Bootstrapping Vision-Language Learning with Decoupled Language Pre-training

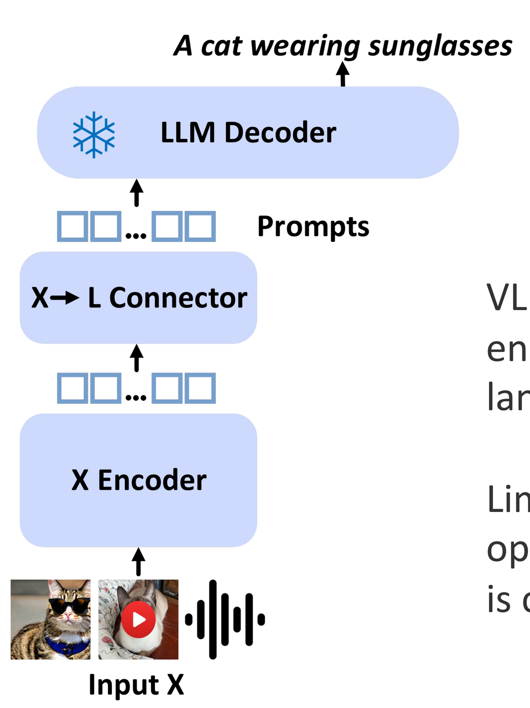
Yiren Jian[1], Chongyang Gao[2] and Soroush Vosoughi [1]

(1) Dartmouth College (2) Northwestern University.

Main Contributions

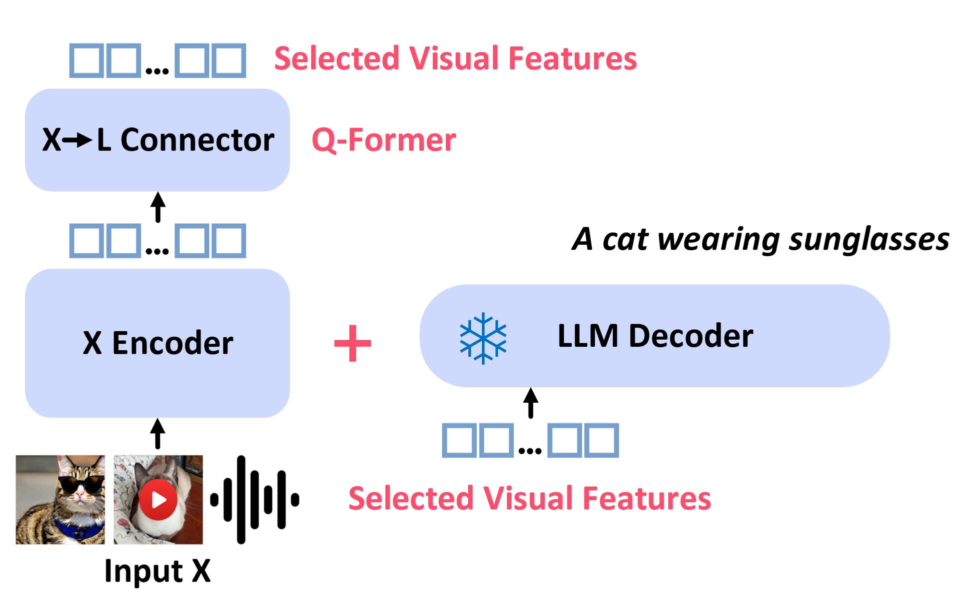
- ✓ Introducing the Prompt-Transformer (P-Former), a model that predicts these ideal prompts, which is trained exclusively on linguistic data, bypassing the need for image-text pairings.
- ✓ Our experiments reveal that our framework significantly enhances the performance of BLIP-2, and effectively narrows the performance gap between models trained with either 4M or 129M image-text pairs.
- ✓ Our framework is modality-agnostic and flexible in terms of architectural design, as validated by its successful application in a video learning task using varied base modules.

Background

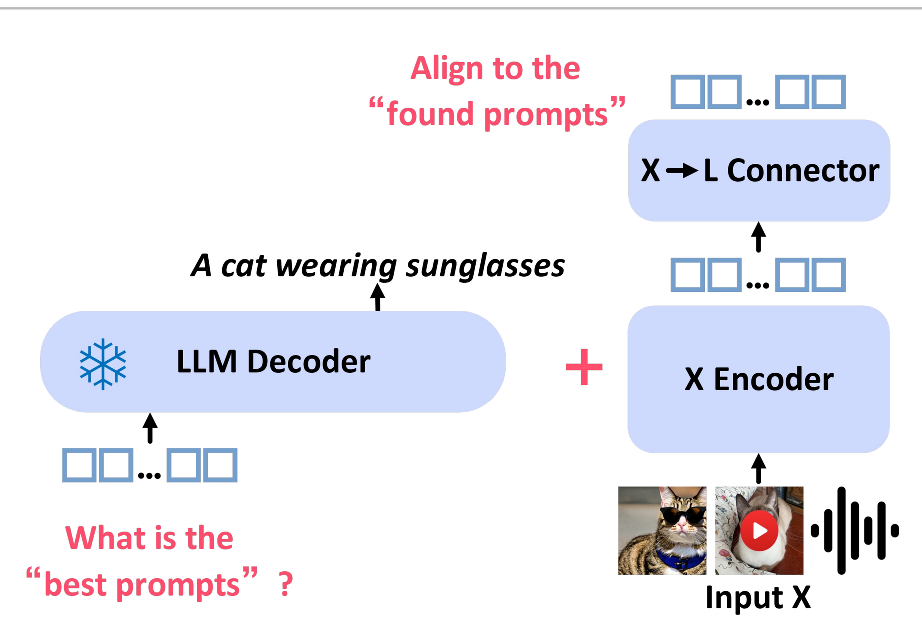


VLMs can be trained end-to-end in an image-conditioned language generation task

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

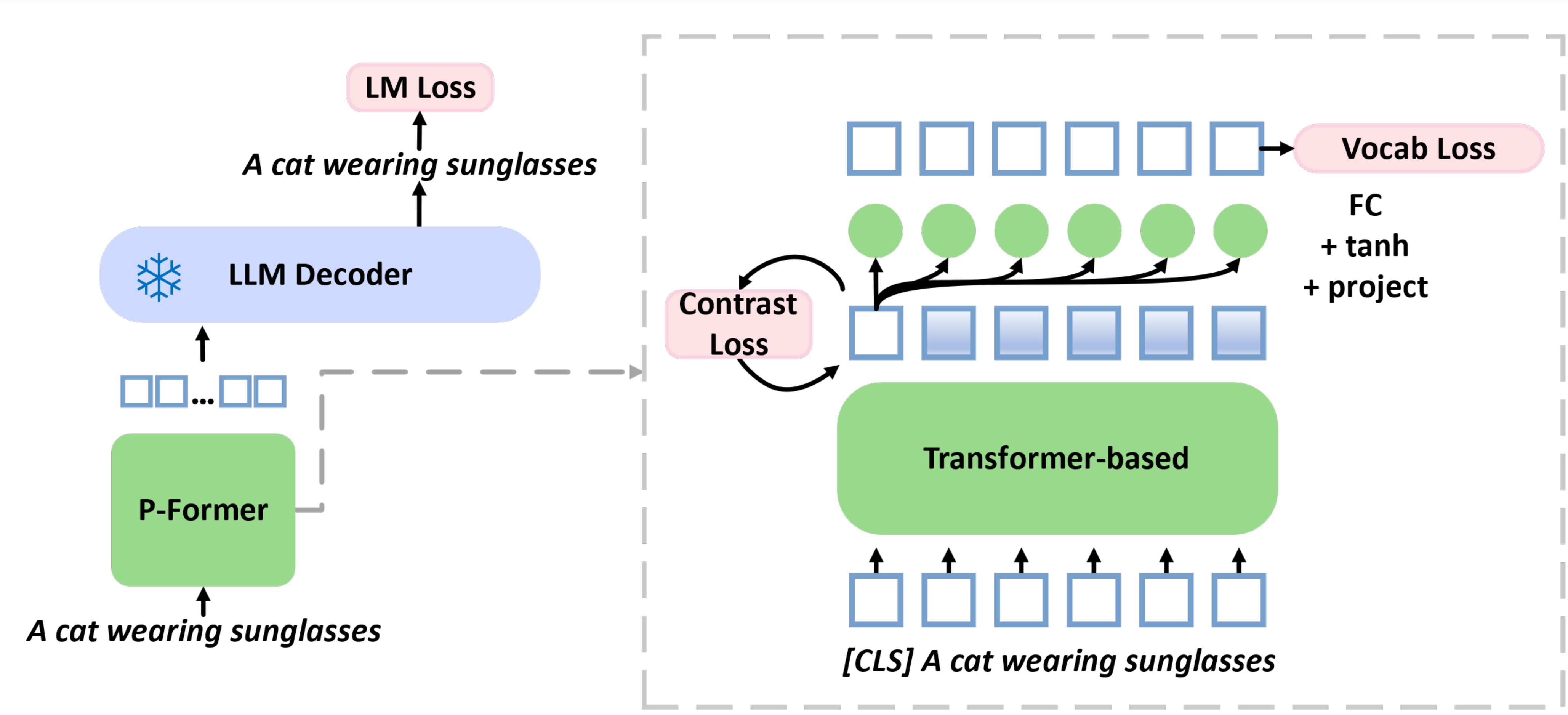


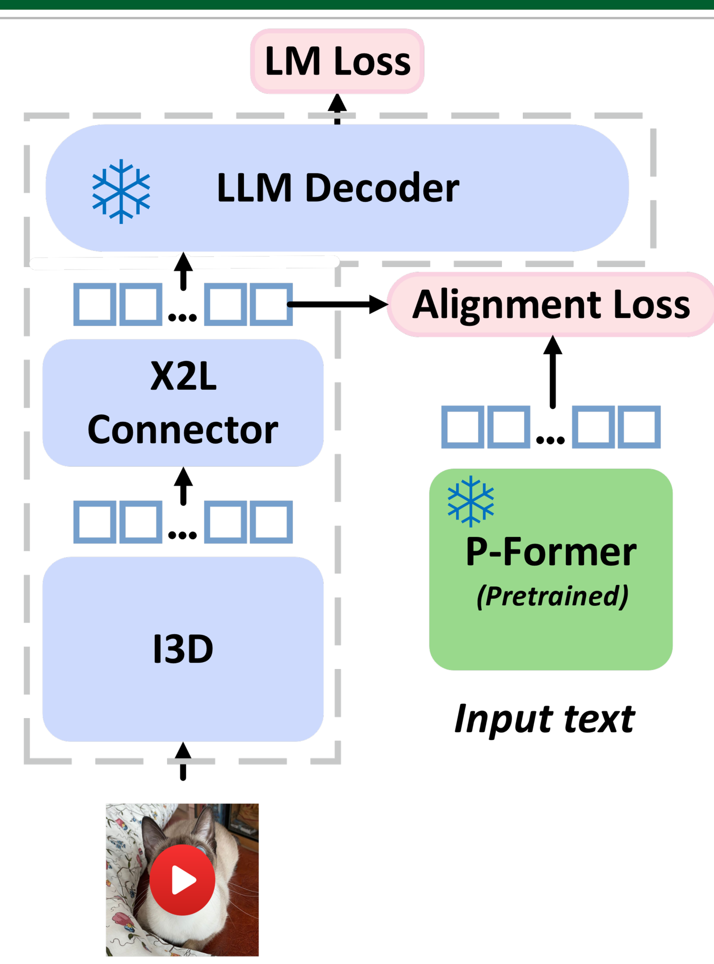
BLIP2 [2] proposes a two-stage training for effective pre-training of VLMs using frozen LLMs.



"Backward-decoupling" during back-propagation.

Training of Prompt-Transformer

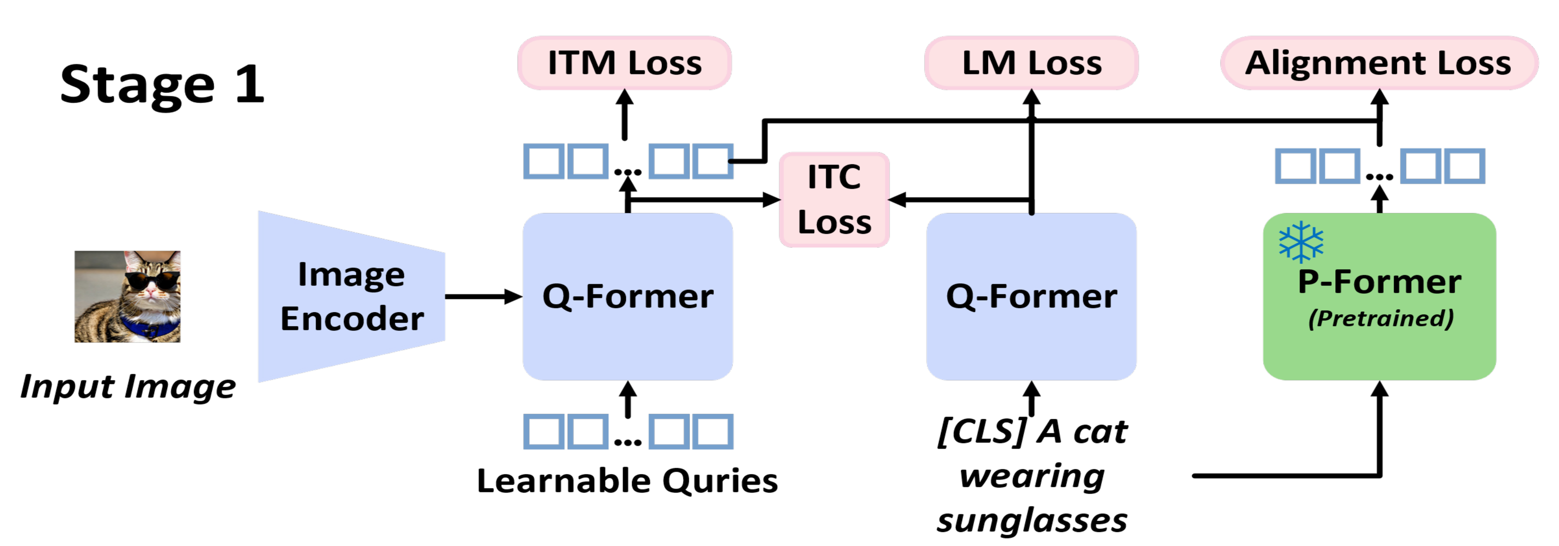


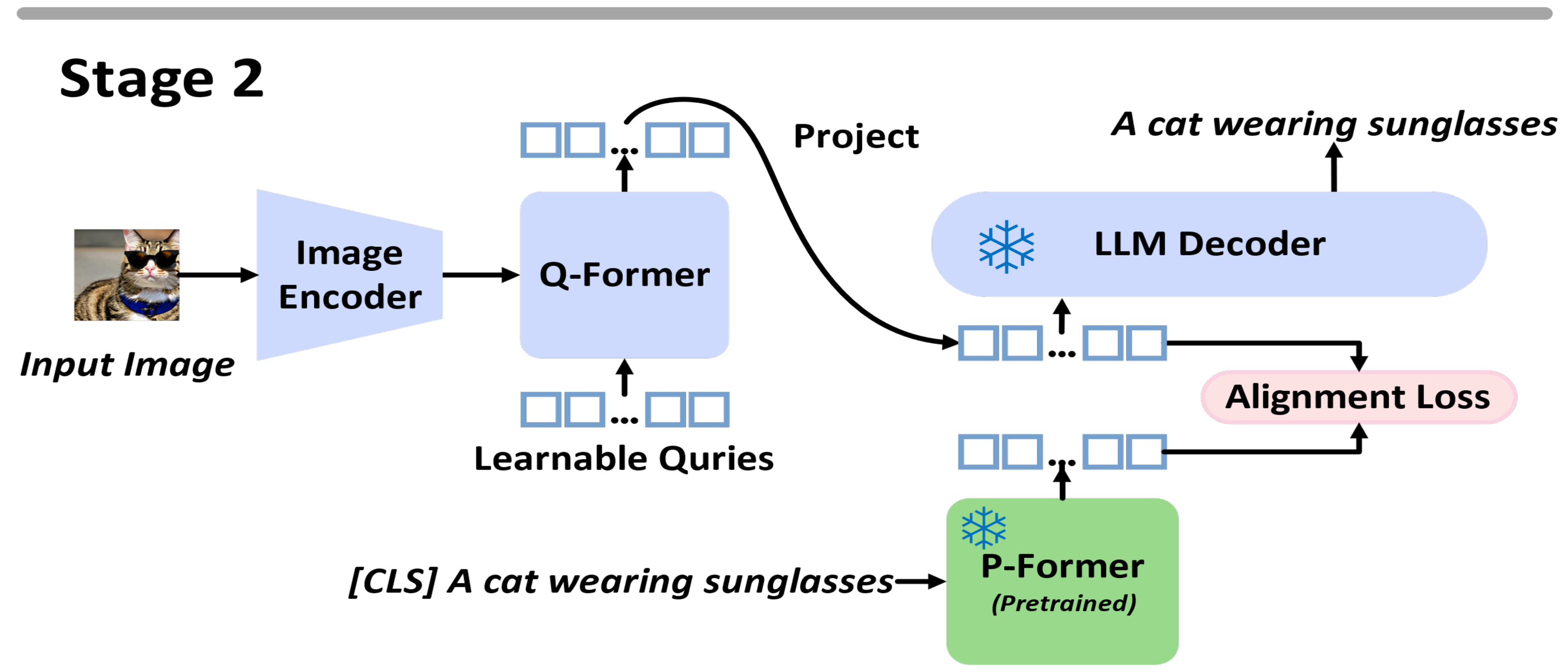


- The P-Former training resembles an autoencoder, with the bidirectional P-Former as the encoder and a causal LLM (frozen) as the decoder.
- The objective is to reconstruct input text auto-regressively. [CLS] representation serves as sentence embeddings, which are projected back to the length of prompts.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

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.

$$L_{BLIP2-stage1} + \omega_1 \times L_{alignment}$$

$$L_{BLIP2-stage1} + \omega_2 \times L_{alignment}$$

Experimental Results

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	-
$OPT_{2.7B}$ BLIP-2 [34]	4M	-	46.8	45.6	25.9	30.5
$OPT_{2.7B}$ Ours	4M	✓	52.6	52.2	30.0	34.0
$OPT_{2.7B}$ BLIP-2† [34]	129M	-	**53.5**	**52.3**	**31.7**	**34.6**

Comparison with different methods on zero-shot VQA.

Our framework significantly enhances the zero-shot VQA performance of BLIP-2 trained with 4M image-text pairs.

Models	#Pretrain Image-Text	NoCaps in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Fine-tuned Karpathy test B@4	COCO 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
$OPT_{2.7B}$ BLIP-2 [34]	4M	115.3	15.0	111.0	14.6	112.5	14.0	111.9	14.5	41.8	140.4
$OPT_{2.7B}$ Ours	4M	118.3	15.3	114.7	14.9	114.1	14.1	115.1	14.8	42.3	141.8
$OPT_{2.7B}$ BLIP-2† [34]	129M	**123.0**	**15.8**	**117.8**	**15.4**	**123.4**	**15.1**	**119.7**	**15.4**	**43.7**	**145.8**

Comparison with different methods on NoCaps and COCO. All methods optimize the CE loss during fine-tuning. C: CEDEr, S: SPICE, B: BLEU

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**

Ablations on *w*1 and *w*2 (using OPT 2.7B).

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	54.9	35.7	40.3
Flan-$T5_{XL}$ BLIP-2†	129M	**62.6**	**39.4**	**44.4**

Experiments using Flan-T5 XL.

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**

Ablations on sentence datasets used to train P-Former (using OPT 2.7B).

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

VATEX English video caption. Baseline is a sequential model, training end-to-end with ITG.

[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.