (No. 62306278), the Zhejiang Provincial Natural Science Foundation (No. LZ23F020004 and No. LQ23F020008), the Young Elite Scientists Sponsorship Program by China Association for Science and Technolog y (No. 2022QNRC001), and the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. FR2402ZD).

References

Aggarwal, P.; and Kale, A. 2020. Towards zero-shot cross-lingual image retrieval. *arXiv preprint arXiv:2012.05107*.

Ahmad, W. U.; Zhang, Z.; Ma, X.; Chang, K.-W.; and Peng, N. 2019. Cross-Lingual Dependency Parsing with Unlabeled Auxiliary Languages. In Bansal, M.; and Villavicencio, A., eds., *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 372–382. Association for Computational Linguistics.

Ansell, A.; Ponti, E. M.; Pfeiffer, J.; Ruder, S.; Glavaš, G.; Vulić, I.; and Korhonen, A. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4762–4781.

Bogolin, S.-V.; Croitoru, I.; Jin, H.; Liu, Y.; and Albanie, S. 2022. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5194–5205.

Cai, R.; Dong, J.; Liang, T.; Liang, Y.; Wang, Y.; Yang, X.; Wang, X.; and Wang, M. 2024. Cross-Lingual Cross-Modal Retrieval with Noise-Robust Fine-Tuning. *IEEE Transactions on Knowledge and Data Engineering*.

Cai, R.; Li, Z.; Wei, P.; Qiao, J.; Zhang, K.; and Hao, Z. 2019. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, volume 2019, 2060. NIH Public Access.

Chang, T.; Yang, X.; Luo, X.; Ji, W.; and Wang, M. 2023. Learning Style-Invariant Robust Representation for Generalizable Visual Instance Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6171–6180.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chen, Z.; Luo, Y.; Qiu, R.; Wang, S.; Huang, Z.; Li, J.; and Zhang, Z. 2021. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8712–8720.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022a. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 246–257.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2022b. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4065–4080.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Leanring Representations*.

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, 70–74.

Fang, X.; Easwaran, A.; Genest, B.; and Suganthan, P. N. 2024. Your Data Is Not Perfect: Towards Cross-Domain Out-of-Distribution Detection in Class-Imbalanced Data. *ESWA*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *CVPR*.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Hedegaard, L.; Alok, A.; Jose, J.; and Iosifidis, A. 2024. Structured pruning adapters. *Pattern Recognition*, 156: 110724.

Huang, P.-Y.; Patrick, M.; Hu, J.; Neubig, G.; Metze, F.; and Hauptmann, A. G. 2021. Multilingual Multimodal Pretraining for Zero-Shot Cross-Lingual Transfer of Vision-Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2443–2459.

Jain, A.; Guo, M.; Srinivasan, K.; Chen, T.; Kudugunta, S.; Jia, C.; Yang, Y.; and Baldridge, J. 2021. MURAL: multimodal, multitask retrieval across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3449–3463.

Li, H.; Zhang, C.; Jia, X.; Gao, Y.; and Chen, C. 2021. Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(2): 1185–1199.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2024. GPT understands, too. *AI Open*, 5: 208–215.

Lu, H.; Fei, N.; Huo, Y.; Gao, Y.; Lu, Z.; and Wen, J.-R. 2022. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, 15692–15701.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Mao, Y.; Huang, K.; Guan, C.; Bao, G.; Mo, F.; and Xu, J. 2024. DoRA: Enhancing Parameter-Efficient Fine-Tuning with Dynamic Rank Distribution. *arXiv preprint arXiv:2405.17357*.

Ni, M.; Huang, H.; Su, L.; Cui, E.; Bharti, T.; Wang, L.; Zhang, D.; and Duan, N. 2021. M3p: Learning universal representations via multitask multilingual multimodal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3977–3986.

Pfeiffer, J.; Vulić, I.; Gurevych, I.; and Ruder, S. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7654–7673.

Portaz, M.; Randrianarivo, H.; Nivaggioli, A.; Maudet, E.; Servan, C.; and Peyronnet, S. 2019. Image search using multilingual texts: a cross-modal learning approach between image and text. *arXiv preprint arXiv:1903.11299*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Rouditchenko, A.; Chuang, Y.-S.; Shvetsova, N.; Thomas, S.; Feris, R.; Kingsbury, B.; Karlinsky, L.; Harwath, D.; Kuehne, H.; and Glass, J. 2023. C2kd: Cross-lingual cross-modal knowledge distillation for multilingual text-video retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26: 824–836.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Dong, J.; Liang, T.; Zhang, M.; Cai, R.; and Wang, X. 2022. Cross-lingual cross-modal retrieval with noise-robust learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 422–433.

Wang, Y.; Wang, F.; Dong, J.; and Luo, H. 2024a. CL2CM: Improving Cross-Lingual Cross-Modal Retrieval via Cross-Lingual Knowledge Transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6): 5651–5659.

Wang, Y.; Wang, S.; Luo, H.; Dong, J.; Wang, F.; Han, M.; Wang, X.; and Wang, M. 2024b. Dual-view Curricular Optimal Transport for Cross-lingual Cross-modal Retrieval. *IEEE Transactions on Image Processing*, 33: 1522–1533.

Wu, C.; Jiang, G.; and Lian, D. 2024. Mitigate Negative Transfer with Similarity Heuristic Lifelong Prompt Tuning. *arXiv preprint arXiv:2406.12251*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 5288–5296.

Ye, Z.; Hu, F.; Lyu, F.; Li, L.; and Huang, K. 2021. Disentangling semantic-to-visual confusion for zero-shot learning. *IEEE Transactions on Multimedia*, 24: 2828–2840.

Yin, G.; Liu, B.; Sheng, L.; Yu, N.; Wang, X.; and Shao, J. 2019. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2327–2336.

Zeng, Y.; Zhou, W.; Luo, A.; Cheng, Z.; and Zhang, X. 2023. Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5731–5746.

Zhang, A.; Wang, H.; Wang, X.; and Chua, T.-S. 2024. Disentangling Masked Autoencoders for Unsupervised Domain Generalization. *arXiv preprint arXiv:2407.07544*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020a. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020b. Feature pyramid transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 323–339. Springer.

Zhang, L.; Hu, A.; and Jin, Q. 2022. Multi-Lingual Acquisition on Multimodal Pre-training for Cross-modal Retrieval. *Advances in Neural Information Processing Systems*, 35: 29691–29704.

Zhang, Z.; Luo, H.; Zhu, L.; Lu, G.; and Shen, H. T. 2023. Modality-Invariant Asymmetric Networks for Cross-Modal Hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 5091–5104.

Zheng, Q.; Dong, J.; Qu, X.; Yang, X.; Wang, Y.; Zhou, P.; Liu, B.; and Wang, X. 2023. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2): 1–21.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, M.; Zhou, L.; Wang, S.; Cheng, Y.; Li, L.; Yu, Z.; and Liu, J. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4155–4165.