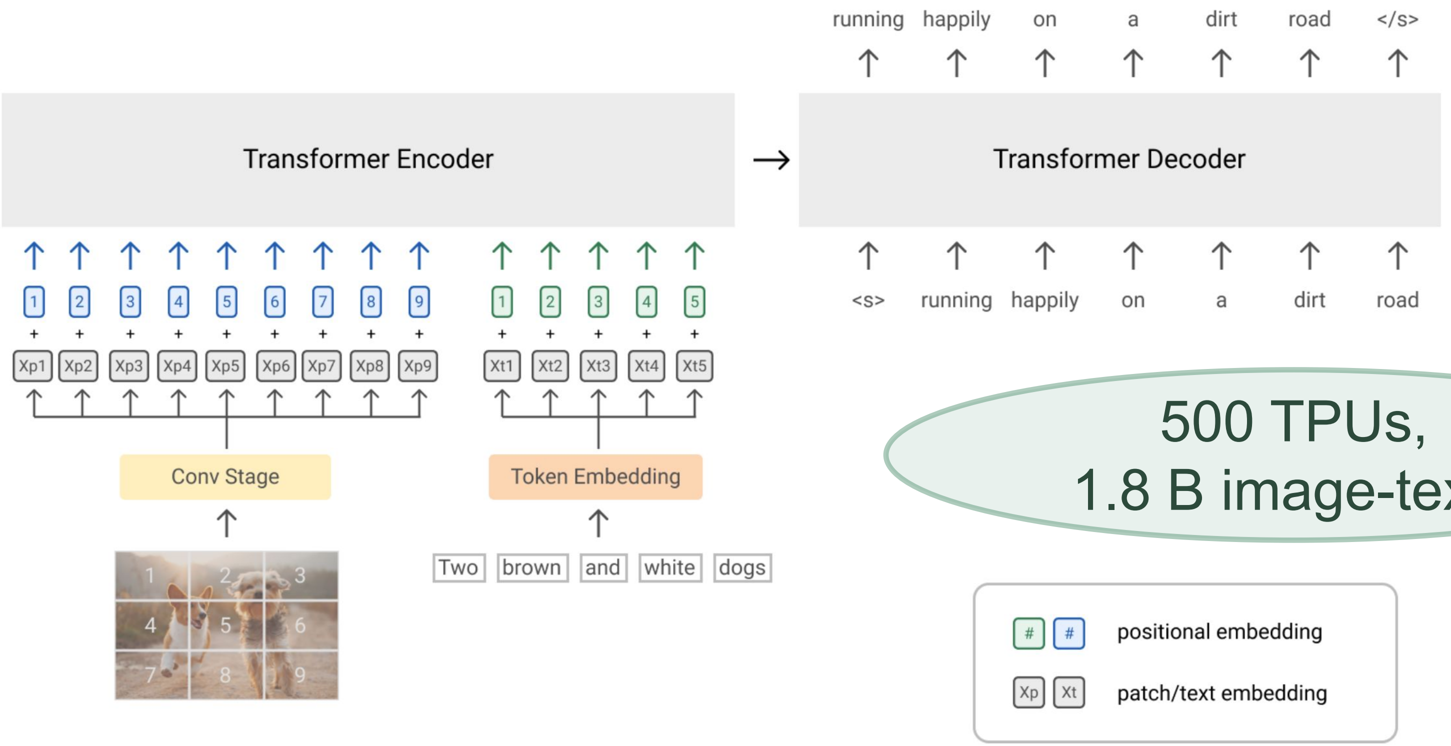
Expedited Training of Visual Conditioned Language Generation via Redundancy Reduction

Yiren Jian[1*], Tingkai Liu[2], Yunzhe Tao[2], Chunhui Zhang[1],,
Soroush Vosoughi[1], Hongxia Yang[2]

1. Dartmouth College 2. ByteDance Inc.
* This work is done during Yiren Jian's internship at ByteDance Inc.

Introduction

- Vision-language generative learning: a growth trajectory
  - SimVLM
  - CoCa
  - BLIP-2

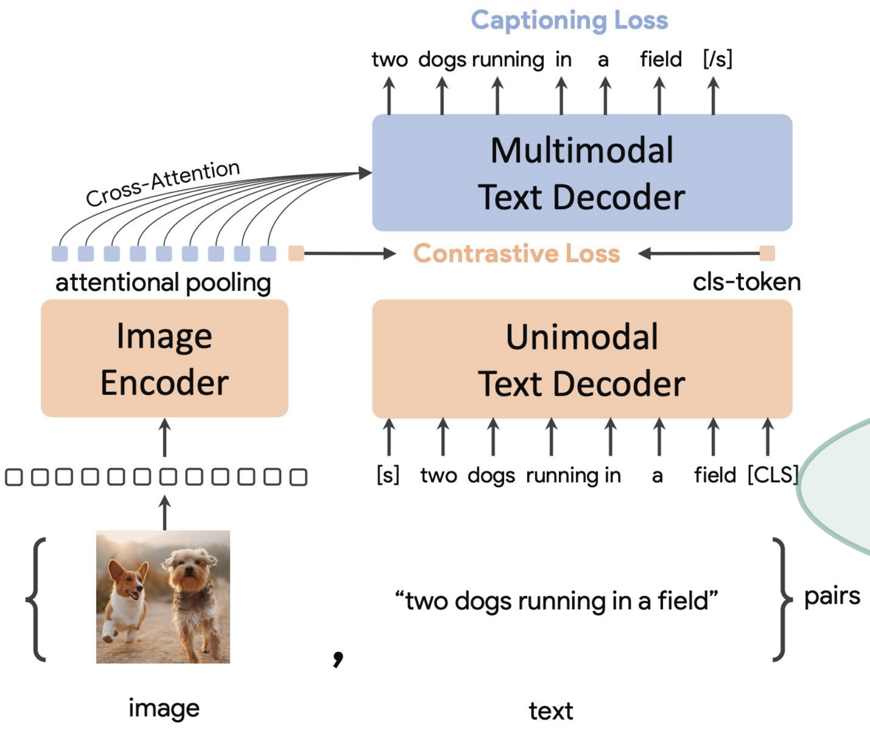


500 TPUs,
1.8 B image-texts

[1] Wang, Zirui, et al. "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision." ICLR 2022

Introduction

- Vision-language generative learning: a growth trajectory
  - SimVLM
  - CoCa
  - BLIP-2

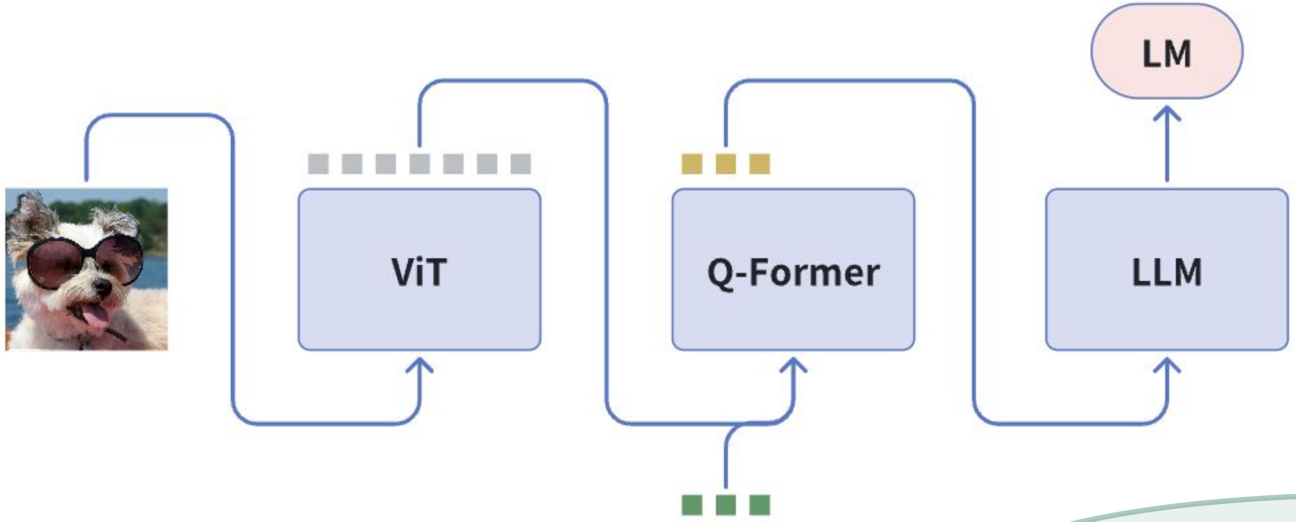


2 048 TPUs,
billion-scale data,
5 days

[2] Yu, Jiahui, et al. "Coca: Contrastive captioners are image-text foundation models." TMLR 2022

Introduction

- Vision-language generative learning: a growth trajectory
  - SimVLM
  - CoCa
  - BLIP-2



8 A100,
10 days

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

Introduction

- Training challenges when connecting *vision-language modalities*
  - SimVLM
  - CoCa

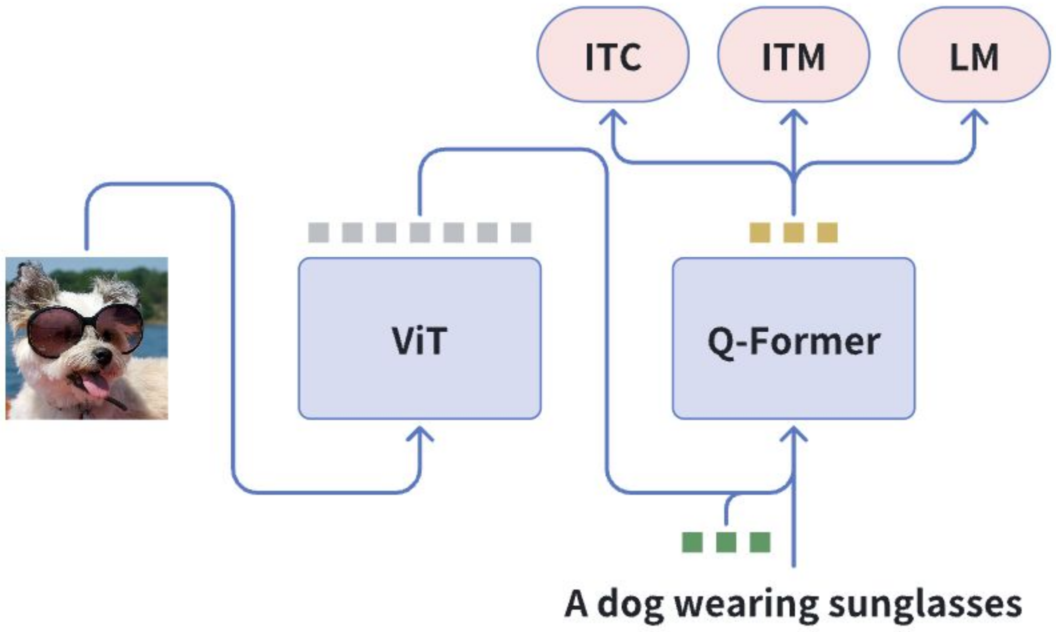
As pioneers, they try to connect vision-language modalities by ***training from scratch*** *on billion-scale image-text pairs.*

Training cost is the challenge!

- BLIP-2 → *Later, BLIP-2 applies existing well-pretrained ViT ① and LLM ②, then align the two backbones, via a novel connector ③ Q-former.*

Introduction

- A closer look on BLIP-2's Q-former: demanding an extra *stage-1* training



10 days on 8x A100 GPUs

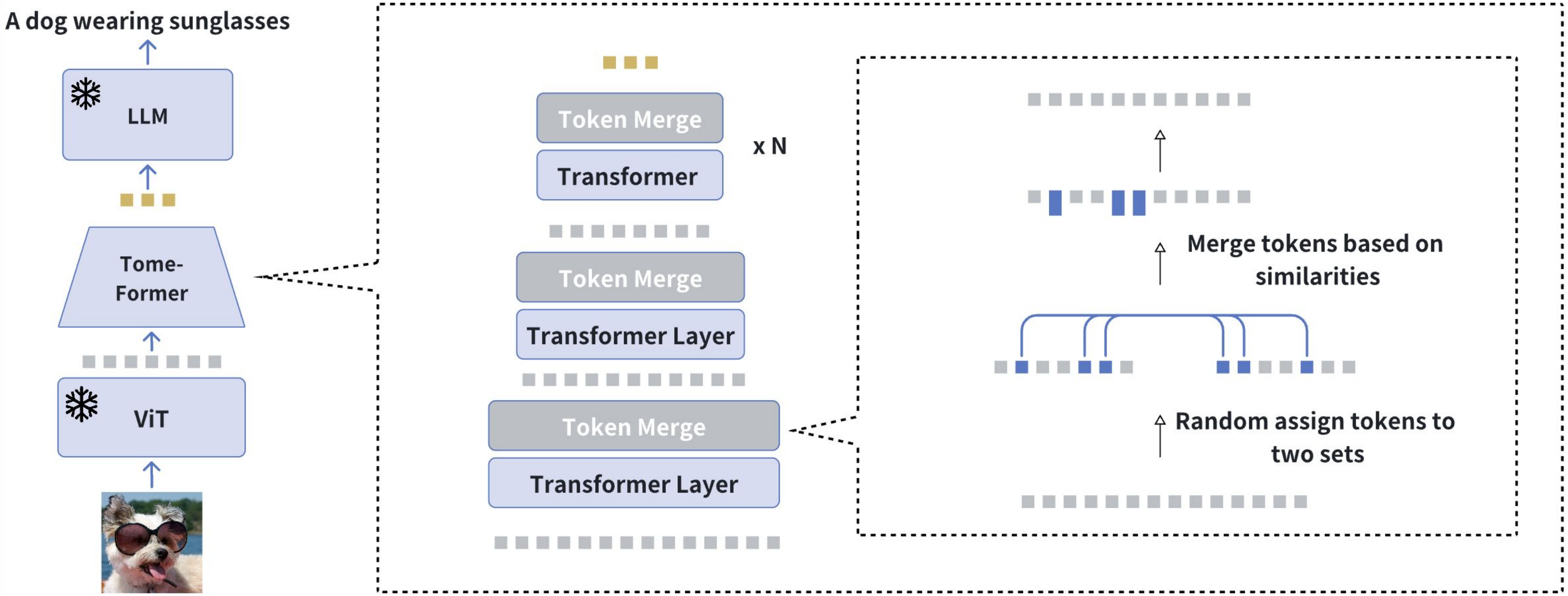
Introduction

- Our question:

how to replace Q-former for further efficiency?

EVLGen: *an end-to-end multimodal alignment*

- On images

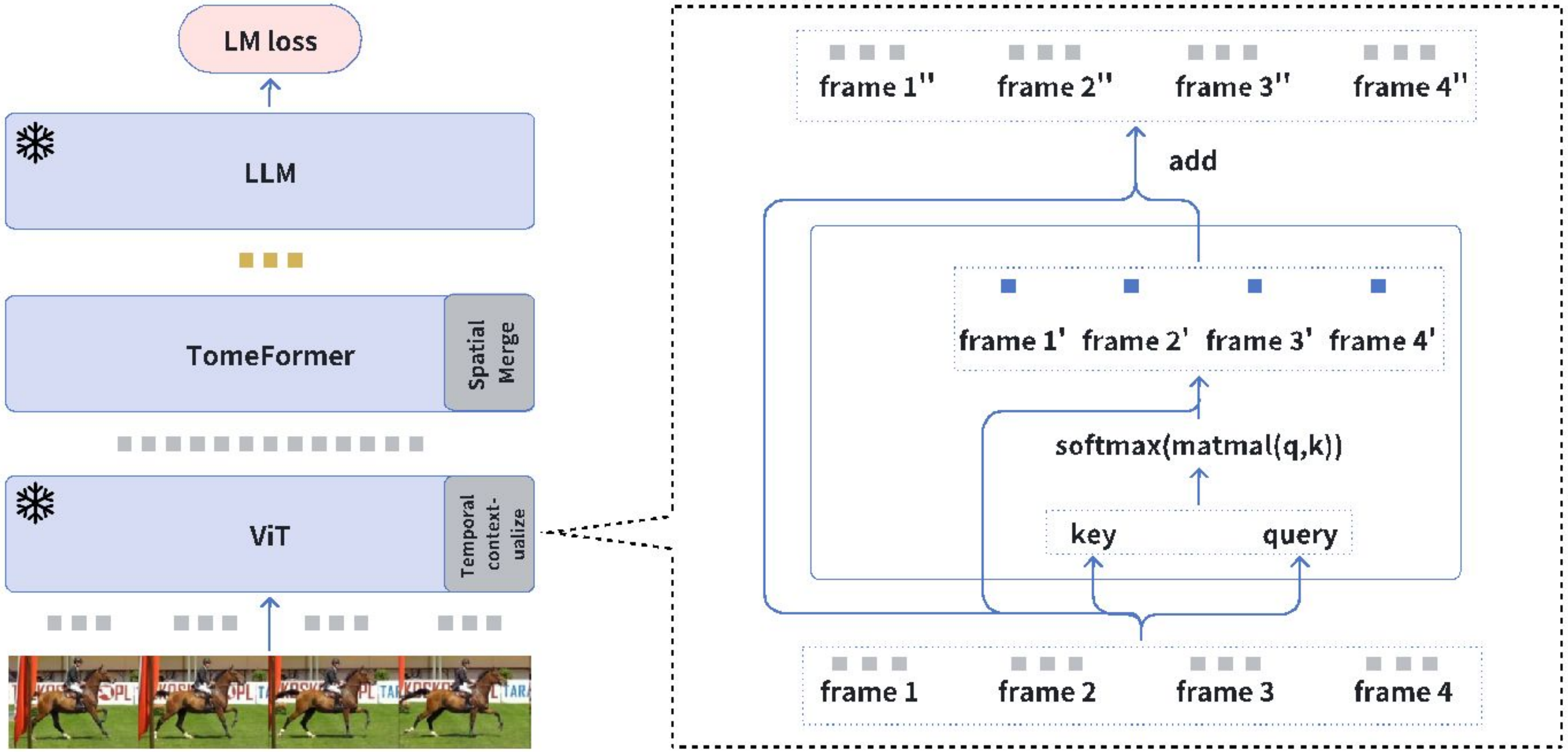


Token Merging [4] Transformer (TomeFormer) aggregates (cosine) similar visual tokens at each layer.

[4] Bolya, Daniel, et al. "Token Merging: Your ViT But Faster." ICLR 2023.

EVLGen: *an <u>end-to-end</u> multimodal alignment*

- On videos



For more spatial redundancy, temporal contextualize can pool multiple frames, then add back to each original frame.

EVLGen: *an <u>end-to-end</u> multimodal alignment*

- Summarization of EVLGen:
  - how it streamlines the pre-training?
    - Vision data (image, video…) is naturally redundant
    - Token-merging reduces learning space
    - the single-stage, single-loss training mechanism

Experiment

- An intuitive case study on token merging

Figure 4: Pre- and post-training visualization of merged tokens in E_2VL_{Gen}. The visual features compressed via token merging exhibit semantic informativeness even prior to training. This inherent characteristic facilitates E_2VL_{Gen}'s ability to converge quickly in an end-to-end training setup.

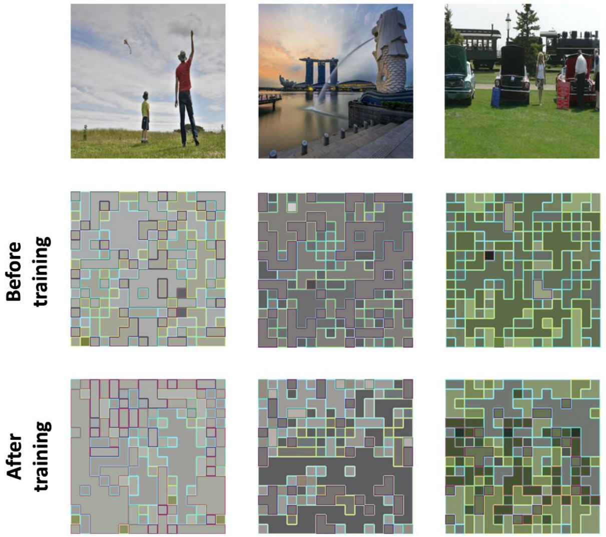


Figure 5: Additional pre- and post-training visualization of merged tokens in E_2VL_{Gen}.

Experiment (8× A100-80G)

- Overall Performance Comparison (1/2 image)

Models	# pre-train image-text	# trainable params	# stage-1 steps	# stage-2 steps	VQAv2 val	GQA test-dev	OK-VQA test	COCO val	Clock time
VL-T5	9.2M	224M	-	-	13.5	6.3	5.8	-	-
FewVLM	9.2M	740M	-	-	47.7	29.3	16.5	-	-
Frozen	3M	40M	-	-	29.6	-	5.9	-	-
VLKD	3M	406M	-	-	42.6	-	13.3	-	-
BLIP-2	104M†	110M+‡	-	80k/250k*	✗	✗	✗	✗	✗
BLIP-2	104M	110M+	250k	80k	44.6	30.6	26.0	137.7	234 hrs
EVL_{Gen}	104M	55M	-	90k	45.9	30.6	25.8	134.0	47 hrs
EVL_{Gen}	11M	110M	-	150k	46.3	30.0	23.0	135.1	80 hrs
EVL_{Gen}	104M	110M	-	150k	46.9	30.8	24.8	137.0	80 hrs
EVL_{Gen}	104M	110M	-	250k	**48.4**	**30.9**	**27.2**	**139.1**	133 hrs

Table 1: Comparison of methods on zero-shot VQA and MSCOCO captioning (CIDEr) tasks without additional fine-tuning. Both BLIP-2 and EVL_{Gen} use OPT-2.7b as the LLM decoder. *: *BLIP-2 without extensive stage-1 pre-training will collapse*. †: We were only able to download approximately 81% of LAION-115M and 78% of CCS-14M from the CapFilt dataset. ‡: BLIP-2 incorporates an additional set of 32 learnable queries, each with a dimension of 768.

Experiment

- Overall Performance Comparison (2/2 image)

	LLM	Model	C	B4	M	R
NoCaps	OPT	BLIP-2	112.2	44.4	29.5	59.7
		EVL_{Gen}	**117.4**	**45.9**	**30.3**	**61.1**
	Vicuna	BLIP-2	115.6	45.3	30.3	60.6
		EVL_{Gen}	**119.0**	**45.9**	**30.6**	**61.5**
Flickr30K	OPT	BLIP-2	77.1	28.7	23.9	51.6
		EVL_{Gen}	**82.0**	**30.0**	**24.5**	**52.4**
	Vicuna	BLIP-2	80.0	30.1	**24.8**	52.1
		EVL_{Gen}	**81.8**	**30.3**	24.5	**52.2**

Table 2: Comparison of different models' performance on zero-shot NoCaps and Flickr30K captioning. C→CIDEr, B4→BLEU-4, M→METEOR, R→ROUGE

Experiment

- Overall Performance Comparison (1/1 video)

Models	C	B4	M	R
Baseline (concat)	65.5	44.4	31.9	64.1
Baseline (mean)	67.8	47.3	32.2	65.0
$\mathrm{EVL}_{\mathrm{Gen}}$-image	68.4	47.6	32.4	65.3
$\mathrm{EVL}_{\mathrm{Gen}}$-video	69.8	48.3	32.6	65.8
$\mathrm{EVL}_{\mathrm{Gen}}$-video-scst	**74.0**	**49.2**	**33.0**	**66.5**
Video-LLaMA	59.3	47.7	29.6	63.7
VideoChat	58.0	46.5	29.5	63.4
VideoCoCa (open)	63.0	48.5	31.4	64.8

Table 3: Comparison of different models' performance on MSR-VTT video captioning. Models are pre-trained using 2 million video-text pairs from WebVid dataset, except for image pre-trained $\mathrm{EVL}_{\mathrm{Gen}}$-image.

Models	C	B4	M	R
Video-LLaMA	121.2	61.6	40.3	77.8
VideoChat	118.4	64.1	41.0	78.7
VideoCoCa (open)	150.9	67.7	45.3	81.9
$\mathrm{EVL}_{\mathrm{Gen}}$-video	**158.2**	**68.4**	**46.8**	**83.1**

Table 4: Comparison of different models' performance on MSVD video captioning.

Experiment (8× A100-80G)

- Training time comparison

Models	Stage 1 (MACs)	Stage 1 steps	Stage 2 (MACs)	Stage 2 steps
BLIP-2	36.7G	250k	6.28G	80k
EVL_{Gen}	-	-	11.9G	250k
EVL_{Gen}	-	-	11.9G	150k
EVL_{Gen55M}	-	-	5.6G	90k

Table 9: **Multiply–accumulate operations** (MACs) comparison of Q-Former (of BLIP-2) and TomeFormer (of EVL_{Gen}) when utilizing OPT-2.7b as the LLM.

Models	Stage 1 time /5k	Stage 2 time /5k	Clock time
BLIP-2	3 hrs 50 min	2 hrs 40 min	234 hrs
EVL_{Gen}	-	2 hrs 45 min	133 hrs
EVL_{Gen}	-	2 hrs 45 min	80 hrs
EVL_{Gen55M}	-	2 hrs 35 min	47 hrs

Table 10: Training time comparison of BLIP-2 and EVL_{Gen} when utilizing OPT-2.7b as the LLM.

1/3 to 1/6 of the training budget required by BLIP-2!

Experiment (8× A100-80G)

- How many tokens can be merged?

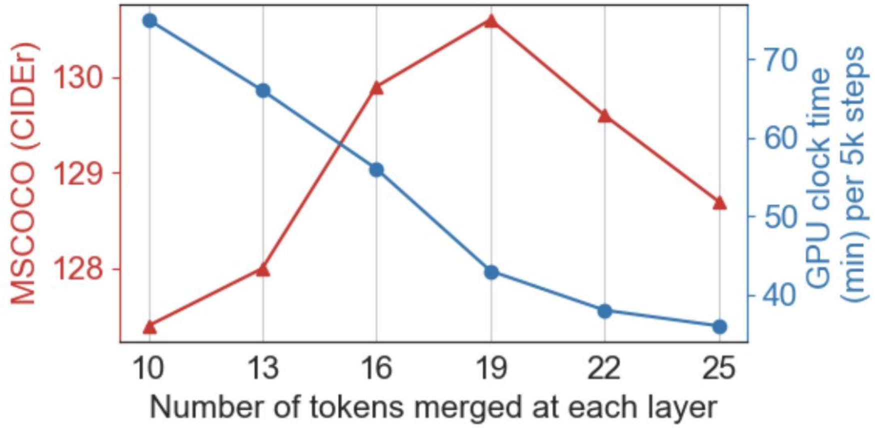


Figure 3: Trade-off between MSCOCO captioning scores (depicted in red) and GPU training time (depicted in blue) as a function of the number of tokens merged (r) in TomeFormer.

Thank you