mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video

Haiyang Xu [*] [1] Qinghao Ye [*] [1] Ming Yan [1] Yaya Shi [1] Jiabo Ye [1] Yuanhong Xu [1] Chenliang Li [1] Bin Bi [1]
Qi Qian [1] Wei Wang [1] Guohai Xu [1] Ji Zhang [1] Songfang Huang [1] Fei Huang [1] Jingren Zhou [1]

Abstract

Recent years have witnessed a big convergence of language, vision, and multi-modal pretraining. In this work, we present mPLUG-2 , a new unified paradigm with modularized design for multi-modal pretraining, which can benefit from modality collaboration while addressing the problem of modality entanglement. In contrast to predominant paradigms of solely relying on sequence-to-sequence generation or encoder-based instance discrimination, mPLUG-2 introduces a multi-module composition network by sharing common universal modules for modality collaboration and disentangling different modality modules to deal with modality entanglement. It is flexible to select different modules for different understanding and generation tasks across all modalities including text, image, and video. Empirical study shows that mPLUG-2 achieves state-of-the-art or competitive results on a broad range of over 30 downstream tasks, spanning multi-modal tasks of image-text and video-text understanding and generation, and uni-modal tasks of text-only, image-only, and video-only understanding. Notably, mPLUG-2 shows new state-of-the-art results of 48.0 top-1 accuracy and 80.3 CIDEr on the challenging MSRVTT video QA and video caption tasks with a far smaller model size and data scale. It also demonstrates strong zero-shot transferability on vision-language and video-language tasks. Code and models will be released in https://github.com/X-PLUG/mPLUG-2.

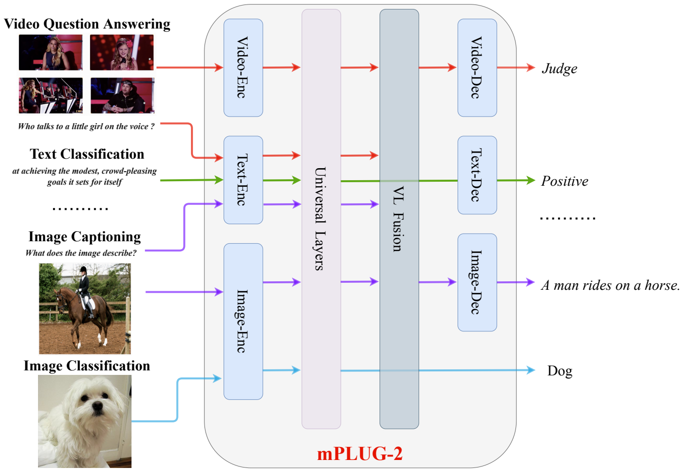


Figure 1. A brief illustration of the new paradigm with modularized design for building multi-modal foundation model.

1. Introduction

Large-scale pre-trained foundation models have been an emerging paradigm for a wide range of artificial intelligence (AI) fields, across language (Devlin et al., 2018; Brown et al., 2020), vision (Dosovitskiy et al., 2020; Liu et al., 2021b) and multi-modality (Radford et al., 2021; Yu et al., 2022; Wang et al., 2022e). With the broad success of Transformer architecture (Vaswani et al., 2017), recent years have featured a trend toward the big convergence of language, vision and multimodal pre-training (Yu et al., 2022; Wang et al., 2022e; Alayrac et al., 2022). One line along this trend proposes to unify the tasks and modalities with a unified sequence-to-sequence generation framework such as T5 (Raffel et al., 2020), OFA (Wang et al., 2022d) and Flamingo (Alayrac et al., 2022). On the other hand, BERT (Devlin et al., 2018), Florence (Yuan et al., 2021) and BEIT-3 (Wang et al., 2022e) models all the tasks as instance discrimination, and adopt the pure encoder-based architecture.

The predominant foundation models propose to share the same single network for multi-modality (Alayrac et al., 2022) to leverage the information from modality collaboration. However, the strategy will suffer from the issue of

[*]Equal contribution [1]DAMO Academy, Alibaba Group, China. Correspondence to: Ming Yan <ym119608@alibaba-inc.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Table 1. **A system-level comparison between mPLUG-2 and existing foundation models in terms of various uni-modal and multi-modal downstream tasks.** ”Cls.” denotes the classification. ”Det.” and ”Seg.” are the short for ”Detection” and ”Segmentation” tasks respectively. ”VG” stands for visual grounding task. Our mPLUG-2 is capable of supporting both uni-modal (i.e., CV and NLP) and multi-modal (i.e., Image-Text and Video-Text) downstream tasks simultaneously with the help of modularization.

	Computer Vision				Natural Language Processing			Image-Text				Video-Text		
Method	Image Cls.	Video Cls.	Det.	Seg.	Text Cls.	QA	Summarization	Retrieval	QA	Captioning	VG	Retrieval	QA	Captioning
BEiT-3	✓		✓	✓				✓	✓	✓				
EVA	✓	✓	✓	✓										
CLIP	✓							✓				✓		
ALBEF								✓	✓		✓			
BLIP								✓	✓	✓		✓	✓	
VATT	✓	✓										✓		
Florence	✓	✓	✓					✓	✓			✓		
CoCa	✓	✓						✓	✓	✓		✓		
VideoCoCa		✓												
Flamingo		✓						✓	✓	✓			✓	✓
GIT2	✓				✓				✓	✓			✓	✓
FLAVA	✓				✓	✓	✓		✓					
OFA	✓				✓	✓	✓	✓	✓	✓	✓			
OmniVL	✓	✓						✓	✓	✓		✓	✓	✓
mPLUG 2.0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

modality entanglement due to the large variance of different modality tasks. The challenge is that multiple modalities may interfere with each other (Huang et al., 2022b), especially when there are many modalities and tasks. It is difficult for a single-module foundation model to balance the gain of modality collaboration and the influence of modality entanglement on a large number of downstream tasks across multiple modalities.

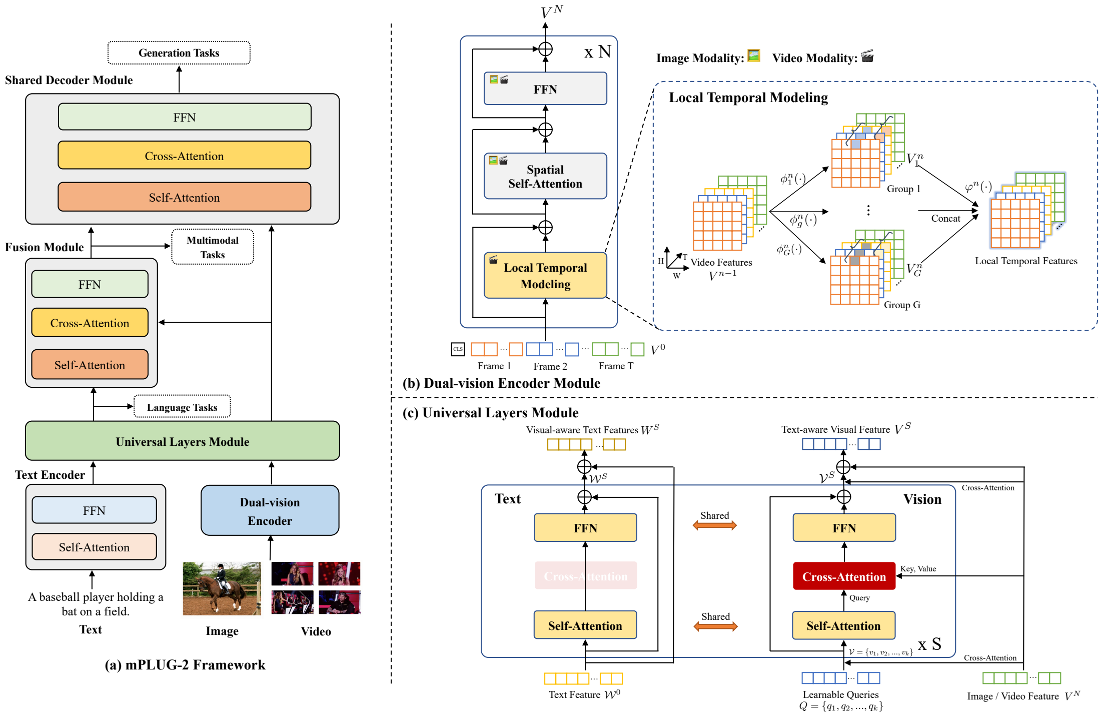
To alleviate the challenge, in this work, we introduce a new unified paradigm of multi-modal foundation models, as shown in Figure 1. It features a module-based network design considering both the modality collaboration and modality entanglement, where mPLUG-2 designs certain shared functional modules to encourage the modality collaboration, while reserving modality-specific modules to tackle the problem of modality entanglement. Different modules are then jointly trained effectively on both the uni-modal and multi-modal datasets according to the task's module design. As a result, different modules can be flexibly selected and combined for the large number of uni-modal and cross-modal understanding and generation tasks accordingly. The details of the supported downstream tasks are given in Table 1. To the best of our knowledge, the proposed method tackles the largest number of different kinds of downstream tasks across text, image and video.

Specifically, we design a unified dual-vision encoder module by disentangling spatial and temporal representations, where video inputs share the standard Transformer module with image inputs for modeling spatial information and an extra local temporal modeling module is used for temporal relation modeling on video-related tasks. Then a novel universal layers module is introduced to serve as a pivot across different modalities, where vision and language modalities are projected to the common language-guided semantic space by sharing self-attention modules. Besides, an extra cross-attention module is used to fuse the universal vision representation with the original fine-grained vision representation. The detailed module design is shown in Figure 2. Finally, different modules of mPLUG-2 are jointly pre-trained with task and modality instructions (Wang et al., 2022d) on both uni-modal and cross-modal tasks. During inference, mPLUG-2 can select different modules for various uni-modal and cross-modal tasks with the modularized Transformer architecture. The selected modules for different tasks can be found in Table 2 in Appendix.

We evaluate the new unified paradigm of mPLUG-2 on over 30 challenging uni-modal and cross-modal understanding and generation benchmarks and it achieves state-of-the-art or competitive results with a similar model size and data scale. Equipping with the module-based network design, mPLUG-2 can be also easily extended to additional tasks by selecting and adding modules. Notably, mPLUG-2 shows new state-of-the-art results of 48.0 top-1 accuracy and 80.3 CIDEr on the challenging MSRVTT video QA and video caption tasks, respectively. mPLUG-2 also demonstrates strong zero-shot transferability on vision-language and video-language tasks.

2. Related Work

Vision-only Foundation Models ConvNets (Szegedy et al., 2017; He et al., 2015) have long been the main stream visual architecture before the emergence of vision transformer (a.k.a. ViT) (Dosovitskiy et al., 2020). Due to the superior capacity of Transformer network, ViT stands out in various downstream tasks (Carion et al., 2020; Xu et al., 2022). Apart from scaling up the naive ViT architecture with large-scale dataset such as JFT-3B (Zhai et al., 2021), SwinV2-G (Liu et al., 2021a) extends the original ViT with hierarchical architectures. In addition, EVA (Fang et al., 2022a) distills the multi-modal knowledge to scale up ViT by leveraging unlabeled images with the large-scale pre-trained image-text model (e.g. CLIP (Radford et al., 2021)). Recently, InternImage (Wang et al., 2022f) revitalizes the convolutional



(a) mPLUG-2 Framework

(b) Dual-vision Encoder Module

(c) Universal Layers Module

Figure 2. The overall framework and module details of mPLUG-2 .

neural networks with deformable convolution and achieves the state-of-the-art performance on various vision downstream tasks. Besides, InternVideo (Wang et al., 2022g) extends to video tasks by assembling two large video models with both generative and discriminative self-supervised video learning.

Language-only Foundation Models Inspired by the successful practice of the BERT (Devlin et al., 2018) in natural language understanding, a massive large-scale language foundation models are proposed for natural language processing. BART (Lewis et al., 2020) is a denoising autoencoder like BERT but with encoder-decoder architecture which shows effectiveness for both text generation and comprehension tasks. Apart from BERT-series methods (Devlin et al., 2018; Lewis et al., 2020; Liu et al., 2019), there are numerous other effective architectures and pre-training objectives. T5 (Raffel et al., 2020) introduce a unified framework that covers all text-based language tasks into a text-to-text format. GPT-3 (Brown et al., 2020) is an auto-regressive language foundation model which includes 175 billion parameters, and shows strong performance on many NLP tasks under the few-shot and zero-shot settings.

Vision-Language Foundation Models Benefiting from a large number of image/video-text pairs in the Internet, the emergence of vision-language foundation models can subsume vision-language pre-training. The success of CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) indicates that the model pre-trained with simple contrastive objectives on noisy image-text pairs can generate powerful vision-language representation. Moreover, ALBEF (Li et al., 2021b), BLIP (Li et al., 2022c) and mPLUG (Li et al., 2022a) extend the task with multi-modal text completion and text generation for auxiliary learning. On the other hand, some foundation models are built through task unification. For instance, Florence (Yuan et al., 2021) unifies the contrastive objectives that can leverage both vision and vision-language data. BEiT-3 (Wang et al., 2022e) ascribe the pre-training task to mask data modeling in terms of text, vision, and vision-language. SimVLM (Wang et al., 2021b), OFA (Wang et al., 2022d), and CoCa (Yu et al., 2022) perform the generative pre-training for vision-language understanding and generation. Moreover, some methods (Li et al., 2023; Ye et al., 2023) leverage the large language model for image-language understanding and generation. Different from predominant foundation models, mPLUG-2 introduces a new modularized transformer framework, which can leverage different compositions of modules for both uni-modal and cross-modal tasks by both sharing common universal modules and disentangling modality-specific ones to address the problem of modality entanglement.

3. Method

3.1. Overall Framework

As shown in Figure 2, mPLUG-2 consists of a dual-vision encoder module for image and video, a text encoder module, a universal layers module that serves as a multi-modal

Table 2. **The modules for each downstream task.**

	Input			Modules							
Tasks	Text	Image	Video	Text Enc	Image Enc	Video Enc	Universal Layers	Fusion Layers	Text Dec	Image Dec	Video Dec
Video-Text Retrieval	✓		✓	✓		✓	✓	✓			
Video-Text Question Answering	✓		✓	✓		✓	✓	✓			✓
Video-Text Captioning	✓		✓	✓		✓	✓	✓			✓
Image-Text Retrieval	✓	✓		✓	✓		✓	✓			
Image-Text Question Answering	✓	✓		✓	✓		✓	✓		✓	
Image-Text Captioning	✓	✓		✓	✓		✓	✓		✓	
Visual Grounding	✓	✓		✓	✓		✓	✓		✓	
Video Classification			✓			✓	✓				
Image Classification		✓			✓		✓				
Image Detection		✓			✓		✓				
Image Segmentation		✓			✓		✓				
Text Classification	✓			✓			✓				
Text Question Answering	✓			✓			✓				
Text Summarization	✓			✓			✓		✓		

pivot shared by all tasks, a multi-modal fusion module and a shared decoder module for uni-modal and cross-modal generation. We first use two uni-modal encoders which encode image/video and text separately to represent the inherent information of the individual modality. For image/video, we adopt the dual-vision encoder to encode visual features with spatial modeling and local temporal modeling. Then, the visual and linguistic representations are fed into the universal module separately, which consists of multiple universal layers. Each universal layer projects different modalities to shared semantic space for cross-modal alignment while preserving the original representation of different modalities. The output of universal layers is applied to conduct uni-modal discrimination tasks. For cross-modal tasks, an additional fusion module will be applied to produce cross-modal representations. Finally, the uni-modal and cross-modal representations can be incorporated as input to a shared Transformer decoder for various generation tasks, which facilitates multi-task pre-training and transfer learning. The modules for different downstream tasks are summarized in Table 2.

Dual-vision Encoder Module To capture the visual information of various vision modalities, we propose dual-vision encoder to model image and video simultaneously. Specially, we split the image and video frames into a sequence of L non-overlapping visual tokens. Every sequence of visual tokens with learnable spatial position embeddings and an extra [CLS] token constitute an input visual sequence. However, modeling the completed visual sequences leads to difficulty in spatio-temporal learning without large-scale video pre-training (Li et al., 2022e; Wang et al., 2022a;b). To alleviate this problem, we decouple the visual representation into the spatial and temporal representation separately by introducing temporal locality. As illustrated in Figure 2(b), we leverage the self-attention (SA) layer and feed-forward layer (FFN) in the Transformer block for spatial modeling, and propose a novel local temporal modeling module (LT) to model the temporal dependency among the spatial representation as:

$$V_{LT}^n = LN(LT(V^{n-1}) + V^{n-1}), \tag{1}$$

$$V_{SA}^n = LN(SA(V_{LT}^{n-1}) + V_{LT}^{n-1}), \tag{2}$$

$$V^n = LN(FFN(V_{SA}^n) + V_{SA}^n), \tag{3}$$

where LN is short for layer normalization. The local temporal modeling module captures the correlation among patches with the same spatial locations through multi-group fusion formulated as:

$$V_g^n = ReLU(A_g^n \phi_g^n(V^{n-1})) \in \mathbb{R}^{T \times \frac{C}{G}} \tag{4}$$

$$LT(V^{n-1}) = \varphi^n(Concat[V_1^n; \cdots; V_G^n]), \tag{5}$$

where $\phi_g^n(\cdot)$ and $\varphi^n(\cdot)$ are linear transformation functions. A_g^n is the learnable temporal relation parameter, which is instantiated as a convolution kernel. T and C are number of frames and size of hidden state. G indicates the number of groups, and $Concat$ denotes concatenation function. By using multi-group fusion, the model is able to learn rich temporal information from distinctive representation subspaces at different temporal locations. As a result, except the local temporal module, the dual-vision encoder module enables weight sharing for images and videos, which effectively and efficiently learns the spatial and temporal representation.

Text Encoder Module For the text encoder module, we use BERT (Devlin et al., 2018) as the text encoder, which transforms the input text and an extra [CLS] token into a sequence of text embeddings. The embedding of [CLS] token is used to summarize the input text.

Universal Layers Module To benefit from modality collaboration, we propose the universal layers to model the vision and language modalities in the shared semantic space while preserving the original representation of the different modalities. Before the universal module, we take a variable number of image or video features V^N from the dual-vision encoders as input to produce a fixed number k of visual tokens $\mathcal{V} = \{v_1, v_2, ..., v_k\}$ to reduce the computational complexity of universal layers. In the i_{th} universal layer, the visual tokens $\mathcal{V}^{i-1}$ and the text representation $\mathcal{W}^{i-1}$ are fed to the shared self-attention layers to align semantics, and then the visual tokens are injected into the original visual feature space by the cross-attention layer to keep the original representation.

$$\mathcal{V}_{SA}^{i} = LN(SA(\mathcal{V}^{i-1}) + \mathcal{V}^{i-1}) \tag{6}$$

$$\mathcal{W}_{SA}^{i} = LN(SA(\mathcal{W}^{i-1}) + \mathcal{W}^{i-1}) \tag{7}$$

$$\mathcal{V}_{CA}^{i} = LN(CA(\mathcal{V}_{SA}^{i}, V^{n}) + \mathcal{V}_{SA}^{i}) \tag{8}$$

$$\mathcal{V}^{i} = LN(FFN(\mathcal{V}_{CA}^{i}) + \mathcal{V}_{CA}^{i}) \tag{9}$$

$$\mathcal{W}^{i} = LN(FFN(\mathcal{W}_{SA}^{i}) + \mathcal{W}_{SA}^{i}) \tag{10}$$

Then $[\mathcal{V}^{i}; \mathcal{W}^{i}]$ is fed into the next universal layer repeatedly to get the final common image and text representation. Finally, the output of the universal layers $[\mathcal{V}^{S}; \mathcal{W}^{S}]$ are combined with the original representations $[V^{N}; W^{M}]$ by the cross-attention layer for the text-aware visual and visual-aware text representation, where S, N, M are the layers of universal module, dual-vision encoder and text encoder respectively.

Fusion Module To effectively capture the cross-modal interaction between vision and language modalities, we use the fusion module as in ALBEF (Li et al., 2021b), which is composed of a stack of Transformer blocks with cross-attention layers. Specifically, the fusion module takes the text embeddings from the universal layers module as the input. Then, the text-aware vision embedding cross-attends to the visual-aware text embeddings in language-shared common space. By cascading the Transformer blocks with cross-attention layers, fusion module is able to yield multi-modal vision-language representations.

Shared Decoder Module To empower the model with the capability of generation, a shared decoder module is introduced to enable the model to generate text with both uni-modal and multi-modal information. In detail, the shared decoder module is a Transformer decoder with arbitrary inputs. For example, image captioning only requires the visual features, while the multi-modal features are used for visual question answering. By taking different types of input, our shared decoder module can adapt to a variety of tasks with text generation. The shared decoder module facilitates multi-task pre-training and transfer learning.

3.2. Unified Pre-training Objectives

We jointly train the multiple modules of mPLUG-2 with the following three objectives.

Language Loss For the text encoder module, we use Masked Language Modeling (MLM) as in BERT (Devlin et al., 2018) to learn the text representation. We randomly mask 15% tokens in the text and the model is asked to predict these masked tokens with the context representations.

Multi-modal Loss For the cross-modal module, we employ the Cross-modal Matching Losses (CML) as in ALBEF (Li et al., 2021b), which consists of Vision-language Matching (VLM) and Vision-language Contrastive Learning (VLC).

Instruction-based Language Model Loss Following Flamingo (Alayrac et al., 2022) and OFA (Wang et al., 2022d), we adopt the Instruction-based Language Model Loss to unify various generation tasks. We use handcrafted instructions to discriminate tasks and modalities, which include Video/Image-Text Pairs, Video/Image Captioning, Video/Image Question Answering, Text Generation, etc.

4. Experiment

4.1. Training Setup

Pre-training Datasets Following previous works (Li et al., 2021b; 2022a), we pre-train our model with the same popular image-text datasets with 14M images including MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), Conceptual Captions 3M (Sharma et al., 2018), Conceptual Captions 12M (Changpinyo et al., 2021), and SBU Captions (Ordonez et al., 2011). For video-text datasets, we adopt the web-sourced video dataset WebVid-2M (Bain et al., 2021a) with 2.5M video-text pairs. The text datasets consists of WikiCorpus (Devlin et al., 2018) (about 20GB) and cleaned common crawl (about 350GB). The collection and cleaning method of the latter is generally the same as that used in c4 (Raffel et al., 2020). The implementation details of pre-training can be found in the Appendix.

4.2. Main Results

We evaluate the new unified paradigm of mPLUG-2 on over 30 benchmarks including vision-language tasks (e.g. multi-modal retrieval, question answering and captioning) (Xu et al., 2016; 2017; Chen & Dolan, 2011), language-only tasks (e.g. text classification, question answering and summarization) (Wang et al., 2018; Rush et al., 2015a), and vision-only tasks (e.g. image classification and video action recognition) (Deng et al., 2009; Kay et al., 2017). Specially, the vision-language benchmarks can be categorized as image-text parts and video-text parts. Details of these datasets can be found in the Appendix.

4.2.1. Multi-modal Tasks

Text-to-video Retrieval We compare mPLUG-2 with several state-of-the-art methods on MSRVTT (Xu et al., 2016), DiDeMo (Anne Hendricks et al., 2017) and LSMDC (Rohrbach et al., 2015) datasets. The results are summarized in Table 3. We can observe that mPLUG-2 outperforms the previous SoTA methods on most of the datasets. In particular, our method yields 5.7% lift in terms of R@1 on LSMDC datasets compared with HiTeA, which indicates that the proposed model can leverage the temporal information presented in fruitful movie clips through the proposed local temporal modeling module in the dual-vision encoder.

Table 3. **Performance comparison on text-to-video retrieval.** All results are reported on R@1/R@5/R@10.

		MSRVTT			DiDeMo			LSMDC		
Method	#PT Data	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Frozen (Bain et al., 2021b)	5M	31.0	59.5	70.5	31.0	59.8	72.4	15.0	30.8	39.8
BridgeFormer (Ge et al., 2022)	5M	37.6	64.8	75.1	37.0	62.2	73.9	17.9	35.4	44.5
Singularity (Lei et al., 2022)	5M	36.8	65.9	75.5	47.4	75.2	84.0	-	-	-
LAVENDER (Li et al., 2022e)	30M	37.8	63.8	75.0	47.4	74.7	82.4	22.2	43.8	53.5
All-in-one (Wang et al., 2022a)	283M	37.9	68.1	77.1	32.7	61.4	73.5	-	-	-
OmniVL (Wang et al., 2022b)	18M	47.8	74.2	83.8	52.4	79.5	85.4	-	-	-
HiTeA (Ye et al., 2022)	17M	46.8	71.2	81.9	**56.5**	**81.7**	**89.7**	28.7	50.3	59.0
mPLUG-2$_{\text{Base}}$	17M	48.3	75.0	83.2	52.3	80.8	87.5	25.5	45.8	55.8
mPLUG-2	17M	**53.1**	**77.6**	**84.7**	56.4	79.1	85.2	**34.4**	**55.2**	**65.1**

Table 4. **Performance comparison on video question answering.** Accuracy is reported for evaluation. mPLUG-2 creates a new state-of-the-art video question answering results on MSRVTT-QA and TGIF-FrameQA with open-vocabulary generation.

Method	#PT Data	MSRVTT-QA	MSVD-QA	TGIF-FrameQA
JustAsk (Yang et al., 2021a)	69M	41.5	46.3	-
LAVENDER (Li et al., 2022e)	30M	45.0	56.6	73.5
All-in-one (Wang et al., 2022a)	283M	46.8	48.3	66.3
MERLOT (Zellers et al., 2021)	180M	43.1	-	69.5
OmniVL (Wang et al., 2022b)	18M	44.1	51.0	-
HiTeA (Ye et al., 2022)	17M	45.9	55.3	73.2
GIT (Wang et al., 2022c)	800M	43.2	56.8	72.8
GIT2 (Wang et al., 2022c)	12.9B	45.6	**58.2**	74.9
FrozenBiLM (Yang et al., 2022)	10M	47.0	54.8	68.6
VideoCoCa (Yan et al., 2022)	3B	46.0	56.9	-
InternVideo (Wang et al., 2022g)	12M	47.1	55.5	72.2
Flamingo (Alayrac et al., 2022)	2.3B	47.4	-	-
mPLUG-2$_{\text{Base}}$	17M	46.3	55.3	72.6
mPLUG-2	17M	**48.0**	58.1	**75.4**

Video Question Answering Table 4 summarizes the video question answering results on MSRVTT-QA (Xu et al., 2017), MSVD-QA (Xu et al., 2017), and TGIF-FrameQA (Jang et al., 2017). It can be observed that mPLUG-2 outperforms all the existing foundation models on MSRVTT-QA and TGIF-FrameQA by a large margin, and it also attains the comparable result with big foundation models GIT2 (Wang et al., 2022c) on MSVD-QA even using significantly smaller amount of pre-trained data. In particular, mPLUG-2 achieves absolute improvement 0.6% on MSRVTT and 0.5% on TGIF-FrameQA. Furthermore, mPLUG-2$_{\text{Base}}$ achieves the comparable results compared to the large models (i.e., VideoCoCa and GIT2) with smaller model size.

Video Captioning Table 55 compares mPLUG-2 with existing methods on video captioning datasets MSRVTT and MSVD. As shown in the table, although pre-trained on less data, mPLUG-2 derives the significant improvement on MSRVTT dataset and comparable performance on MSVD dataset. On MSRVTT Caption, our method surpasses SoTA method VideoCoCa (Yan et al., 2022) and GIT2 (Wang et al., 2022c) by 4.4% on CIDEr and 3.0% on BLEU@4. Moreover, we can notice mPLUG-2 outperforms HiTeA with the same amount of pre-training data, which shows that mPLUG-2 is able to generate stronger video-language representation.

Visual Grounding We compare mPLUG-2 with existing state-of-the-art methods on visual grounding datasets including RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016). Table 7 shows that mPLUG-2 achieves comparable performance to the state-of-the-art methods. Our method achieve 0.97% absolute improvement compared with the second best method on RefCOCO "testB" split without using object detection data for pre-training. Queries in "testB" split can refer to various visual concepts but only people in "testA". The improvement demonstrates that the introduction of universal layers can help model the visual concepts in the image.

Image-Text Retrieval We evaluate mPLUG-2 on image-text retrieval datasets MSCOCO and Flickr30k. As shown in Table 6, both mPLUG-2$_{\text{Base}}$ and mPLUG-2 achieves comparable or better performance than state-of-the-art methods. Florence (Yuan et al., 2021) and BLIP (Li et al., 2022c) use 0.9B and 129M data for pre-train respectively. In contrast, our mPLUG-2 only requires 17M data. It demonstrates that mPLUG-2 is data-efficient.

Visual Question Answering We report the performance of mPLUG-2 on visual question answering test sets. mPLUG-2 surpasses state-of-the-art method Florence (Yuan et al., 2021) 0.95% on test-dev and 0.77% on test-std. The scale of the pre-trained data used in our model is 89.11% less than that in Florence. It shows that our mPLUG-2 can learn multi-modal representations efficiently and effectively.

Image Captioning We compare mPLUG-2 with existing state-of-the-art methods on MSCOCO (Lin et al., 2014). Following (Li et al., 2020b), we train mPLUG-2 on the COCO Caption with cross-entropy loss and test on the same Karpathy split. As shown in Table 9, our mPLUG-2 achieves new SoTA results on COCO Caption. Moreover, our method achieves competitive results with big foundation models, such as LEMON (Hu et al., 2021) and BLIP (Li et al., 2022c) which use more than nearly 10x amount of pre-training data. Specifically, our mPLUG-2 outperforms BLIP on COCO caption by an obvious 1.2 point margin on BLEU@4, and 1 point on CIDEr.

Table 5. **Performance comparison on video captioning.** B@4: BLEU@4; M: METEOR; R: ROUGE-L; C: CIDEr.

		MSRVTT				MSVD			
Method	#PT Data	B@4	M	R	C	B@4	M	R	C
UniVL (Luo et al., 2020)	136M	42.2	28.2	61.2	49.9	-	-	-	-
SwinBERT (Lin et al., 2022)	-	41.9	29.9	62.1	53.8	58.2	41.3	77.5	120.6
CLIP4Caption (Tang et al., 2021)	-	46.1	30.7	63.7	57.7	-	-	-	-
MV-GPT (Seo et al., 2022)	69M	48.9	38.7	64.0	60.0	-	-	-	-
LAVENDER (Li et al., 2022e)	30M	-	-	-	60.1	-	-	-	150.7
HiTeA (Ye et al., 2022)	17M	49.2	30.7	65.0	65.1	71.0	45.3	81.4	146.9
VideoCoca (Yan et al., 2022)	3B	53.8	-	68.0	73.2	-	-	-	-
GIT (Wang et al., 2022c)	0.8B	53.8	32.9	67.7	73.9	79.5	51.1	87.3	180.2
GIT2 (Wang et al., 2022c)	12.9B	54.8	33.1	68.2	75.9	**82.2**	**52.3**	**88.7**	**185.4**
mPLUG-2$_{\text{Base}}$	17M	52.2	32.1	66.9	72.4	69.3	45.1	81.9	148.2
mPLUG-2	17M	**57.8**	**34.9**	**70.1**	**80.3**	75.0	48.4	85.3	165.8

Table 6. **Performance comparison on image-text retrieval.** All results are reported on R@1/R@5/R@10.

		MSCOCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
Method	#PT Data	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
E2E-VLP (Xu et al.)	4M	-	-	-	-	-	-	86.2	97.5	98.92	73.6	92.4	96.0
UNITER (Chen et al., 2020)	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
OSCAR (Li et al., 2020b)	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO (Li et al., 2020a)	4M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
VLMo (Wang et al., 2021a)	4M	78.2	94.4	97.4	60.6	84.4	91.0	95.3	99.9	**100.0**	84.5	97.3	98.6
ALIGN (Jia et al., 2021)	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	**100.0**	84.9	97.4	98.6
ALBEF (Li et al., 2021b)	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	**100.0**	85.6	97.5	98.9
Florence (Yuan et al., 2021)	0.9B	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	**98.1**	-
BLIP (Li et al., 2022c)	129M	82.4	95.4	97.9	65.1	86.3	91.8	**97.4**	99.8	99.9	87.6	97.7	99.0
mPLUG-2$_{\text{Base}}$	17M	81.2	95.6	**98.1**	65.3	86.9	92.4	96.9	**100.0**	**100.0**	**88.2**	97.8	99.0
mPLUG-2	17M	**82.5**	**95.7**	98.0	**65.7**	**87.1**	**92.6**	97.2	**100.0**	**100.0**	88.1	97.6	**99.1**

Table 7. **Evaluation results on visual grounding (ReferCOCO and ReferCOCOg).** We use the accuracy@0.5 (a prediction is right if the IoU between the grounding-truth box and the predicted bounding box is larger than 0.5) to measure model performance.

Model	RefCOCO			RefCOCOg	
	val	testA	testB	val-u	test-u
UNITER (Chen et al., 2020)	81.41	87.04	74.17	74.86	75.77
VILLA (Gan et al., 2020)	82.39	87.48	74.84	76.18	76.71
MDETR (Kamath et al., 2021)	86.75	89.58	81.41	81.64	80.89
UNICORN (Yang et al., 2021b)	88.29	90.42	83.06	83.44	83.93
OFA_{Large} (Wang et al., 2022d)	90.05	**92.93**	85.26	84.54	**85.20**
mPLUG-2	**90.33**	92.80	**86.05**	**84.70**	85.14

4.2.2. Language Only Tasks

Natural Language Understanding We evaluate mPLUG-2 on 6 tasks of the GLUE benchmark (Wang et al., 2018) for natural language understanding. Table 10 shows that mPLUG-2 achieves comparable performance to the state-of-the-art natural language and multimodal pretrained models including RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021b). Our method with DeBERTa achieves improvement compared with DeBERTa (He et al., 2021b) on three tasks, which also demonstrates the effectiveness of universal modules for modality collaboration.

Natural Language Generation We evaluate mPLUG-2 on Gigaword abstractive summarization (Rush et al., 2015b) for natural language generation. As shown in Table 11, mPLUG-2 achieves the comparable result with the state-of-the-art models.

Table 8. **Performance comparison on visual question answering.** Accuracy is reported for evaluation.

Method	#PT Data	test-dev	test-std
UNITER (Chen et al., 2020)	4M	72.70	72.91
UNIMO (Li et al., 2020a)	4M	73.79	74.02
E2E-VLP (Xu et al.)	4M	73.25	73.67
OSCAR (Li et al., 2020b)	4M	73.16	73.44
ALBEF (Li et al., 2021b)	4M	74.54	74.70
BLIP (Li et al., 2022c)	14M	77.54	77.62
SimVLM (Wang et al., 2021b)	1.8B	80.03	80.34
Florence (Yuan et al., 2021)	0.9B	80.16	80.36
OFA_{Large} (Wang et al., 2022d)	18M	80.30	80.50
VLMo (Wang et al., 2021a)	-	79.94	79.98
GIT (Wang et al., 2022c)	0.8B	78.56	78.81
mPLUG-2$_{\text{Base}}$	17M	79.27	79.32
mPLUG-2	17M	**81.11**	**81.13**

4.2.3. Vision Only Tasks

Video Action Recognition Video action recognition is the most representative of video understanding since it requires the model to understand the spatio-temporal cues revealed in the video. Table 12 summarizes the performance of different approaches on Kinetics 400, Kinetics 600, and Kinetics 700 datasets. Our mPLUG-2 surpasses the most of SoTA methods. For example, compared with Florence pre-trained on 900M vision-text pairs, mPLUG-2 improves the Top-1 accuracy by 1.9% on Kinetics 600 and 0.6% on Kinetics 400. Meanwhile, we can notice that the performance of mPLUG-2 is better than OmniVL with similar amount of pre-training data, which shows the effectiveness of the dual-vision encoder module for video representation learning.

Table 9. **Performance comparison on image captioning.** B@4: BLEU@4; M: METEOR; R: ROUGE-L; C: CIDEr.

Method	#PT Data	COCO Caption			
		B@4	M	C	S
Encoder-Decoder	12M	-	-	110.9	-
E2E-VLP (Xu et al.)	4M	36.2	-	117.3	-
VinVL (Zhang et al., 2021b)	5.65M	38.5	30.4	130.8	
OSCAR (Li et al., 2020b)	6.5M	-	-	-	-
LEMON$_{large}$ (Hu et al., 2021)	200M	40.6	30.4	135.7	23.5
BLIP (Li et al., 2022c)	129M	40.4	-	136.7	-
mPLUG-2	17M	**41.6**	**30.9**	**137.7**	**23.7**

Table 10. **Experimental results on the GLUE benchmark.**

Model	SST-2	RTE	MRPC	QQP	MNLI	QNLI
Multimodal Pretrained Baseline Models						
VisualBERT (Li et al., 2019)	89.4	56.6	71.9	89.4	81.6	87.0
UNITER (Chen et al., 2020)	89.7	55.6	69.3	89.2	80.9	86.0
VL-BERT (Su et al., 2019)	89.8	55.7	70.6	89.0	81.2	86.3
VilBERT (Lu et al., 2019)	90.4	53.7	69.0	88.6	79.9	83.8
LXMERT (Tan & Bansal, 2019)	90.2	57.2	69.8	75.3	80.4	84.2
Uni-Perceiver (Zhu et al., 2021)	90.2	64.3	86.6	87.1	81.7	89.9
SimVLM (Wang et al., 2021b)	90.9	63.9	75.2	90.4	83.4	88.6
FLAVA (Singh et al., 2021)	90.9	57.8	81.4	90.4	80.3	87.3
UNIMO (Li et al., 2020a)	96.8	-	-	-	89.8	-
OFA (Wang et al., 2022d)	96.6	**91.0**	91.7	92.5	90.2	94.8
Natural-Language-Pretrained SOTA Models						
BERT (Devlin et al., 2018)	93.2	70.4	88.0	91.3	86.6	92.3
RoBERTa (Liu et al., 2019)	96.4	86.6	90.9	92.2	90.2	93.9
XLNet (Yang et al., 2019)	**97.0**	85.9	90.8	92.3	90.8	94.9
ELECTRA (Clark et al., 2020)	96.9	88.0	90.8	92.4	90.9	95.0
DeBERTa (He et al., 2021b)	96.8	88.3	91.9	92.3	**91.1**	**95.3**
mPLUG-2$_{\text{Base}}$	93.5	85.2	87.3	91.3	87.6	93.2
mPLUG-2	95.1	88.0	90.1	**92.7**	90.2	94.5
mPLUG-2$_{\text{Deberta}}$	96.2	89.4	**92.1**	92.6	90.8	94.8

Image Classification We further evaluate the performance of mPLUG-2 in terms of image classification on ImageNet-1K. As we can see in Table 13, We can see that mPLUG-2 achieves comparable results or even surpass the SoTA methods on ImageNet-1K without using the ImageNet data for pre-training. Besides, to effectively evaluate the robustness and generalization ability of mPLUG-2 , we perform the evaluation on 5 ImageNet variants (i.e. IN-V2, IN-Real., IN-Adversarial, IN-Rendition, and IN-Sketch). Following standard evaluation procedure (Fang et al., 2022a), all these models are first fine-tuned on the original ImageNet-1K training set and directly tested on the 6 variants without further fine-tuning. As shown in Table 13, mPLUG-2 not only achieves the highest accuracy on ImageNet-1K validation set but also obtains the relative small gap (i.e., $\Delta_{\downarrow}$), which reflects the excellent robustness and generalization capability of mPLUG-2 with the help of the universal layer module by learning language-shared representation.

Table 11. **Experimental results on Gigaword abstractive summarization.** We report performance on the ROUGE evaluation.

Model	Gigaword		
	ROUGE-1	ROUGE-2	ROUGE-L
BERTSHARE (Rothe et al., 2020)	38.13	19.81	35.62
MASS (Song et al., 2019)	38.73	19.71	35.96
UniLM (Dong et al., 2019)	38.45	19.45	35.75
PEGASUS (Zhang et al., 2020)	39.12	19.86	36.24
ProphetNet (Qi et al., 2020)	39.55	20.27	36.57
UNIMO (Li et al., 2020a)	39.71	20.37	36.88
OFA (Wang et al., 2022d)	**39.81**	20.66	**37.11**
mPLUG-2	39.65	**20.67**	36.89

Table 12. **Comparison with the state-of-the-art on video action recognition under fine-tuning settings.**

Method	Kinetics 400		Kinetics 600		Kinetics 700	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
TimeSformer-L (Bertasius et al., 2021)	80.6	94.7	82.2	95.6	-	-
ViViT-H (Arnab et al., 2021)	84.8	95.8	85.8	96.5	-	-
VideoSwin-L (Liu et al., 2022b)	84.9	96.7	86.1	97.3	-	-
OmniVL (Wang et al., 2022b)	79.1	94.5	-	-	-	-
TokenLearner (Ryoo et al., 2021)	85.4	96.3	86.3	97.0	-	-
VATT (Akbari et al., 2021)	82.1	95.5	83.6	96.6	-	-
MoViNet (Kondratyuk et al., 2021)	81.5	-	84.8	-	79.4	-
Florence (Yuan et al., 2021)	86.5	97.3	87.8	97.8	-	-
CoVeR (Zhang et al., 2021a)	86.3	97.2	87.9	97.8	78.5	94.2
mPLUG-2$_{\text{Base}}$	83.6	96.0	86.7	97.2	74.6	91.2
mPLUG-2	**87.1**	**97.7**	**89.8**	**98.3**	**80.4**	**94.9**

Table 13. **Comparison with state-of-the-art methods in terms of robustness and generalization capability evaluation on ImageNet-1K variants.** We test the model on various ImageNet-1K validation set without any further fine-tuning. "Avg." indicates the average Top-1 accuracy on 6 different ImageNet-1K variants. "$\Delta_{\downarrow}$" stands for the gap between averaged Top-1 accuracy of 6 variants and the accuracy of original ImageNet-1K validation (the lower the better).

Method	IN-1K	IN-V2	IN-ReaL	IN-Adv.	IN-Ren.	IN-Ske.	Avg.	$\Delta_{\downarrow}$
ConvNeXt (Liu et al., 2022a)	87.5	77.7	90.5	70.8	67.0	53.7	74.5	13.0
SwinV2-G (Liu et al., 2021a)	87.5	77.3	90.2	73.9	67.7	52.3	74.8	12.7
MAE (He et al., 2021a)	87.8	79.2	90.3	76.7	66.5	50.9	75.2	12.6
DeiT3 (Touvron et al., 2022)	87.7	79.1	90.2	79.2	70.6	54.9	77.0	10.7
Eff-L2-NS (Tan & Le, 2019)	88.4	**80.5**	**90.6**	**84.8**	74.7	47.6	**77.8**	**10.6**
OFA (Wang et al., 2022d)	85.6	-	-	-	-	-	-	-
mPLUG-2	**88.5**	78.1	89.5	73.2	**75.6**	**61.2**	77.7	10.7

Table 14. **Evaluation of the proposed instructional-based learning on downstream tasks.** For retrieval task, we report the average of Recall@1, Recall@5, and Recall@10. For QA and caption task, Top-1 Accuarcy and CIDEr are reported.

Instruction	MSRVTT-Ret.	MSVD-QA	MSRVTT-Cap.
	72.8	54.1	71.8
✓	**73.5** (+0.7)	**55.3** (+1.2)	**72.4** (+0.6)

Table 15. **Evaluation of different temporal modeling modules in the dual-vision encoder module.** For retrieval task, we report the average of Recall@1, Recall@5, and Recall@10. For QA and caption task, Top-1 Accuarcy and CIDEr are reported.

Temporal Module	MSRVTT-Ret.	MSVD-QA	MSRVTT-Cap.
Temporal Self-Attention	70.3	55.1	71.1
Temporal Convolution	71.4 (+1.1)	55.0 (-0.1)	71.7 (+0.6)
Local Temporal Modeling	**73.5** (+3.2)	**55.3** (+0.2)	**72.4** (+1.3)

4.3. Discussion

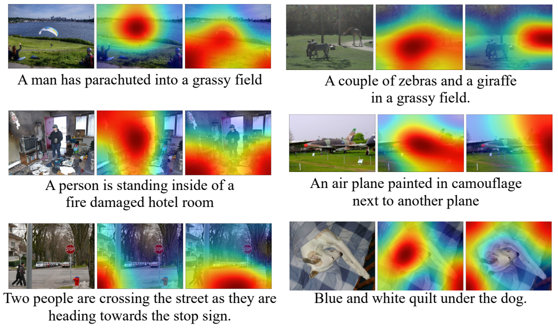
Impact of Instruction-based Learning The instructional-based learning is able to distinguish different types of tasks with specific instructions. Table 14 demonstrates the effectiveness of instructional-based learning. In the table, we can observe that instructional-based learning improves the performance of retrieval and question answering by at least 0.7% and 1.2% in Average Recall and accuracy respectively. With the help of instructional-based learning, mPLUG-2 is capable of utilizing the different modules when different instructions are used to boost the performance.

Table 16. **Evaluation of the impact of universal layer** in terms of boosting vision task's performance.

Method	ImageNet	CIFAR10	CIFAR100	Cars	DTD	SUN	Food101	Average
CLIP-ViT-L/14	86.2	98.6	92.2	91.6	81.9	80.7	94.4	89.4
+Universal Layers	86.6 (+0.4)	99.3 (+0.7)	93.1 (+0.9)	94.4 (+2.8)	85.1 (+3.2)	80.4 (-0.4)	95.4 (+1.0)	**90.6** (+1.2)

Table 17. **Evaluation of the impact of the universal layer** in terms of boosting language and vision-language task's performance.

Model	SST-2	RTE	MRPC	QQP	MNLI	QNLI	VQA test-dev
BERT$_{base}$	91.7	71.4	86.3	90.8	84.3	89.3	78.6
+Joint Training	92.5	82.3	86.6	90.6	86.2	92.1	78.9
+Universal Layers	**93.5**	**85.2**	**87.3**	**91.3**	**87.6**	**93.2**	**79.3**


A man has parachuted into a grassy field

A couple of zebras and a giraffe in a grassy field.

A person is standing inside of a fire damaged hotel room

An air plane painted in camouflage next to another plane

Two people are crossing the street as they are heading towards the stop sign.

Blue and white quilt under the dog.

Figure 3. Grad-CAM visualizations for latent queries in the universal layers.

Impact of Local Temporal Modeling Module To validate the effectiveness of our proposed local temporal modeling module in the dual-vision encoder, we conduct experiments with the different temporal modeling structures. Specially, we have tried out the temporal self-attention and temporal convolution for comparison. The results are summarized in Table 15. We can notice that the local temporal modeling module outperforms temporal self-attention module by introducing modeling temporal locality. Meanwhile, with the help of the multi-group fusion mechanism, the local temporal modeling module can learn the diverse temporal representations in distinctive representation subspaces while the temporal convolution is restricted in the same temporal representation spaces, thus leading to the better performance.

Impact of Universal Layer To validate the effectiveness of our proposed universal layer module, we ablation this module for all uni-modal and multi-modal tasks. As shown in Table 16 and Table 17, we set Row 1/2/2 as the baseline of the vision/language/vision-language task in this experiment, respectively. We can find that compared with the baseline the shared universal layer is beneficial for all modality tasks by encouraging collaboration between modalities.

In Figure 5, we visualize the Grad-CAM on the cross-attention map in the first universal layer. For each sample, we present two cross-attention maps that attend to different visual concepts. The results show that the universal layer can encourage modality collaboration and modality entanglement between visual patch features and language features by attending the areas of various visual concepts in the image.

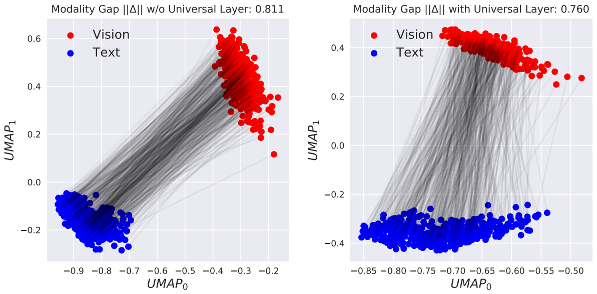


Figure 4. The UMAP visualization of generated vision and language embeddings from pre-trained mPLUG-2 . The black lines refer to vision-language pairs.

Universal Layer for Modality Collaboration Here we investigate the influence of universal layer in terms of modality collaboration. We randomly sample some vision-language pairs, and sketch the UMAP visualization of the generated embeddings from pre-trained mPLUG-2 in the Figure 4. We can observe that with the help of universal layer, the distance between vision and text samples are more closer instead of solely two concentrated clusters. Besides, we quantitatively compute the modality gap $\|\Delta\|$ (Liang et al., 2022), where the Δ is the difference between the center of vision embeddings and text embeddings. It can be observed that the model with universal layer would encourage the collaboration between vision and language modalities thus yielding lower modality gap compared to the model without universal layer.

5. Conclusion

This paper presents mPLUG-2 , a new unified paradigm with modularized design for building multi-modal foundation models. mPLUG-2 introduces a module-based network design that shares common universal modules for modality collaboration and disentangles modality-specific modules to address the problem of modality entanglement. Experimental results show that the new unified paradigm of mPLUG-2 can achieve strong performances on a broad range of over 30 tasks across the text, image and video modalities. It is also easy to extend mPLUG-2 to more tasks by selecting and adding modules.

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. Vqa: Visual question answering. *Int. J. Comput. Vision*, 123(1):4–31, May 2017.

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., and Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., and Russell, B. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., and Schmid, C. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 6816–6826. IEEE, 2021.

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1708–1718, 2021a.

Bain, M., Nagrani, A., Varol, G., and 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*, pp. 1728–1738, 2021b.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *ICML*, volume 2, pp. 4, 2021.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J. (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346, pp. 213–229, 2020.

Changpinyo, S., Sharma, P., Ding, N., and 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*, pp. 3558–3568, 2021.

Chen, D. and Dolan, W. B. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

Chen, T., Saxena, S., Li, L., Fleet, D. J., and Hinton, G. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Clark, K., Luong, M., Le, Q. V., and Manning, C. D. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H. Unified language model pre-training for natural language understanding and generation. In *NeurIPS 2019*, pp. 13042–13054, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022a.

Fang, Y., Wang, W., Xie, B., Sun, Q.-S., Wu, L. Y., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. *ArXiv*, abs/2211.07636, 2022b.

Fu, T.-J., Li, L., Gan, Z., Lin, K., Wang, W. Y., Wang, L., and Liu, Z. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.

Gan, Z., Chen, Y., Li, L., Zhu, C., Cheng, Y., and Liu, J. Large-scale adversarial training for vision-and-language representation learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Ge, Y., Ge, Y., Liu, X., Li, D., Shan, Y., Qie, X., and Luo, P. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16167–16176, 2022.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T., Cubuk, E. D., Le, Q. V., and Zoph, B. Simple copy-paste is a strong data augmentation method for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 2918–2928. Computer Vision Foundation / IEEE, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2021a.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021b.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L. Scaling up vision-language pretraining for image captioning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17959–17968, 2021.

Huang, J., Li, Y., Feng, J., Sun, X., and Ji, R. Clover: Towards a unified video-language alignment and fusion model. *arXiv preprint arXiv:2207.07885*, 2022a.

Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pp. 9226–9259. PMLR, 2022b.

Jang, Y., Song, Y., Yu, Y., Kim, Y., and Kim, G. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, A., Suleyman, M., and Zisserman, A. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017.

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., and Gong, B. Movinets: Mobile video networks for efficient video recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 16020–16030. Computer Vision Foundation / IEEE, 2021.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Lei, J., Berg, T. L., and Bansal, M. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7871–7880. Association for Computational Linguistics, 2020.

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., da Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., and Si, L. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *ArXiv*, abs/2205.12005, 2022a.

Li, D., Li, J., Li, H., Niebles, J. C., and Hoi, S. C. Align and prompt: Video-and-language pre-training with entity prompts. *arXiv preprint arXiv:2112.09583*, 2021a.

Li, D., Li, J., Li, H., Niebles, J. C., and Hoi, S. C. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4953–4963, 2022b.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021b.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022c.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., and Qiao, Y. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *ArXiv*, abs/2211.09552, 2022d.

Li, L., Gan, Z., Lin, K., Lin, C.-C., Liu, Z., Liu, C., and Wang, L. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022e.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W., and Gao, J. Grounded language-image pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965, 2021c.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020a.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020b.

Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., and Ling, H. Cbnet: A composite backbone network architecture for object detection. *IEEE Transactions on Image Processing*, 31:6893–6906, 2021.

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022.

Lin, K., Li, L., Lin, C.-C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., and Wang, L. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17949–17958, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. Swin transformer v2: Scaling up capacity and resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11999–12009, 2021a.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021b.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022a.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 3192–3201. IEEE, 2022b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2019.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., and Zhou, M. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

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

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., and Ji, R. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*, 2022.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*, 2019.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and 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*, pp. 2641–2649, 2015.

Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 2401–2410, 2020.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3202–3212, 2015.

Rothe, S., Narayan, S., and Severyn, A. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280, 2020.

Rush, A. M., Chopra, S., and Weston, J. A neural attention model for abstractive sentence summarization. In *Conference on Empirical Methods in Natural Language Processing*, 2015a.

Rush, A. M., Chopra, S., and Weston, J. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, 2015b.

Ryoo, M. S., Piergiovanni, A. J., Arnab, A., Dehghani, M., and Angelova, A. Tokenlearner: What can 8 learned tokens do for images and videos? *ArXiv*, abs/2106.11297, 2021.

Seo, P. H., Nagrani, A., Arnab, A., and Schmid, C. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17959–17968, 2022.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., and Sun, J. Objects365: A large-scale, high-quality dataset for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8429–8438, 2019.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. 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)*, pp. 2556–2565, 2018.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15617–15629, 2021.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. MASS: masked sequence to sequence pre-training for language generation. In *ICML 2019*, pp. 5926–5936, 2019.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

Tang, M., Wang, Z., Liu, Z., Rao, F., Li, D., and Li, X. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4858–4862, 2021.

Touvron, H., Cord, M., and J'egou, H. Deit iii: Revenge of the vit. In *European Conference on Computer Vision*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*, 2018.

Wang, A. J., Ge, Y., Yan, R., Ge, Y., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X., and Shou, M. Z. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022a.

Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.-G., and Yuan, L. Omnivl: One foundation model for image-language and video-language tasks. *ArXiv*, abs/2209.07526, 2022b.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. Git: A generative image-to-text transformer for vision and language. *ArXiv*, abs/2205.14100, 2022c.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 2022d.

Wang, W., Bao, H., Dong, L., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021a.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O., Singhal, S., Som, S., and Wei, F. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv*, abs/2208.10442, 2022e.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., hua Hu, X., Lu, T., Lu, L., Li, H., Wang, X., and Qiao, Y. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *ArXiv*, abs/2211.05778, 2022f.

Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., and Qiao, Y. Internvideo: General video foundation models via generative and discriminative learning. *ArXiv*, abs/2212.03191, 2022g.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021b.

Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., and Huang, F. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pp. 503–513.

Xu, J., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Xu, J., Mello, S. D., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. Groupvit: Semantic segmentation emerges from text supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18113–18123, 2022.

Yan, S., Zhu, T., Wang, Z., Cao, Y., Zhang, M., Ghosh, S., Wu, Y., and Yu, J. Video-text modeling with zero-shot transfer from contrastive captioners. *ArXiv*, abs/2212.04979, 2022.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1686–1697, 2021a.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5754–5764, 2019.

Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., and Wang, L. Crossing the format boundary of text and boxes: Towards unified vision-language modeling. *ArXiv*, abs/2111.12085, 2021b.

Ye, Q., Xu, G., Yan, M., Xu, H., Qian, Q., Zhang, J. C., and Huang, F. Hitea: Hierarchical temporal-aware video-language pre-training. *ArXiv*, abs/2212.14546, 2022.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J. C., and Huang, F. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178, 2023.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *ArXiv*, abs/2205.01917, 2022.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85. Springer, 2016.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J. S., Cao, J., Farhadi, A., and Choi, Y. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1204–1213, 2021.

Zhang, B., Yu, J., Fifty, C., Han, W., Dai, A. M., Pang, R., and Sha, F. Co-training transformer with videos and images improves action recognition. *ArXiv*, abs/2112.07175, 2021a.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. Pegasus: Pretraining with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pp. 11328–11339. PMLR, 2020.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529, 2021b.

Zhu, X., Zhu, J., Li, H., Wu, X., Wang, X., Li, H., Wang, X., and Dai, J. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. *arXiv preprint arXiv:2112.01522*, 2021.

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27, 2015.

Table 18. **System-level comparisons with the state-of-the-art results on COCO dataset for object detection and instance segmentation.** We report the standard boudning box AP (AP$_{box}$) and mask AP (AP$_{mask}$). The detector are Cascade Mask R-CNN (Cascade), Dynamic Head (DyHead), Hybrid Task Cascade (HTC), and its extension (HTC++).

Method	Detector	AP$_{box}$	AP$_{mask}$
Mask R-CNN (He et al., 2017)	-	46.3	40.1
DETR (Carion et al., 2020)	-	44.9	33.0
Pix2seq (Chen et al., 2021)	-	45.0	-
Copy-Paste (Ghiasi et al., 2021)	Cascade	57.0	48.9
Swin-L (Liu et al., 2021b)	HTC++	58.0	50.4
CBNetV2 (Liang et al., 2021)	HTC	59.6	51.8
GLIP (Li et al., 2021c)	DyHead	60.8	-
SwinV2-L (Liu et al., 2021a)	HTC++	60.2	**52.1**
Florence (Yuan et al., 2021)	DyHead	**62.0**	-
mPLUG-2	Cascade	46.9	40.6

Table 19. **Zero-shot evaluation on text-to-video retrieval.** All results are reported on R@1/R@5/R@10.

		MSRVTT			DiDeMo			LSMDC		
Method	# PT Data	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Frozen (Bain et al., 2021b)	5M	18.7	39.5	51.6	21.1	46.0	56.2	9.3	22.0	30.1
ALPRO (Li et al., 2021a)	5M	24.1	44.7	55.4	23.8	47.3	57.9	-	-	-
Singularity (Lei et al., 2022)	5M	28.4	50.2	59.5	36.9	61.6	69.3	-	-	-
VIOLET (Fu et al., 2021)	183M	25.9	49.5	59.7	23.5	49.8	59.8	-	-	-
Florence (Yuan et al., 2021)	900M	37.6	63.8	72.6	-	-	-	-	-	-
mPLUG (Li et al., 2022a)	14M	38.1	59.2	68.2	-	-	-	-	-	-
HiTeA (Ye et al., 2022)	17M	34.4	60.0	69.9	43.2	69.3	79.0	18.3	36.7	44.2
OmniVL (Wang et al., 2022b)	18M	42.0	63.0	73.0	40.6	64.6	74.3	-	-	-
mPLUG-2	17M	**47.1**	**69.7**	**79.0**	**45.7**	**71.1**	**79.2**	**24.1**	**43.8**	**52.0**

Table 20. **Zero-shot evaluation on video question answering.** Accuracy is reported.

Method	# PT Data	MSRVTT-QA	MSVD-QA
Just Ask (Yang et al., 2021a)	69M	2.9	7.5
LAVENDER (Li et al., 2022e)	5M	4.5	11.6
MERLOT Reserve (Zellers et al., 2021)	1B	5.8	-
FrozenBiLM (Yang et al., 2022)	10M	6.4	11.7
BLIP (Li et al., 2022c)	129M	19.2	35.2
mPLUG (Li et al., 2022a)	400M	21.1	37.2
HiTeA (Ye et al., 2022)	5M	21.7	37.4
mPLUG-2	17M	**43.8**	**55.3**

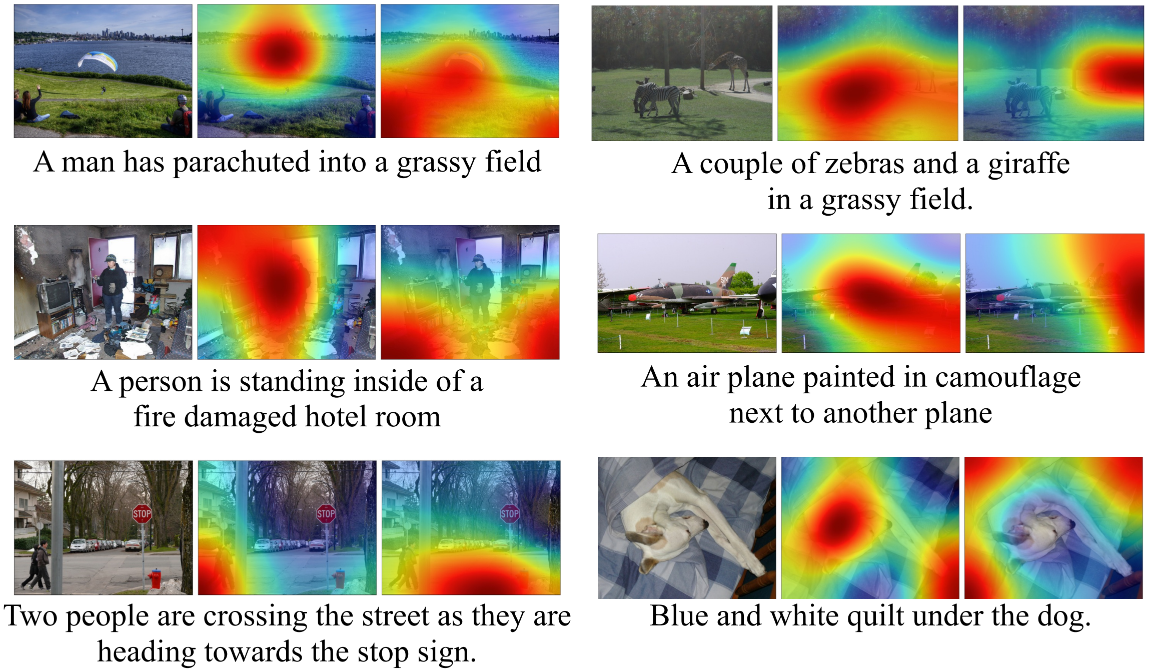
A. More Results

A.1. Detection and Segmentation

We evaluate the object detection and instance segmentation performance of mPLUG-2 on COCO dataset (Lin et al., 2014), which is widely used for object-level detection and segmentation with 80 common categories. Table 18 reports the results on COCO dataset. We observe that mPLUG-2 outperform typical state-of-the-art resnet-based detection methods (e.g., DETR (Carion et al., 2020) and Pix2seq (Chen et al., 2021)). There is a performance gap between foundation model optimized for computer vision (e.g., Florence (Yuan et al., 2021) and Swin-Transformer (Liu et al., 2021b)) and mPLUG-2 . Note that mPLUG-2 does not pre-trained with vision only task and data. Lower performance than models pre-trained on ImageNet is to be expected.

A.2. Zero-Shot Transferability

Text-to-Video Retrieval For testing the transferability of pre-trained mPLUG-2 , we conduct the zero-shot evaluation on Text-to-Video Retrieval and the results are summarized in Table 19. We can find that mPLUG-2 obtains SoTA results on both MSRVTT, DiDeMo and LSMDC datasets, and outperforms previous methods by a large margin, such as 5.1 point of R@1 on the MSRVTT dataset. The results prove that our mPLUG-2 has excellent zero-shot transferability.


A man has parachuted into a grassy field

A couple of zebras and a giraffe in a grassy field.

A person is standing inside of a fire damaged hotel room

An air plane painted in camouflage next to another plane

Two people are crossing the street as they are heading towards the stop sign.

Blue and white quilt under the dog.

Figure 5. Grad-CAM visualizations for latent queries in the universal layers.

Video Question Answering We testing the transferability of pre-trained mPLUG-2 on Video QA and the results are summarized in the Table 20. It can be observed that mPLUG-2 achieves the best zero-shot performance on both MSRVTT-QA and MSVD-QA datasets, which demonstrates the strong zero-shot transferability of our model under the help of universal module and instructional-based learning.

A.3. Visualization of Universal Layer

In Figure 5, we visualize the Grad-CAM on the cross-attention map in the first universal layer. For each sample, we present two cross-attention maps that attend to different visual concepts. The results show that the universal layer can encourage modality collaboration and modality entanglement between visual patch features and language features by attending the areas of various visual concepts in the image.

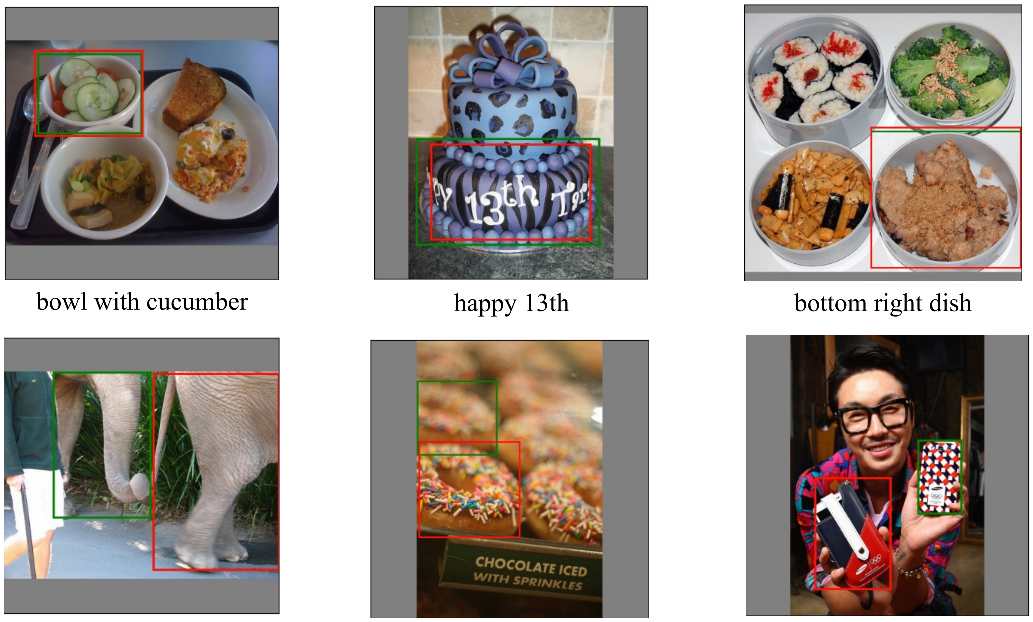
A.4. Visual Grounding

Visualization We visualize several cases of visual grounding task in Figure 6. The first row shows that our mPLUG-2 can understand various visual concepts and their relationships. It also can make fine-grained alignment between vision and language. The second row presents several failure cases. In the first sample, "trunk" is an ambiguous which result in a incorrect prediction. In the second sample, mPLUG-2 fail to recognize the blurred "donut". In the thrid sample, mPLUG-2 does not realize the left and right are reversed in a mirror and predict the "left" item.

B. Implementation Details

B.1. Pre-training

Our models are implemented in the PyTorch framework (Paszke et al., 2019). In detail, we instantiate the text encoder with BERT (Devlin et al., 2018) model pre-traiend on Wikipedia and Bookcorpus (Zhu et al., 2015). The visual encoder is initialized from CLIP-ViT (Radford et al., 2021) pre-trained on 400M noisy image-text pairs. For the base size of model namely mPLUG-2$_{\text{Base}}$, we use the ViT-B/16 for vision encoder and BERT-Base (Devlin et al., 2018) as the text encoder as well as the text decoder. For mPLUG-2 , we scale up the vision and text encoders with ViT-L/14 (Dosovitskiy et al., 2020) and BERT-Large (Devlin et al., 2018) respectively. $C = 768$ and $C = 1024$ for mPLUG-2$_{\text{Base}}$ and mPLUG-2 . We set $S = 2$ for universal layers for the good empirical performance, and choose $G = C$ for multi-group mechanism in the



bowl with cucumber

happy 13th

bottom right dish

trunk

second left donut

the item in his left hand

Figure 6. The visualization of visual grounding task. Green denotes the ground-truth box and red denotes the predicted bounding box.

local temporal modeling module empirically. The number of layers for fusion module is set to 3 for mPLUG-2$_{\text{Base}}$ and 6 for mPLUG-2 , while the number of shared decoder layer is set to 12 for both mPLUG-2$_{\text{Base}}$ and mPLUG-2 . We pre-train the model for 30 epochs with the total batch size of 1024 on 8 NVIDIA A100 GPUs for mPLUG-2$_{\text{Base}}$ and batch size of 512 on 16 NVIDIA A100 GPUs. We use AdamW (Loshchilov & Hutter, 2019) optimizer with the weight decay factor 0.02 and betas (0.9, 0.98) for stabilizing the learning. The learning rate is firstly warmed up to lr_{max} in the first 5000 iterations then decays following the cosine annealing schedule. lr_{max} is set to 1e-4 for mPLUG-2$_{\text{Base}}$ and 5e-5 for mPLUG-2 . During the pre-training, we randomly crop the images and video frames into 224×224 resolution and sparsely sample 4 frames for each video while preserving their order in-between. For vision-text contrastive learning, the queue size and the momentum coefficient are set to 65,536 and 0.995 respectively.

B.2. Downstream Tasks

B.2.1. Vision Only Tasks

Video Action Recognition We first train mPLUG-2 on the Kinetics-710 dataset (Li et al., 2022d) for 40 epochs which is the combination of Kinetics-400, Kinetics-600 and Kinetics-700 by removing the videos represented in the validation and test sets. Specially, the base learning rate is set to 1e-5 for mPLUG-2$_{\text{Base}}$ and 5e-6 for mPLUG-2 with batch size 256 and 128 respectively. Then fine-tuning on Kinetics-400, Kinetics-600, and Kinetics-700 individually for 5 epochs with the same learning rate and batch size.

Image Classification We finetune mPLUG-2 for 30 epochs with the learning rate of 6e-5 and a batch size of 4096. We use the RandomCrop, HorizontalFlip, RandAug and RandErase transformations for data augmentation.

Object Detection and Segmentation We keep the same setting as EVA (Fang et al., 2022b) to train mPLUG-2 on object detection and segmentation tasks. The different is that we do not pre-train mPLUG-2 on Object365 (Shao et al., 2019) before fine-tuning on MSCOCO.

B.2.2. Language Only Tasks

Natural Language Understanding Following (Wang et al., 2022d), we select the best hyperparameters in a suitable range for fine-tuning. We tune the training epochs among 5, 7, 10, learning rate among 3e-5, 5e-5, 6e-5, 7e-5, 1e-4, batch size among 32, 64, 128. We report the best performance on the development set for each task.

Natural Language Generation Following (Wang et al., 2022d), we finetune mPLUG-2 for 50,000 steps with a learning rate of 3e-5 and a batch size of 256. During reference, we beam size with 5 and max generation length with 512.

B.2.3. Video-Text Multi-modal Tasks

For all video-language downstream tasks, we resize video frames to 224 × 224. During fine-tuning, we randomly sample 12 frames for text-to-video size video frames, 16 frames for video question answering and video captions. We perform uniform sampling during inference. We use RandomCrop with minimum ratio 0.5 and HorizontalFlip with 0.5 probability for data augmentation.

Text-to-Video Retrieval We train mPLUG-2_{Base} and mPLUG-2 on the training set of MSRVTT/DiDeMo/LSMDC for 10 epochs with a learning rate of 2e-5 and batch size of 192.

Video Question Answering We train mPLUG-2_{Base} and mPLUG-2 on the training set of MSRVTT-QA/MSVD-QA/TGIF-FrameQA for 10 epochs with a learning rate of 2e-5 and batch size of 128.

Video Captioning For the video caption task, we use a prefix prompt ”What does the video describe?” to improve the quality of generated captions. We set the same training parameters for both the MSRVTT and MSVD datasets. Specifically, we fine-tune mPLUG-2_{Base} and mPLUG-2 with cross-entropy loss on the training set for 10 epochs with a learning rate of 2e-5 and a batch size of 128. Then, we perform CIDEr optimization for extra 5 epochs with a learning rate of 1e-6 and a batch size of 16. Finally, we evaluate the test set with a beam size of 5 and max generation length of 25.

B.2.4. Image-Text Multi-modal Tasks

We resize image frames to 336/576/384/336 for the retrieval/vqa/captioning/grounding tasks. We use ResizedCrop with a minimum ratio of 0.5 and HorizontalFlip with 0.5 probability for data augmentation. We perform center crop during inference.

Image-Text Retrieval We train mPLUG-2 on the training set of MSCOCO/Flickr30K for 8 epochs with a learning rate of 1e-5 and batch size of 512.

Visual Question Answering We train mPLUG-2_{Base} on the VQA dataset for 8 epochs with a learning rate of 3e-5 and batch size of 512.

Image Captioning For the image caption task, we use a prefix prompt ”What does the image describe ?” to improve the quality of generated captions. we first fine-tune mPLUG-2 with cross-entropy loss on COCO training set for 5 epochs with a learning rate of 1e-5 and a batch size of 256. Then we evaluate on the COCO Caption Karpathy validation split and reuse it to predict the Nocaps validation set directly. During inference, we use beam search with a beam size of 5 and set the maximum generation length as 25.

Visual Grounding We first train the model with RefCOCO series datasets with a learning rate of 2e-5 for 120 epochs. Then we continue fine-tuning the model on each dataset with a learning rate of 2e-6 epochs for 30 epochs. We limit the query length to 20/40 for RefCOCO and RefCOCOg, respectively.

B.3. Dataset Description

Text-to-Video Retrieval We evaluate mPLUG-2 on three popular text-to-video retrieval datasets including MSRVTT (Xu et al., 2016), DiDeMo (Anne Hendricks et al., 2017), and LSMDC (Rohrbach et al., 2015).

- **MSRVTT** consists of 10K YouTube sourced videos with 200K text descriptions. Following (Li et al., 2022e; Luo et al., 2022; Huang et al., 2022a), the dataset is divided into 9K and 1K videos for training and testing.
- **DiDeMo** consists of 10K videos from Flickr and each video with 4 descriptions. Following (Li et al., 2022b; Ma et al., 2022; Li et al., 2022e), we concatenate all descriptions of a video as a paragraph, and evaluate the paragraph-to-video retrieval performance. The dataset is separated into 8K for training, 1K for validation and 1K videos for test.

- **LSMDC** consists of 118,081 video clips from 202 movies. Following the standard splits from (Rohrbach et al., 2015), the dataset is divided into 101K and 1K videos for training and testing.

Video Question Answering We evaluate mPLUG-2 on three popular video question answering datasets including MSRVTT-QA (Xu et al., 2017) MSVD-QA (Xu et al., 2017), and TGIF-FrameQA (Jang et al., 2017).

- **MSRVTT-QA** is based on the MSRVTT dataset (Xu et al., 2016). The QA pairs are automatically generated by from the descriptions. This benchmark composed of 243K open-ended questions over 10K videos.
- **MSVD-QA** is based on the MSVD datasets (Chen & Dolan, 2011) with automatically generated QA pairs. It consists 2K videos with 47K questions.
- **TGIF-FrameQA** collects the answerable with just a single frame in the video, and is divided into training set with 35K questions and test set with 14K questions.

Video Captioning We use MSRVTT (Xu et al., 2016) and MSVD (Chen & Dolan, 2011) for video captioning evaluation.

- **MSRVTT** is composed of 10K videos with 20 captions per video as described above. We take the same data split as text-to-video retrieval task.
- **MSVD** contains 1970 YouTube short video clips. Following the standard splits from (Lin et al., 2022; Li et al., 2022e), we separate the dataset into 1,200 train, 100 validation and 670 test videos.

Visual Question Answer We evaluate our method on the VQA 2.0 dataset (Agrawal et al., 2017).

- **VQA 2.0** is a dataset containing open-ended questions about images and at least 3 questions (5.4 questions on average) per image. It contains 83k/41k/81k images for training/validation/test.

Image-Text Retrieval Two popular image-text retrieval benchmarks, COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) are used to evaluate the model. We adopt the widely-used Karpathy split (Karpathy & Fei-Fei, 2015) for both COCO and Flickr30K.

- **COCO** has over 330k images and 5 independent human generated captions are be provided for each image. It contains 113k/5k/5k images for training/validation/testing.
- **Flickr30K** contains 31k images from Flickr, each image with 5 human annotated sentences. It contains 29k/1k/1k images for training/validation/testing.

Image Captioning We evaluate our method on COCO (Lin et al., 2014) datasets.

- **COCO** takes the same data split as the image-text retrieval task.

Natural Language Understanding To verify the natural language understanding ability of our mPLUG-2 , we select 6 language understanding datasets from GLUE (Wang et al., 2018) benchmark, including both single-sentence classification tasks and sentence-pair classification tasks.

- **SST-2** The Stanford Sentiment Treebank consists of sentences from movie reviews and human-annotated sentiment. The task is to predict the sentiment of a given sentence.
- **RTE** The Recognizing Textual Entailment dataset comes from a series of annual textual entailment challenges.
- **MRPC** The Microsoft Research Paraphrase Corpus consists of a corpus of sentence pairs collected from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent.

- **QQP** The Quora Question Pairs dataset is a collection of question pairs from the community question-answering website Quora. The task is to predict whether a pair of questions are semantically equivalent.
- **MNLI** The Multi-Genre Natural Language Inference Corpus consists of sentence pairs (premise, hypothesis) with textual entailment annotations. The task is to predict the entailment between the premise and the hypothesis.
- **QNLI** The Stanford Question Answering Dataset is a question-answering dataset, where one of the sentences in the paragraph (drawn from Wikipedia) contains the answer to the corresponding question (written by an annotator). The task is to determine whether the context sentence contains the answer to the question.

Natural Language Generation We use Gigaword dataset (Rush et al., 2015a) for text summarization task to verify the natural language generation ability of our mPLUG-2 .

- **Gigaword** Headline-generation on a corpus of article pairs from Gigaword consisting of around 4 million articles. It contrains 3803957, 189651 and 1951 samples for training/validation/testing.

Video Action Recognition We adopt three popular benchmarks Kinetics 400/600/700 dataset (Kay et al., 2017) to evaluate our model.

The videos in these three benchmarks are collected from YouTube. Each video clip lasts around 10 seconds and is labeled with a single action class. The videos include human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging.

- **Kinetics 400** consists of 240K training videos and 20K validation videos that span 400 human action categories.
- **Kinetics 600** consists of 392K training videos and 30K validation videos spanning 600 action categories.
- **Kinetics 700** consists of 545K training videos and 35K validation videos spanning 700 action categories.

Image Classification We evaluate performance of mPLUG-2 in terms of image classification on ImageNet-1K (Deng et al., 2009).

- **ImageNet-1K** contains 1.28M training images and 50K validation images from 1,000 classes.