Bootstrapping Vision-Language Learning with Decoupled Language Pre-training

Yiren Jian

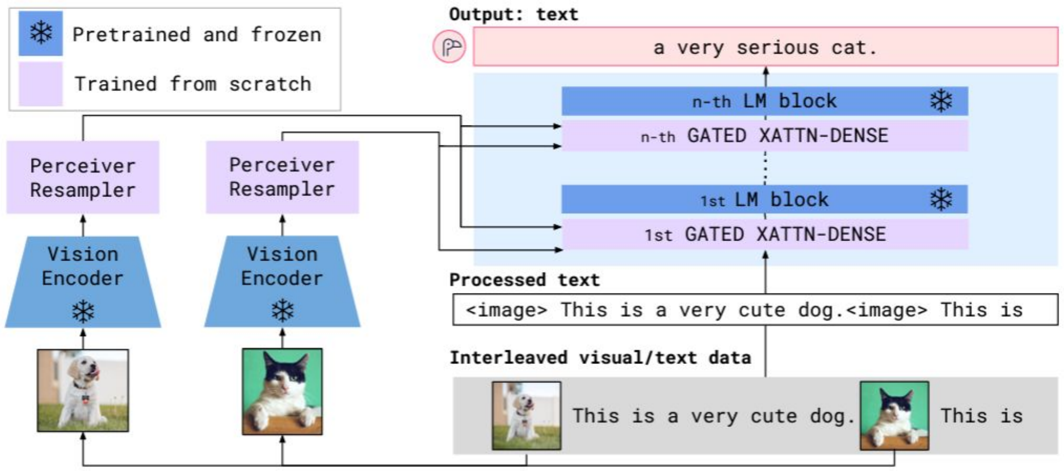
Chongyang Gao

Soroush Vosoughi

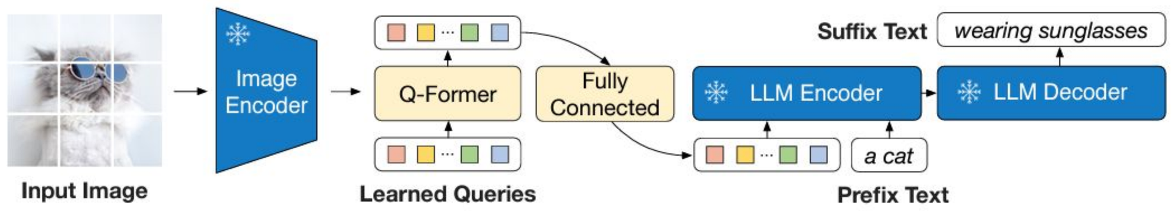
DARTMOUTH
Department of Computer Science

Northwestern | McCORMICK SCHOOL OF ENGINEERING
Computer Science

Background: Vision-Language Models (VLMs) with Frozen LLM



Flamingo [1]

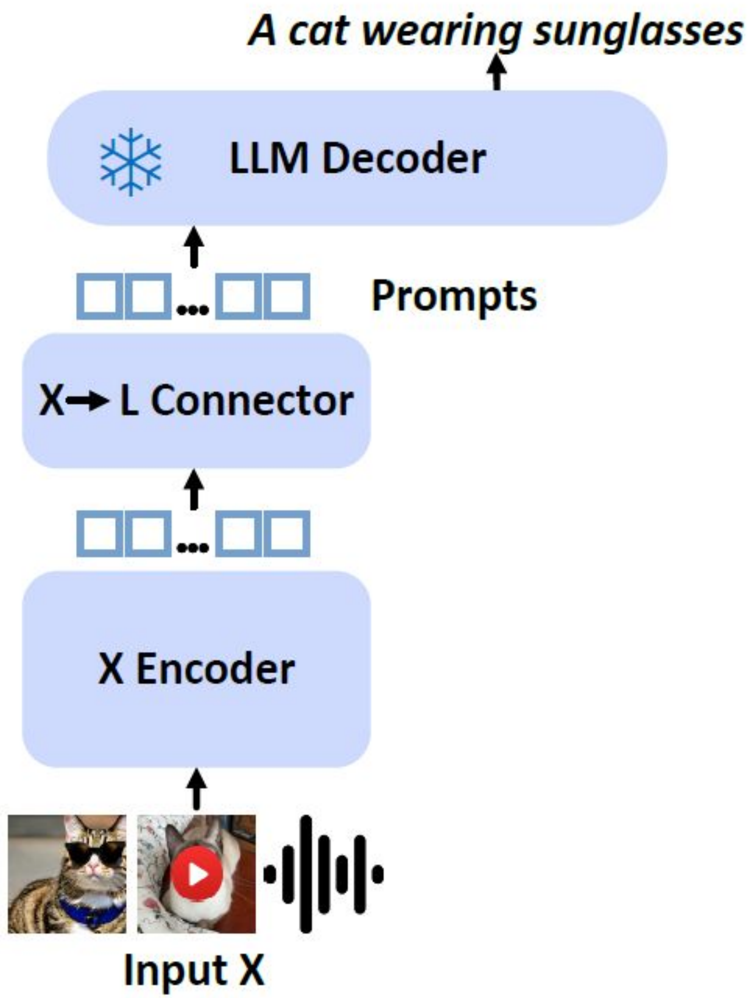


BLIP-2 [2]

Vision-language models are the foundation for various tasks including visual-question-answering (VQA), image captioning, image-text retrieval, and visual reasoning. Current paradigm pre-trains VLMs with frozen LLMs using image-text pairs.

[1] Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." NeurIPS, 2022.
[2] Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.

End-to-End Training of VLMs

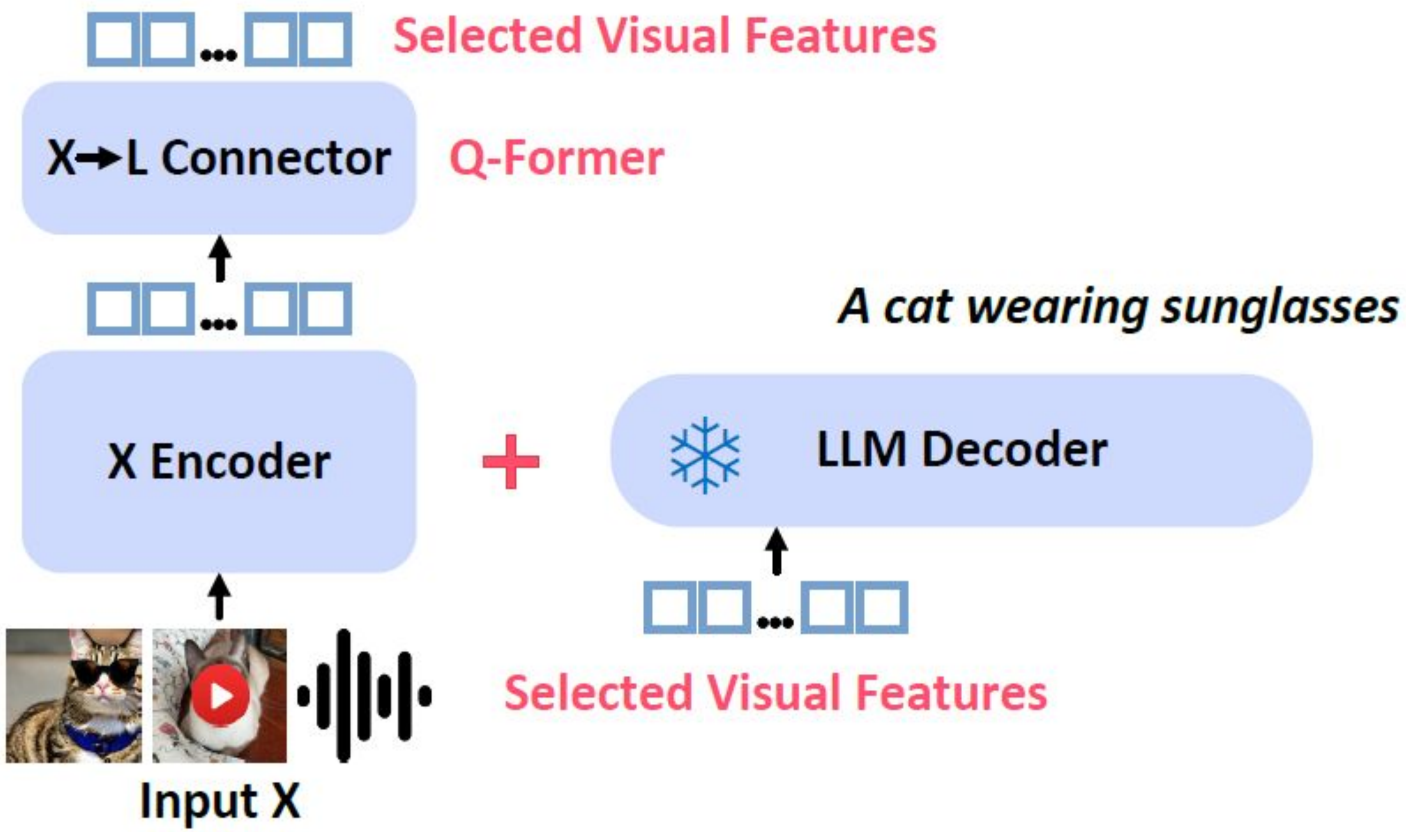


VLMs can be trained end-to-end in an image-conditioned language generation task

- limitation: End-to-end optimization of such a model is challenging [1]

[1] Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.

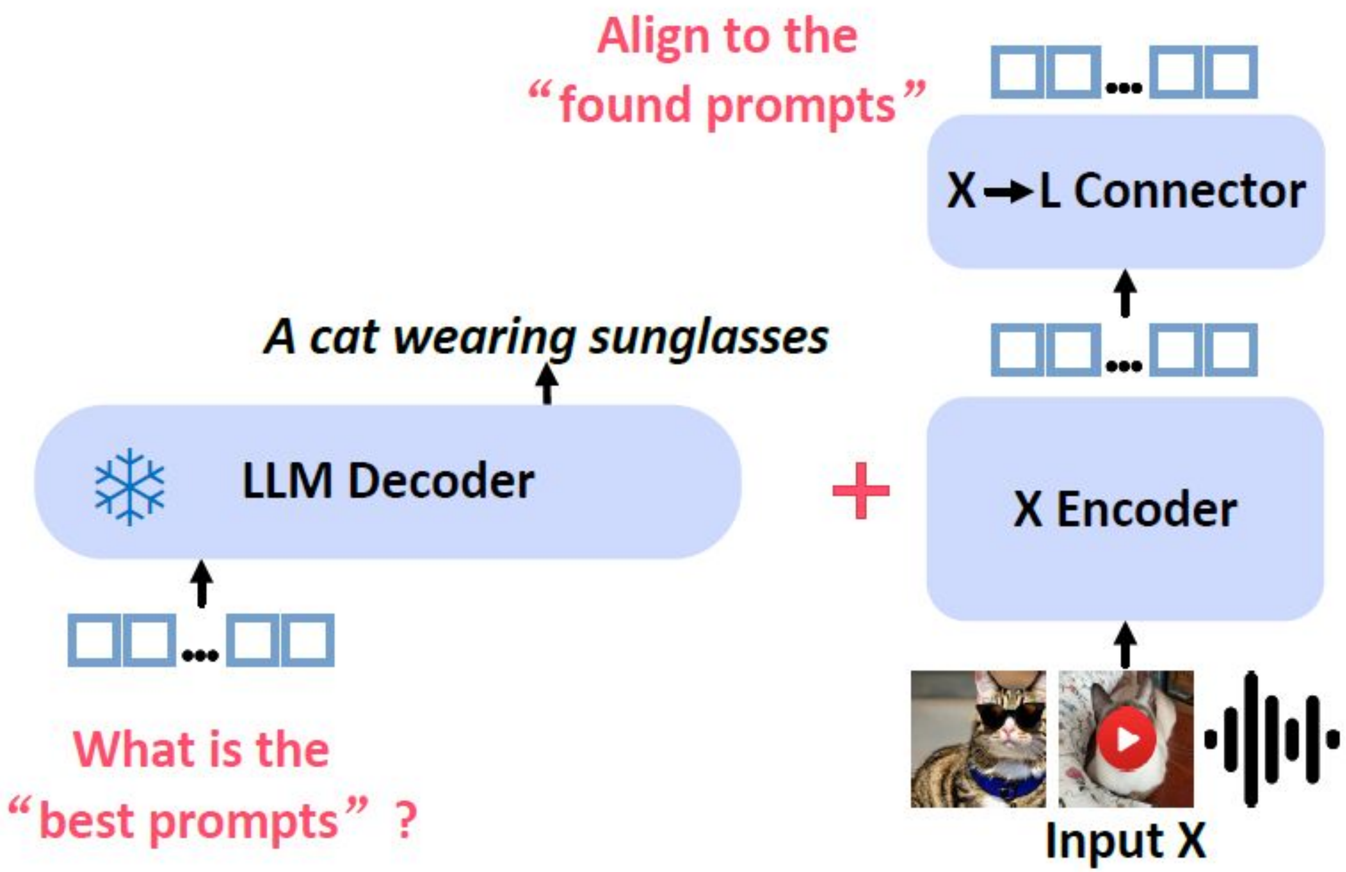
VLM Training: Two-Stage Approach



BLIP2 proposes a two-stage training for effective pre-training of VLMs using frozen LLMs.

1. Representation learning of a **Q-Former** to extract most text-informative visual features.
2. Aligning the selected visual features to the corresponding text.

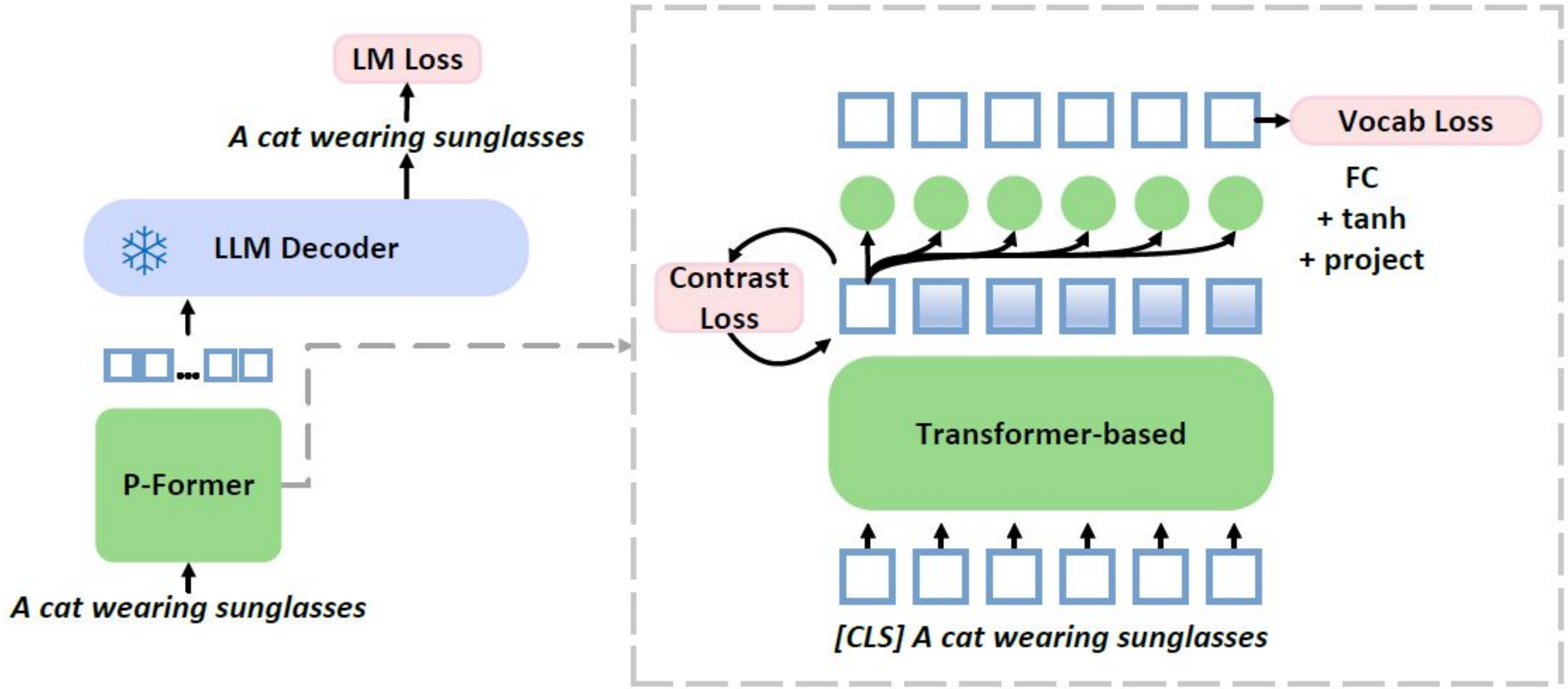
VLM Training: Finding the Ideal Soft-Prompts



We provide a novel insight to mitigate the challenges in end-to-end optimization by introducing ''backward-decoupling'' during back-propagation.

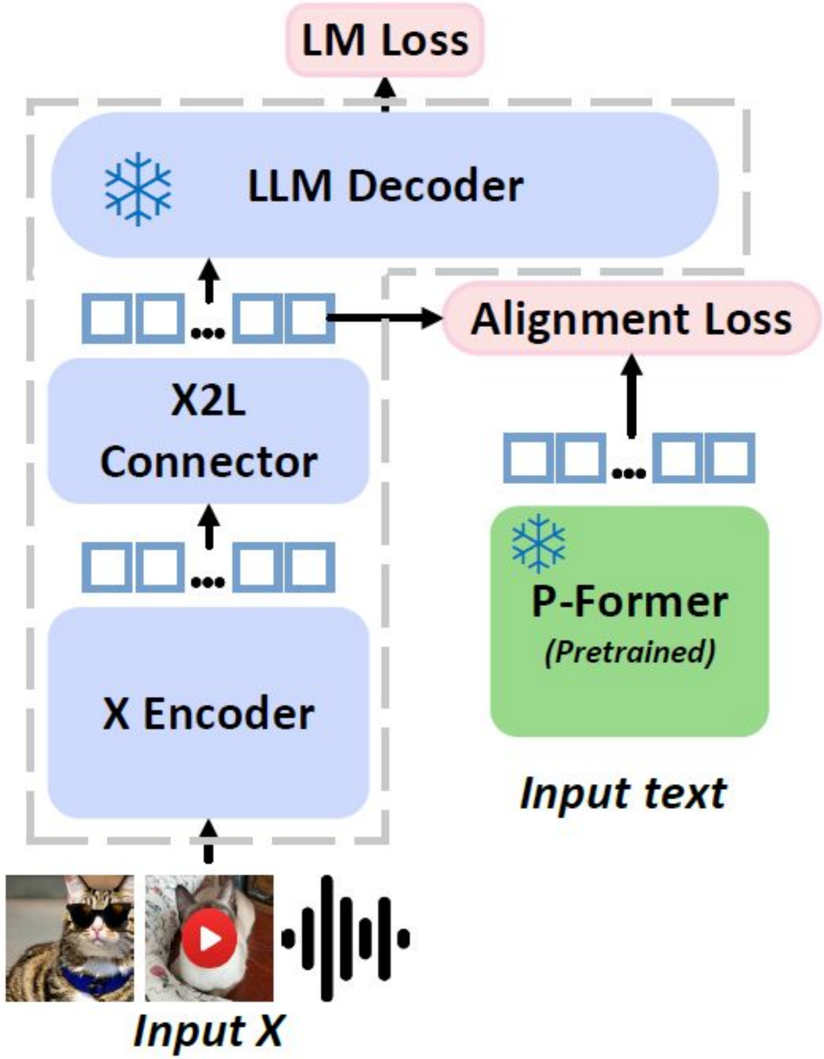
1. learning an ideal soft-prompt of the LLM, given the target text.
2. Aligning the visual features to the learned soft-prompts.

Training of Prompt-Transformer (P-Former) for Soft-Prompts Predictions



- The P-Former training resembles an autoencoder, with the bidirectional P-Former as the encoder and a causal LLM (frozen) as the decoder.
- The objective is to reconstruct input text auto-regressively. [CLS] representation serves as sentence embeddings, which are projected back to the length of prompts.
- **This training process is purely based on text, allowing the P-Former to benefit from text outside the image-text pair dataset.**

Training VLMs with P-Former



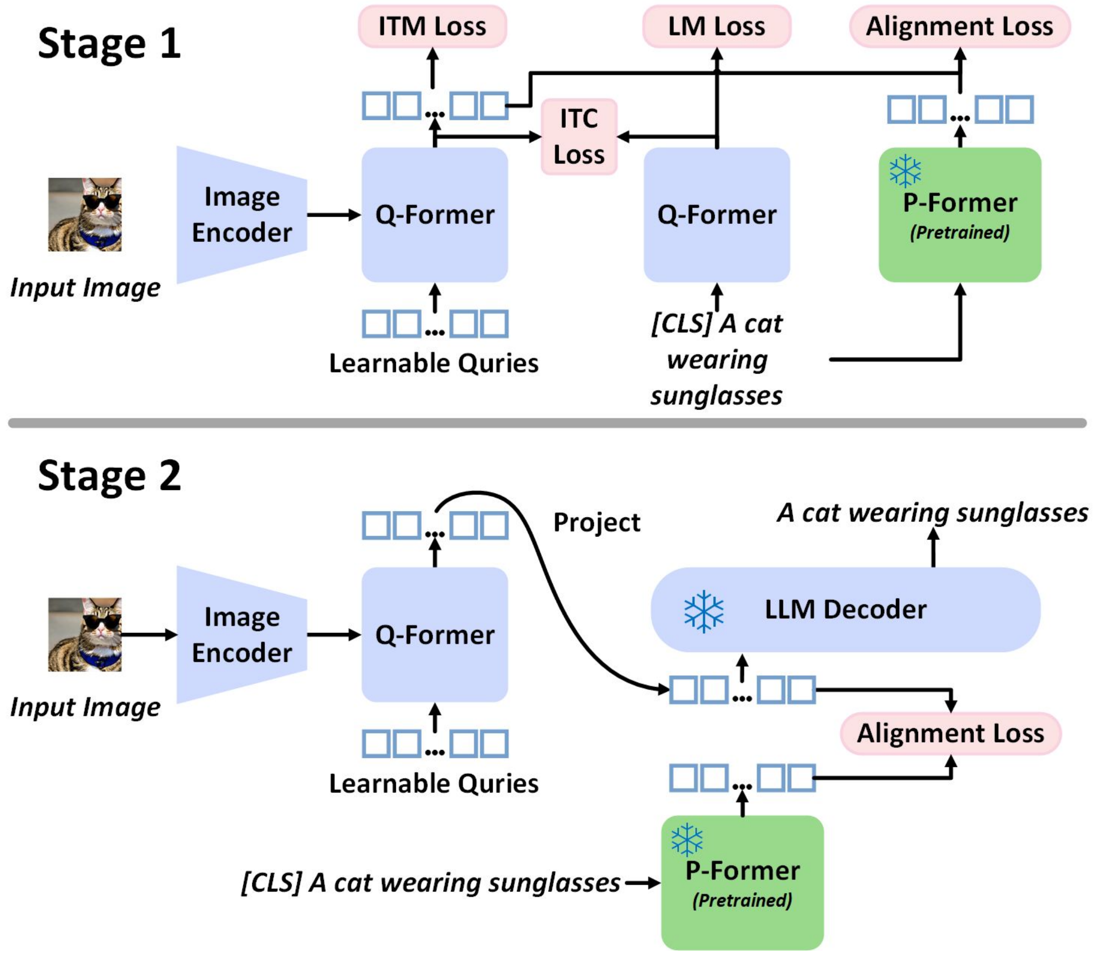
Overview of bootstrapping VL pre-training with the trained P-Former. **The alignment loss introduced by P-Former is agnostic to input modalities, encoders, and X-to-language connection modules.**

Training BLIP2 with P-Former

The overview of applying trained P-Former to the BLIP2 pre-training framework.

$$\mathcal{L}_{\text{BLIP2-stage1}} + \omega_1 \times \mathcal{L}_{\text{alignment}}$$

$$\mathcal{L}_{\text{BLIP2-stage2}} + \omega_2 \times \mathcal{L}_{\text{alignment}}$$



Experimental Results on Zero-Shot VQA

Models	#Pretrain Image-Text	Pretrain Uni-Text	VQAv2 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
FewVLM [24]	9.2M	-	47.7	-	16.5	29.3
Frozen [56]	3M	-	29.6	-	5.9	-
VLKD [9]	3M	-	42.6	44.5	13.3	-
Flamingo3B [2]	1.8B	-	-	49.2	41.2	-
$\text{OPT}_{2.7\text{B}}$ BLIP-2 [34]	4M	-	46.8	45.6	25.9	30.5
$\text{OPT}_{2.7\text{B}}$ Ours	4M	✓	<u>52.6</u>	<u>52.2</u>	<u>30.0</u>	<u>34.0</u>
$\text{OPT}_{2.7\text{B}}$ BLIP-2† [34]	129M	-	**53.5**	**52.3**	**31.7**	**34.6**

Table 1: Comparison with different methods on zero-shot VQA †: numbers taken from Li et al. [34].

Our proposed framework significantly enhances the zero-shot VQA performance of BLIP-2 trained with 4M image-text pairs. Remarkably, the gap between the BLIP-2 trained with 4M and 129M image-text pairs is largely bridged by our method

Experimental Results on Image Captioning

Models	#Pretrain Image-Text	NoCaps Zero-shot (validation set)								COCO Fine-tuned Karpathy test	
		in-domain		near-domain		out-domain		overall			
		C	S	C	S	C	S	C	S	B@4	C
OSCAR [38]	4M	-	-	-	-	-	-	80.9	11.3	37.4	127.8
VinVL [69]	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
BLIP [33]	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7
OFA [58]	20M	-	-	-	-	-	-	-	-	43.9	145.3
Flamingo [2]	1.8B	-	-	-	-	-	-	-	-	-	138.1
SimVLM [61]	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
$\text{OPT}_{2.7\text{B}}$ BLIP-2 [34]	4M	115.3	15.0	111.0	14.6	112.5	14.0	111.9	14.5	41.8	140.4
$\text{OPT}_{2.7\text{B}}$ Ours	4M	118.3	15.3	114.7	14.9	114.1	14.1	115.1	14.8	42.3	141.8
$\text{OPT}_{2.7\text{B}}$ BLIP-2† [34]	129M	**123.0**	**15.8**	**117.8**	**15.4**	**123.4**	**15.1**	**119.7**	**15.4**	**43.7**	**145.8**

Table 2: Comparison with different captioning methods on NoCaps and COCO. All methods optimize the cross-entropy loss during fine-tuning. C: CIDEr, S: SPICE, B: BLEU. †: numbers taken from Li et al. [34].

Our framework improves BLIP-2 in all metrics, with greater improvements in CIDEr compared to SPICE.

Ablation Studies: Alignment Loss on Two Stages

ω_1	ω_2	VQAv2 val	OK-VQA test	GQA test-dev
0	0	46.8	25.9	30.5
10	0	51.4	29.2	32.8
0	100	50.4	28.7	33.0
10	100	**52.6**	**30.0**	**34.0**

Table 4: Ablations on ω_1 and ω_2 of Equation 8 and 9 (using $\text{OPT}_{2.7\text{B}}$ as LLMs).

$$\mathcal{L}_{\text{BLIP2-stage1}} + \omega_1 \times \mathcal{L}_{\text{alignment}}$$

$$\mathcal{L}_{\text{BLIP2-stage2}} + \omega_2 \times \mathcal{L}_{\text{alignment}}$$

The alignment loss introduced by the P-Former proves beneficial in both stages of VL pre-training,

Ablation Studies: Different LLM Decoders

Models	#Pretrain Image-Text	VQAv2 val	OK-VQA test	GQA test-dev
Flan-$T5_{XL}$ BLIP-2‡	4M	48.3	31.5	36.4
Flan-$T5_{XL}$ ours‡	4M	<u>54.9</u>	<u>35.7</u>	<u>40.3</u>
Flan-$T5_{XL}$ BLIP-2†	129M	**62.6**	**39.4**	**44.4**

Table 5: Experiments using Flan-$T5_{XL}$ as LLM. ‡: using much less GPUs/epochs compared to Sec.4.1. †: from Li et al. [34].

Besides the OPT language decoders, we verify the effectiveness of our framework with another LLM.

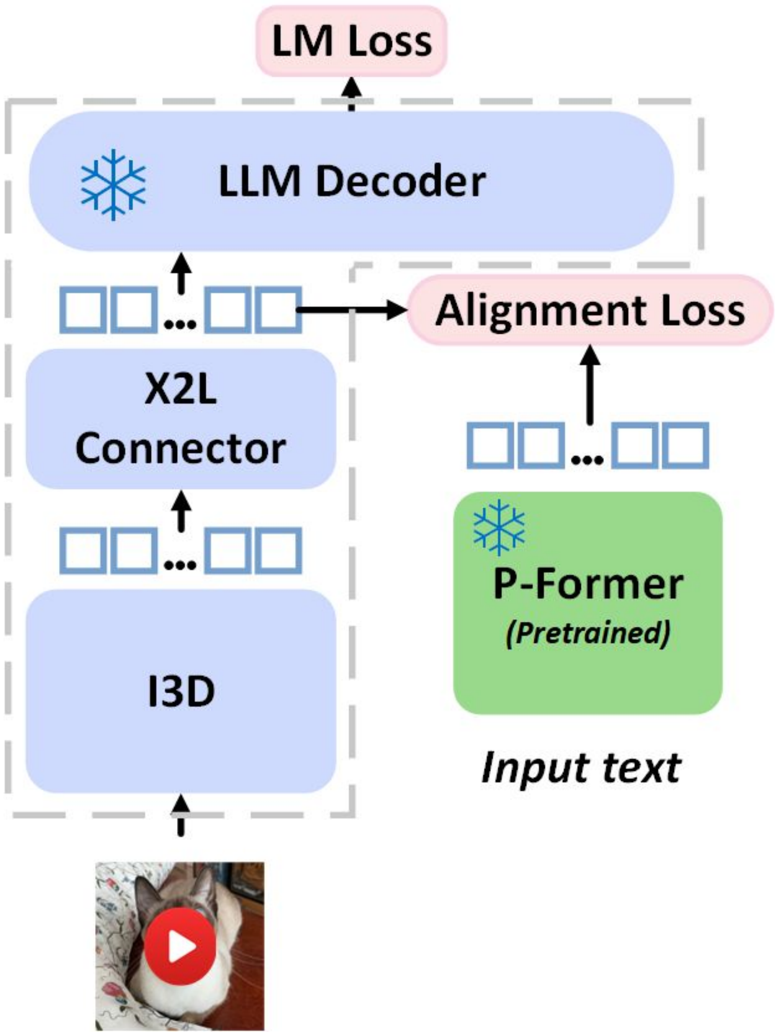
Ablation Studies: Pre-training Datasets for P-Former

P-Former	#Pretrain Sentences	VQAv2 val	OK-VQA test	GQA test-dev
×	-	46.8	25.9	30.5
✓	4M	51.7	28.2	32.3
✓	12M	**52.6**	**30.0**	**34.0**

Table 6: Ablations on sentence datasets used to train P-Former (using $OPT_{2.7B}$ as LLMs). The first row w/o P-Former is baseline BLIP-2.

Both the implicit decoupling of BLIP-2's two-stage training into a 3-stage training (pre-training of P-Former), and the employment of additional unimodal sentences contribute to the improved outcomes

Ablation Studies: Video-Language Tasks



	BLEU-4	CIDEr	ROUGE
NITS-VC [53]	20.0	24.0	42.0
ORG-TRL [71]	32.1	49.7	48.9
$\mathcal{L}_{\text{ITG}}$	29.3	56.6	48.2
$\mathcal{L}_{\text{ITG}} + \mathcal{L}_{\text{alignment}}$	**30.9**	**60.9**	**49.1**

Table 7: VATEX English video captioning. Baseline is a sequential model (I3D $\rightarrow$ Transformer $\rightarrow$ $\text{OPT}_{2.7\text{B}}$), training end-to-end with ITG.

Our framework is modality-agnostic with respect to the visual encoder and vision-to-language adaptor, making it applicable to other modalities, such as video.

Bootstrapping Vision-Language Learning with Decoupled Language Pre-training

Thank you!