BLIP

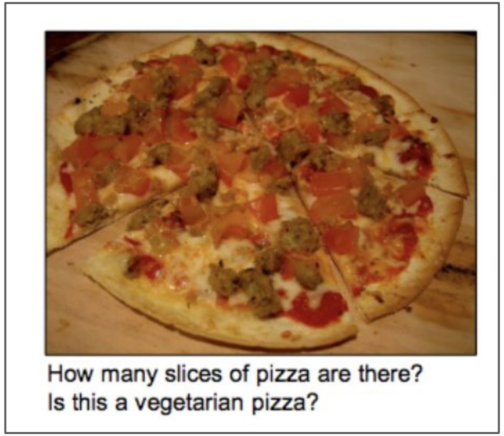
Bootstrapping Language-Image Pre-training
for Unified Vision-Language Understanding and Generation

Arzhantsev Andrei, 202

main sourse: https://arxiv.org/pdf/2201.12086.pdf

Vision-language tasks

- Image-Text Retrieval
- Image Captioning
- Visual Question Answering
- Visual Dialog
- Natural Language Visual Reasoning
- Memes explanation

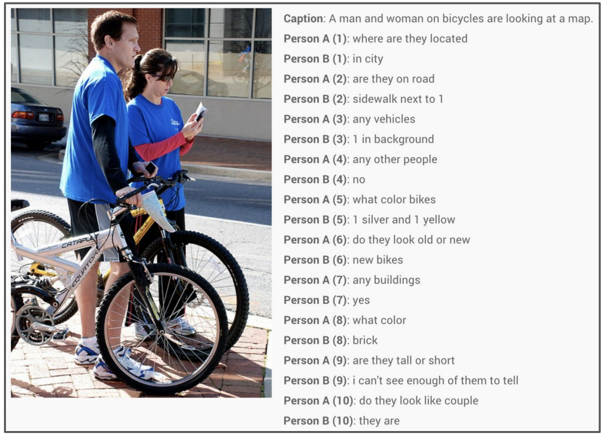


VQA

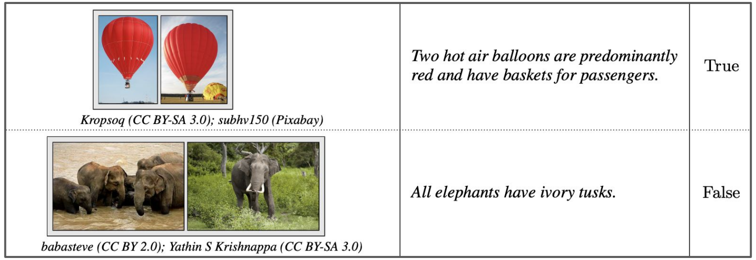
NoCaps

Flickr30k

VisDial

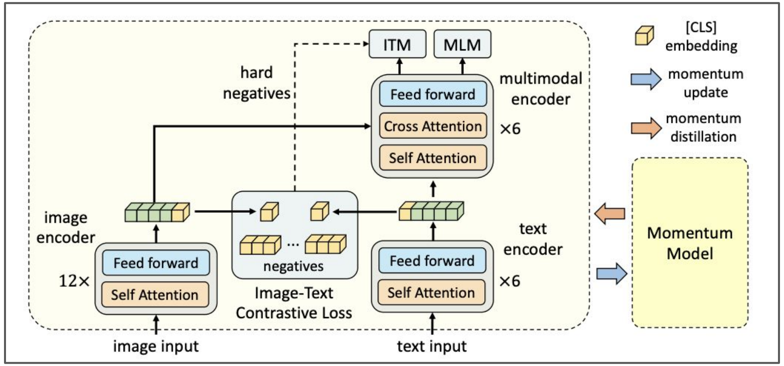


NLVR2

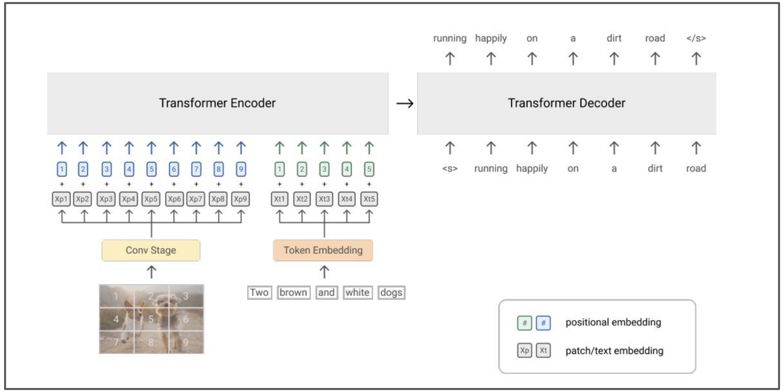


Model perspective

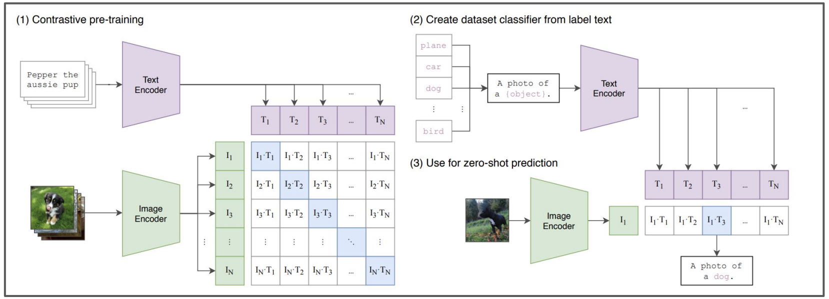
- Encoder-based models
  - CLIP: https://arxiv.org/pdf/2103.00020.pdf
  - ALBEF: https://arxiv.org/pdf/2107.07651.pdf
  - not optimal for text generation tasks
- Encoder-Decoder models
  - VL-T5: https://arxiv.org/pdf/2102.02779.pdf
  - SimVLM: https://arxiv.org/pdf/2108.10904.pdf
  - not optimal for image-text retrieval tasks



ALBEF



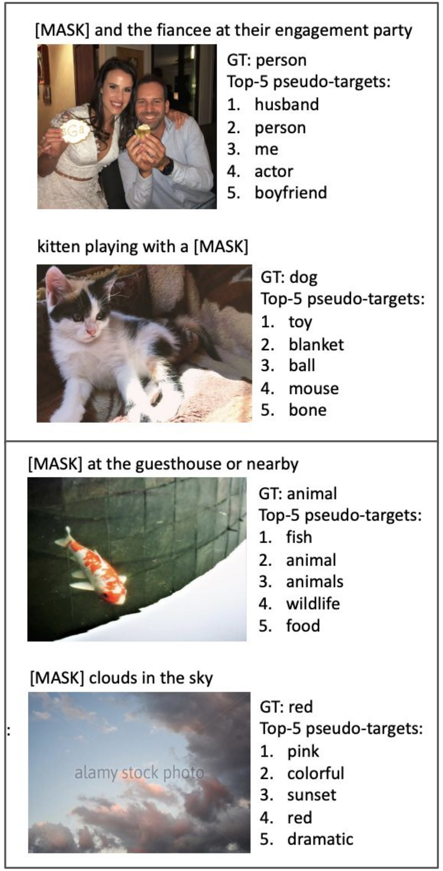
SimVLM



CLIP

Data perspective

- Pre-train on relevant image-text pairs
  - few data, noisy data is bad
- Knowledge Distillation
  - https://arxiv.org/pdf/1503.02531.pdf
  - (ALBEF) Momentum Distillation
  - performs worse than CapFILT used in BLIP
- Data Augmentation
  - well known for CV tasks
  - recently used for NLP tasks
  - no analogue for VL tasks



Momentum distillation in ALBEF

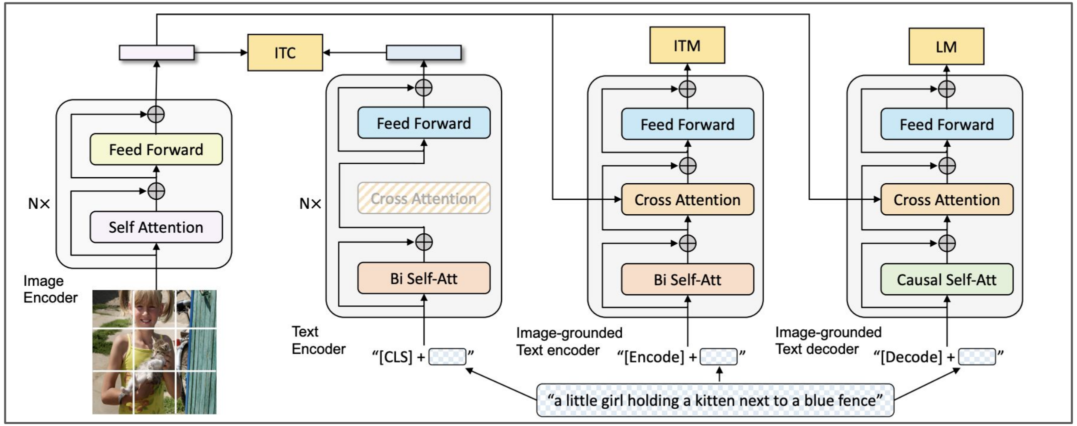
BLIP: Captioning and Filtering (CapFilt)

- Human-annotated Image-Text pairs (Ih-Th)
- Large dataset of web Image-Text pairs (Iw-Tw)
- Captioner
  - Generates captions (Ts) given web images (Iw)
  - Initialized from pre-trained MED model
  - Image-grounded text decoder
  - Finetuned with the LM objective (from Ih-Th)
- Filter
  - Removes noisy image-text pairs
  - Initialized from pre-trained MED model
  - Image-grounded text encoder
  - finetuned with the ITC (contrastive) and ITM (matching) objectives to learn whether a text matches an image
- With CapFilt we have
  - Ih-Th and new filtered Iw-Tw or Iw-Ts

Filtering from both web texts and synthetic texts

BLIP: Multimodal mixture of Encoder-Decoder (MED)

- Unimodal encoder
  - separately encodes image and text
  - Text encoder same as BERT
- Image-grounded text encoder
  - injects visual information
  - one additional CA between SA and FF
- Image-grounded text decoder
  - generates captions given images
  - uses casual SA



Pre-training model architecture

BLIP: Multimodal mixture of Encoder-Decoder (MED)

- Image-Text Contrastive Loss (ITC)
  - aligns feature space of the visual transformer and text transformer
  - learns a similarity function

$$s(I,T) = g_v(\boldsymbol{v}_{\text{cls}})^\top g'_w(\boldsymbol{w}'_{\text{cls}}) \text{ and } s(T,I) = g_w(\boldsymbol{w}_{\text{cls}})^\top g'_v(\boldsymbol{v}'_{\text{cls}}).$$

$$p_m^{\text{i2t}}(I) = \frac{\exp(s(I,T_m)/\tau)}{\sum_{m=1}^{M} \exp(s(I,T_m)/\tau)}, \quad p_m^{\text{t2i}}(T) = \frac{\exp(s(T,I_m)/\tau)}{\sum_{m=1}^{M} \exp(s(T,I_m)/\tau)}$$

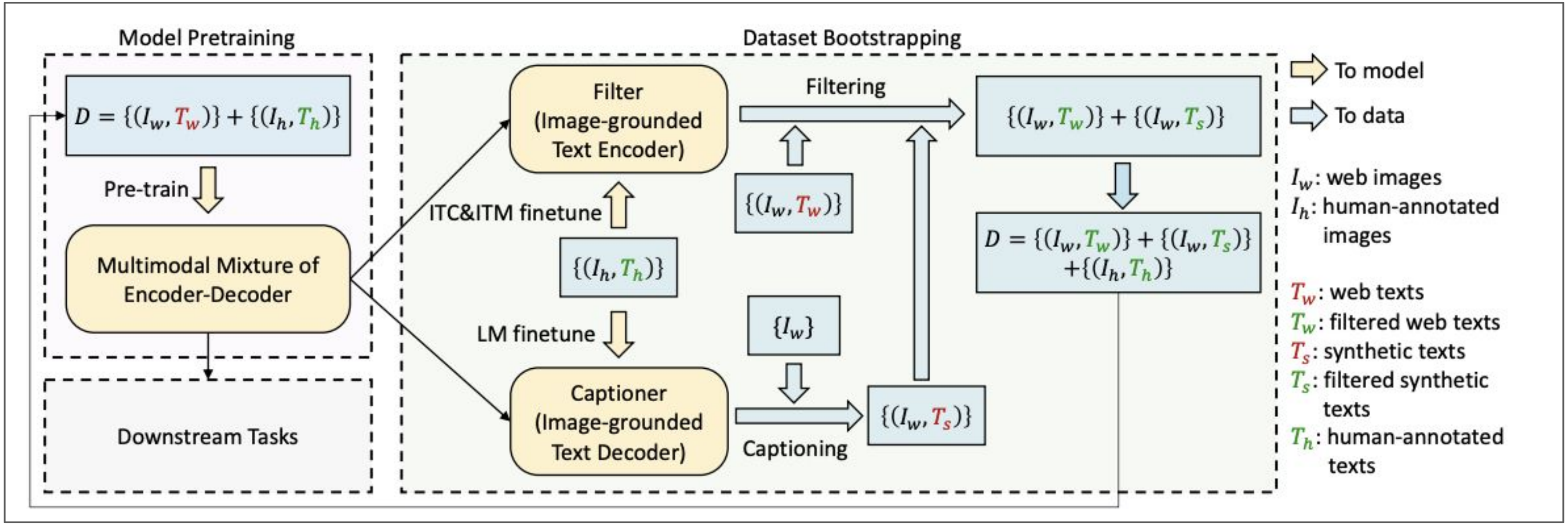
$$\mathcal{L}_{\text{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\left[\text{H}(\boldsymbol{y}^{\text{i2t}}(I), \boldsymbol{p}^{\text{i2t}}(I)) + \text{H}(\boldsymbol{y}^{\text{t2i}}(T), \boldsymbol{p}^{\text{t2i}}(T))\right]$$

- Image-Text Matching Loss (ITM)
  - learns image-text multimodal representation
  - binary classification task

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T)\sim D}\text{H}(\boldsymbol{y}^{\text{itm}}, \boldsymbol{p}^{\text{itm}}(I,T))$$

- Language Modeling Loss (LM)
  - cross entropy loss
  - aims to generate textual descriptions given an image
  - trains the model to maximize the likelihood of the text in an autoregressive manner

BLIP: Full schema



Effect of CapFilt

The efficacy of CapFilt on downstream tasks,
including image-text retrieval and image captioning

Pre-train dataset	Bootstrap		Vision backbone	Retrieval-FT (COCO)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
	C	F		TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	SPICE
COCO+VG +CC+SBU (14M imgs)	✗	✗	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✗	✓$_B$		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	✓$_B$	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	✓$_B$	✓$_B$		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	✗	✗	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	✓$_B$	✓$_B$		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	✓$_L$	✓$_L$		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	ViT-L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	✓$_L$	✓$_L$		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

Experiments

Method	Pre-train #Images	NoCaps validation in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Caption Karpathy test B@4	C
Enc-Dec (Changpinyo et al., 2021)	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL† (Zhang et al., 2021)	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
LEMON$_{\text{base}}$† (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
LEMON$_{\text{base}}$† (Hu et al., 2021)	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	**40.3**	**133.3**
BLIP	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
BLIP	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
BLIP$_{\text{CapFilt-L}}$	129M	**111.8**	**14.9**	**108.6**	**14.8**	**111.5**	**14.2**	**109.6**	**14.7**	39.7	**133.3**
LEMON$_{\text{large}}$† (Hu et al., 2021)	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7
SimVLM$_{\text{huge}}$ (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP$_{\text{ViT-L}}$	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

Image captioning

Method	Pre-train #Images	VQA test-dev	VQA test-std	NLVR^2 dev	NLVR^2 test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	-
ALBEF	14M	75.84	76.04	82.55	83.14
SimVLM$_{\text{base}}$†	1.8B	77.87	78.14	81.72	81.77
BLIP	14M	77.54	77.62	**82.67**	82.30
BLIP	129M	78.24	78.17	82.48	**83.08**
BLIP$_{\text{CapFilt-L}}$	129M	**78.25**	**78.32**	82.15	82.24

VQA and NLVR

Method	Pre-train # Images	COCO (5K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10
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
VILLA (Gan et al., 2020)	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR (Li et al., 2020)	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO (Li et al., 2021b)	5.7M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
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., 2021a)	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
BLIP	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	**100.0**	87.2	97.5	98.8
BLIP	129M	**81.9**	95.4	97.8	**64.3**	85.7	91.5	**97.3**	**99.9**	**100.0**	87.3	97.6	**98.9**
BLIP$_{\text{CapFilt-L}}$	129M	81.2	**95.7**	**97.9**	64.1	**85.8**	**91.6**	97.2	**99.9**	**100.0**	**87.5**	**97.7**	**98.9**
BLIP$_{\text{ViT-L}}$	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

Image-text retrieval

Method	MRR↑	R@1↑	R@5↑	R@10↑	MR↓
VD-BERT	67.44	54.02	83.96	92.33	3.53
VD-ViLBERT†	69.10	55.88	85.50	93.29	3.25
BLIP	**69.41**	**56.44**	**85.90**	**93.30**	**3.20**

Visual Dialog

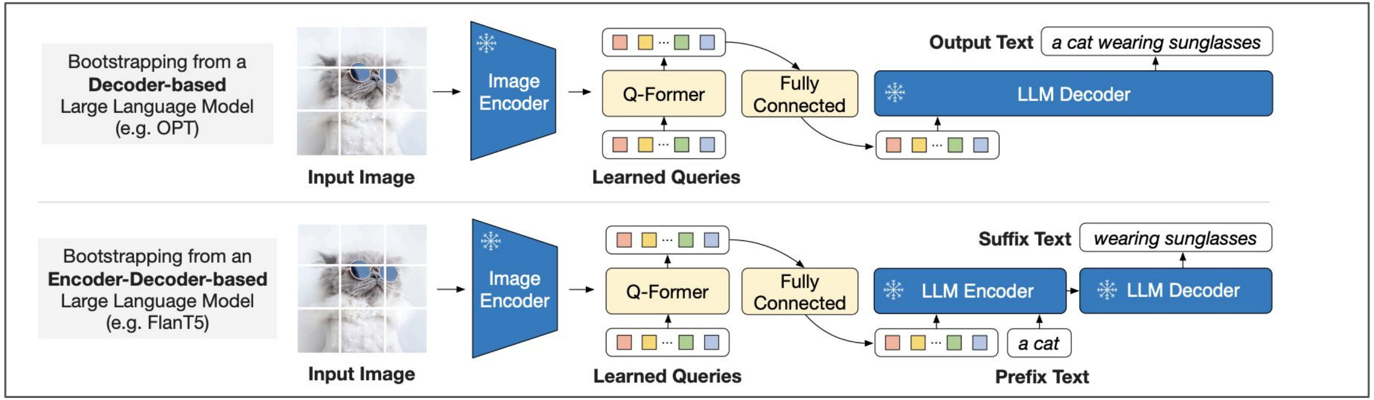
and even more

BLIP: potential directions for future

1. Multiple rounds of dataset bootstrapping
2. Generate multiple synthetic captions per image to further enlarge the pre-training corpus
3. Model ensemble by training multiple different captioners and filters and combining their forces in CapFilt

BLIP-2 (June 2023)

- Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models https://arxiv.org/pdf/2301.12597.pdf
- Novelties:
  - Q-Former - trainable module to connect a frozen image encoder and a frozen LLM
  - new objective in model pretraining: Image-grounded Text Generation
  - …



BLIP-2's second-stage vision-to-language generative pre-training

That’s all

Method	R1↑	R5↑	R10↑	MdR↓
zero-shot				
ActBERT (Zhu & Yang, 2020)	8.6	23.4	33.1	36
SupportSet (Patrick et al., 2021)	8.7	23.0	31.1	31
MIL-NCE (Miech et al., 2020)	9.9	24.0	32.4	29.5
VideoCLIP (Xu et al., 2021)	10.4	22.2	30.0	-
FiT (Bain et al., 2021)	18.7	39.5	51.6	10
BLIP	**43.3**	**65.6**	**74.7**	**2**
finetuning				
ClipBERT (Lei et al., 2021)	22.0	46.8	59.9	6
VideoCLIP (Xu et al., 2021)	30.9	55.4	66.8	-

Table 10. Comparisons with state-of-the-art methods for text-to-**video** retrieval on the 1k test split of the MSRVTT dataset.

Method	MSRVTT-QA	MSVD-QA
zero-shot		
VQA-T (Yang et al., 2021)	2.9	7.5
BLIP	19.2	35.2
finetuning		
HME (Fan et al., 2019)	33.0	33.7
HCRN (Le et al., 2020)	35.6	36.1
VQA-T (Yang et al., 2021)	41.5	46.3

Table 11. Comparisons with state-of-the-art methods for **video** question answering. We report top-1 test accuracy on two datasets.

Not contained examples and experiment tables