

# Expedited Training of Visual Conditioned Language Generation via Redundancy Reduction

**Yiren Jian\***  
Dartmouth College

**Tingkai Liu**  
ByteDance Inc.

**Yunzhe Tao**  
ByteDance Inc.

**Chunhui Zhang**  
Dartmouth College

**Soroush Vosoughi**  
Dartmouth College

**Hongxia Yang**  
ByteDance Inc.

## Abstract

In this paper, we introduce  $EVL_{Gen}$ , a streamlined framework designed for the pre-training of visually conditioned language generation models with high computational demands, utilizing frozen pre-trained large language models (LLMs). The conventional approach in vision-language pre-training (VLP) typically involves a two-stage optimization process: an initial resource-intensive phase dedicated to general-purpose vision-language representation learning, focused on extracting and consolidating relevant visual features. This is followed by a subsequent phase that emphasizes end-to-end alignment between visual and linguistic modalities. Our novel one-stage, single-loss framework bypasses the computationally demanding first training stage by gradually merging similar visual tokens during training, while avoiding model collapse caused by single-stage training of BLIP-2 type models. The gradual merging process effectively condenses visual information while preserving semantic richness, resulting in rapid convergence without compromising performance. Our experimental findings demonstrate that our approach accelerates the training of vision-language models by a factor of 5 without a noticeable impact on overall performance. Furthermore, we illustrate that our models significantly narrow the performance gap to current vision-language models using only 1/10 of the data. Finally, we showcase how our image-text models can seamlessly adapt to video-conditioned language generation tasks through novel soft attentive temporal token contextualizing modules. Code is available at <https://github.com/yiren-jian/EVLGen>.

## 1 Introduction

The landscape of vision-language modeling has undergone significant transformations in recent years, with CLIP (Radford et al., 2021) serving as a landmark development. It distinguished itself through

\*This work was done during an internship at ByteDance Inc.

unparalleled zero-shot classification capabilities and efficiency in image-text retrieval tasks from prior works (Zhang et al., 2022a). Successive models like ALBEF (Li et al., 2021a), X-VLM (Zeng et al., 2022), and VLMO (Bao et al., 2022) further broadened the scope, addressing a myriad of tasks such as retrieval, visual entailment, and closed-set Visual Question Answering (VQA), among others.

Recently, the field has been enriched by the advent of generative models designed for complex image-to-language tasks. Notable contributions include CoCa (Yu et al., 2022), SimVLM (Wang et al., 2022c), Frozen (Tsimpoukelli et al., 2021), and Flamingo (Alayrac et al., 2022), targeting tasks like image and video captioning and open-set VQA. These models all rely on billion-scale datasets for training from scratch to bridge the substantial modality gap between vision and language.

As a result, the resource-intensive requirements (i.e., thousands of TPUs) of these training-from-scratch Vision-Language Models (VLMs) led to the conceptualization of BLIP-2 (Li et al., 2023a): this model alleviates computational costs (e.g., only requiring  $16\times$  fewer GPUs) by integrating existing well-pretrained vision encoders (ViT) with language decoders (LLM), and then tuning their joint operation. A central innovation in aligning vision and language modules in BLIP-2 is *Q-former*, a multimodal connector equipped with learnable queries for enhancing cross-attention mechanisms. This architectural choice, however, prevents the full model from end-to-end training and therefore *still* demands an additional pre-training regimen for *Q-former*, referred to as *BLIP-2's Stage 1*. The stage involves three learning objectives—image-text contrastive, image-text matching, and language generation—and necessitates multiple forward passes for facilitating the *Q-former's* optimization.

Despite its efficiency gains over CoCa, BLIP-2's training still imposes considerable computational costs. This poses challenges for research en-

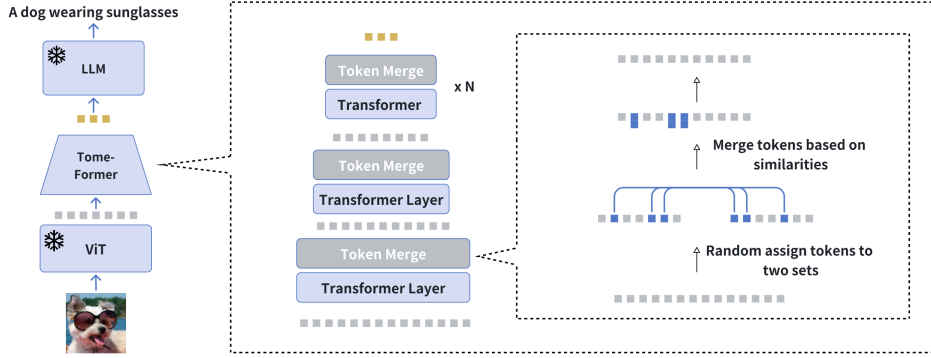


Figure 1: Overview of our  $EVL_{Gen}$ .  $EVL_{Gen}$  employs a streamlined, single-stage training mechanism with a unified loss. Here, visual tokens (in grey) are progressively aggregated based on their inherent similarities at each layer of the TomeFormer architecture. The final set of merged tokens (in orange) serves as semantically rich but computationally efficient soft prompts, guiding the LLM to generate a corresponding caption for the input image.

vironments with limited computational resources, such as university labs. Our experiments indicate that the Stage-1 training of BLIP-2 took approximately eight days on eight A100-80G GPUs (See Appendix B for training configurations). This computational burden has consequently restricted research to using the pre-trained Q-former, hindering the exploration of alternative ViTs in VLMs. This limitation is evident in subsequent works such as InstructBLIP (Dai et al., 2023b), VideoChat (Li et al., 2023b), Video-LLaMA (Zhang et al., 2023b), X-LLM (Chen et al., 2023a).

The prospect of reducing BLIP-2’s computational cost through end-to-end, single-stage training is compelling. Such an approach would remove the complexities associated with resource allocation and hyper-parameter tuning inherent in multi-stage training. Yet, direct end-to-end training with BLIP-2 poses substantial challenges, corroborated by both original findings from BLIP-2 and our own empirical analyses. We hypothesize that these challenges emanate from the intrinsic design of the Q-former. Specifically, the inclusion of randomly initialized learnable queries and cross-attention mechanisms complicates the optimization landscape, especially when the aim is to minimize the representational disparity between visual and linguistic modalities.

In this paper, we propose a token merging Transformer (TomeFormer) as an efficient vision-language connector. TomeFormer employs a systematic token-merging (Bolya et al., 2023) strategy that is both intuitive and effective. By connecting a pre-trained ViT as the visual encoder and a frozen LLM as the language decoder, we intro-

duce a new VLM “**Expedited Visual Language Generation model**” ( $EVL_{Gen}$ ), facilitates a streamlined, single-stage training process. It requires only a singular learning objective and a single forward pass per optimization step. This stands in contrast to BLIP-2’s multi-stage training, laden with multiple objectives and several forward passes.

Further, we introduce a *soft attentive temporal* contextualization mechanism within the ViT for effective video-language modeling. This uncovers more shared semantic features across temporal frames, thereby improving the efficiency of the spatial token merging process. It eliminates the need for modality realignment, contrasting approaches such as the temporal Q-former (Zhang et al., 2023b), or the addition of new learnable temporal queries (Li et al., 2023b). Our strategy simplifies the optimization challenges tied to working with relatively smaller video-text datasets, compared to their image-text counterparts. Remarkably, we demonstrate that even without video pre-training, our temporal token contextualize approach can effectively train robust video-language models. This differs from recent work in video-language models (Yan et al., 2022; Chen et al., 2023b) that depend on pre-training models using vast million-scale video-text datasets. In summary, our contributions are:

- For reducing vision redundancy within the vision language connector, we adopt Token Merging, initially designed to enhance ViT inference speed without training. Concurrently, we present a novel temporal token contextualization scheme for video modeling.
- Our proposed VLM featuring TomeFormer com-

petes effectively with BLIP-2, while requiring just a fraction of the computational resources. Given the reliance on BLIP-2’s pre-trained model in contemporary studies, our approach widens the exploratory scope for various ViTs.

- We introduce a straightforward spatial attentive temporal modeling technique that allows for the seamless adaptation of pre-trained image-text models to video tasks. This approach eliminates the need for complex modality re-alignment, a common requirement in alternative methods.

## 2 Related Work

**Image-Language Models** CoCa (Yu et al., 2022), trained on billions of image-text pairs, represents a state-of-the-art approach in generative tasks like open VQA and visual captioning. To mitigate the computational demands of pre-training, BLIP-2 (Li et al., 2023a) employs frozen pre-trained ViT and LLM components, focusing on training a specialized connector between visual and linguistic modalities called the Q-former. Due to the computationally intensive nature of training BLIP-2, subsequent models in visual instruction (Dai et al., 2023b; Zhu et al., 2023; Li et al., 2023b) have predominantly utilized the pre-trained Q-former, which is aligned with the *eva-vit-g* model supplied by BLIP-2. Additional related works on image-language modeling are further discussed in Appendix A.

**Video-Language Models** While many image-text models can be adapted for video-text tasks through simple feature pooling (e.g., VideoCoCa (Yan et al., 2022)), the field has seen specialized models that incorporate temporal dynamics. Building on the foundation of BLIP-2, VideoLLaMA (Zhang et al., 2023b) enhances its architecture by introducing additional temporal Q-former layers between the spatial Q-former and the LLM components of BLIP-2. Inspired by BLIP-2, most recent works such as VideoChat (Li et al., 2023b), PandaGPT (Su et al., 2023), Valley (Luo et al., 2023), and Video-ChatGPT (Muhammad Maaz and Khan, 2023) leverage frozen LLMs in their video-language models.

**Token Merging** Token Merging (ToMe) (Bolya et al., 2023) aims to improve the inference speed of pre-trained ViTs without requiring re-training. At each Transformer layer, tokens are divided into two sets and subsequently merged based on similarity,

effectively reducing the token count and thereby accelerating inference. This method maintains classification and generation quality.

In our work, we repurpose ToMe to condense the visual features used as language prompts in the LLM. We integrate a standard Transformer with ToMe capabilities, resulting in a model we term TomeFormer. This model serves as an effective connector between visual and language domains, preserving semantic richness while reducing token count. Importantly, this integration of ToMe does not introduce any additional parameters. Inspired by spatial ToMe, we introduce a novel soft temporal ToMe variant within the vision encoder, thereby adding temporal modeling capabilities to our image-text models. Additional related works on token redundancy are further discussed in Appendix A.

## 3 Methods

We begin by presenting our image-text model and then describe the adaptations made to this pre-trained model for video-related tasks.

### 3.1 Preliminary

We follow BLIP-2’s efficient training paradigm, i.e., utilizing frozen but well-pretrained ViTs and LLMs while *solely training a vision-to-language connector*. However, BLIP-2 still remains a challenge, since it necessitates an extra Stage-1 as a pre-training phase for the unstable Q-former (i.e., the vision-to-language connector), before the final end-to-end fine-tuning.

Our observation underscores the pivot role of BLIP-2 Stage-1 pre-training (which takes approximately 8 days on eight A100 GPUs): without it, the BLIP-2 model collapses, as evidenced in Table 1. To avoid this extra stage-1, we replace Q-former with a novel vision-to-language connector, which is designed to discover vision redundancy and then significantly accelerate visual-language alignment, often resulting in enhanced performance.

### 3.2 EVL<sub>Gen</sub>-Image

We introduce EVL<sub>Gen</sub>-Image (abbreviated as EVL<sub>Gen</sub>, shown in Figure 1), an optimized vision-language generative pre-training model. EVL<sub>Gen</sub> utilizes a ViT for visual encoding and an LLM for linguistic decoding. The key innovation is the incorporation of a standard Transformer, augmented with spatial Token Merging, to act as the connector between the visual and linguistic modalities.

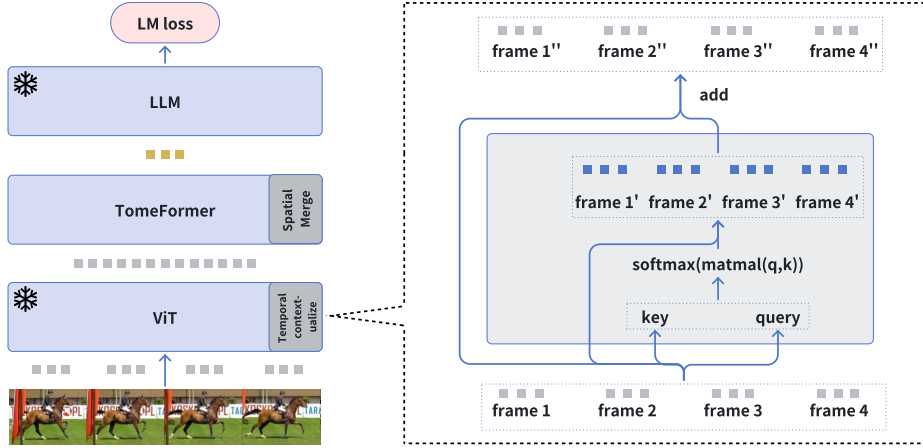


Figure 2: Overview of  $EVL_{Gen}\text{-Video}$ : In addition to TomeFormer’s spatial token merging capabilities, our design introduces Temporal Attentive Soft Token Contextualizing for nuanced temporal modeling. Each frame’s output is calculated as a learnable weighted average of other frames in the video. This approach maintains the integration of pre-existing, well-trained image-text models. For instance, when the input consists of static videos with identical frames,  $EVL_{Gen}\text{-Video}$  operates as if it were an image-text model. Importantly, this architecture avoids the need for complex modality realignment, a requirement in alternative designs that insert a temporal Q-former between the visual encoder and the language model. It also significantly enriches the shared semantic information distributed among these frame tokens, laying the groundwork for more efficient token merging in future spatial merging steps.

Formally, our framework includes a vision encoder  $E_{\text{vision}}$ , which ingests an input image  $I$  and encodes it into a fixed set of visual tokens:  $[v_1, v_2, \dots, v_L] = E_{\text{vision}}(I)$ . Here,  $L$  denotes the number of image patches. Subsequently, we employ a Transformer equipped with token-merging modules (further technical details are provided in Appendix C), termed as *TomeFormer* ( $T_{v \rightarrow 1}$ ) as the vision-to-language connector. This module effectively compresses the token count:

$$[v'_1, v'_2, \dots, v'_{L'}] = T_{v \rightarrow 1}(f_{\text{proj}_1}([v_1, v_2, \dots, v_L])). \quad (1)$$

In this equation,  $L'$  is considerably smaller than the initial token count  $L^1$ . The LLM decoder then employs these compressed tokens as soft prompts for text generation:

$$\text{output} = D_{\text{LLM}}(f_{\text{proj}_2}([v'_1, v'_2, \dots, v'_{L'}])). \quad (2)$$

Projection functions  $f_{\text{proj}_1}$  and  $f_{\text{proj}_2}$  are used to ensure dimension compatibility. The training objective is to minimize the cross-entropy between the output and ground truth caption:

$$\mathcal{L} = \text{CrossEntropyLoss}(\text{output}, \text{cap}_{\text{gt}}). \quad (3)$$

Three main advantages of using TomeFormer are:

<sup>1</sup>We merge a fixed number of tokens at each layer of the TomeFormer. Finally, 256 visual tokens are reduced to 28 tokens. Ablation on merged tokens at each layer is studied in Section 5.

- Efficient token reduction, facilitating the transformation of loosely structured visual data into a more concise yet informative representation.
- Computational efficiency, as the uncompressed ViT output consists of 256 tokens, plus a [CLS] token. Without compression, the subsequent vision-to-language connector would be computationally expensive in terms of both memory and processing power.
- Semantic richness of the compressed tokens. Unlike BLIP-2, which requires an extensive pre-training phase for Q-former, TomeFormer naturally merges semantically similar tokens. Our empirical evidence confirms that TomeFormer-equipped models train more efficiently compared to alternatives like BLIP-2.

### 3.3 $EVL_{Gen}\text{-Video}$

Although many image-text models can be adapted for video-text tasks with minor modifications, such adaptations either result in inadequate temporal modeling (as in VideoCoCa or InstructBLIP) or require re-alignment with substantial video-text pairs due to additional learnable Q-formers (as in VideoChat and Video-LLaMA).

In this paper, we propose a novel module called *Temporal Attentive Soft Token Contextualizing* to enhance the ViT backbone with temporal modeling capabilities (depicted in Figure 2). A key feature

of temporal soft contextualizing is that it is equivalent to the identity operator when the input is static images instead of videos. Thus, our approach maintains the integration of pre-existing, well-trained image-text models, thus avoiding the additional need for modality realignment, a requirement in alternative designs that insert a temporal Q-former between the visual encoder and LLM.

Formally, let  $v$  be a video feature tensor with dimensions  $[B \times N \times L \times D]$ , where  $B$  is the batch size,  $N$  is the number of frames,  $L$  is the sequence length (i.e., the number of patches in a single video frame), and  $D$  is the hidden dimension. Initially, we reshape  $v$  into  $[(B \times N) \times L \times D]$  which is subsequently fed into the self-attention layer of the ViT for *spatial modeling* as:

$$v' = \text{self-attn}(v.\text{reshape}(B \times N, L, D)). \quad (4)$$

For *temporal modeling*,  $v'$  is reshaped to  $[(B \times L), N, D]$ . We then project this into key and query matrices  $k$  and  $q$  and compute  $v''$  using our *Temporal Attentive Soft Token Contextualizing* as follows:

$$k = W_{\text{key}}(v'.\text{reshape}(B \times L, N, D)), \quad (5)$$

$$q = W_{\text{query}}(v'.\text{reshape}(B \times L, N, D)), \quad (6)$$

$$v'' = v' + \text{softmax}(\text{matmul}(q, k)) \cdot v'. \quad (7)$$

The softmax operation models temporal weights and *softly* fuses tokens among multiple frames. This is distinct from spatial token merging, which employs average pooling and reduces the token count. Here, the weighted average pooling is applied to multiple frames for contextualization. It preserves the original count of tokens while enhancing the shared semantic content that is spread across various frames. Therefore, it allows a higher rate of token merging in the subsequent spatial merging processes.

## 4 Experiments

Our experimental setup is as follows:

- **Pre-training Data** Our model is pre-trained using the MSCOCO (Lin et al., 2014) and CapFilt (Li et al., 2022) datasets, which include BLIP’s pseudo-labeled Conceptual Captioning (Sharma et al., 2018), SBU (Ordonez et al., 2011), and LAION (Schuhmann et al., 2022) datasets—similar to the data sources utilized in BLIP-2. Note that we intentionally exclude the VG (Krishna et al., 2017) dataset from our pre-training procedure, as it mainly consists of localized captions.

- **Models** To facilitate a direct and fair comparison with BLIP-2, we employ the same ViT, textttteva-vit-g (Fang et al., 2023). For the language model decoders, we explore both opt-2.7b (Zhang et al., 2022b) and vicuna-7b (Chiang et al., 2023). Our TomeFormer is initialized using bert-base-uncased, ensuring parameter count parity with BLIP-2’s Q-former.
- **Pre-training Details** Our pre-training setup closely mirrors the configurations of BLIP-2. We utilize a maximum learning rate of  $1e^{-4}$  and a minimum learning rate of  $1e^{-5}$ . The learning rate follows a schedule that begins with a linear warm-up phase of 5000 steps starting from  $1e^{-6}$  and then transitions to a cosine decay schedule. Weight decay is set at 0.05. The training is conducted with a batch size of 1600, distributed over either  $8 \times \text{A100-80G}$  or  $32 \times \text{V100-32G}$ .
- **Downstream Tasks** EVL<sub>Gen</sub>-Image is evaluated without additional fine-tuning on a variety of tasks, including MSCOCO captioning, VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), and OK-VQA (Marino et al., 2019). For video tasks, EVL<sub>Gen</sub>-Video is evaluated on fine-tuned MSR-VTT (Xu et al., 2016) and MSVD (Chen and Dolan, 2011) captioning. We use the standard train/val/test splits.

### 4.1 Evaluation on Image-Text Benchmarks

We conducted comparative evaluations between EVL<sub>Gen</sub> and BLIP-2 on multiple image-text benchmarks, including zero-shot VQAv2, GQA, OK-VQA, and MSCOCO captioning. It is essential to note that BLIP-2 demands an extensive Stage-1 pre-training phase involving 250,000 optimization steps. This phase incorporates three distinct loss functions and necessitates multiple forward passes through the model, a process crucial for BLIP-2 to prevent model divergence.

Table 1 summarizes the results of our experiments. Our primary insights can be distilled into the following key points:

- Utilizing the same training set of 104M image-text pairs and an equal number of optimization steps (250K), EVL<sub>Gen</sub> consistently outperforms BLIP-2 across nearly all evaluated tasks.
- Remarkably, EVL<sub>Gen</sub> maintains competitive performance even when its training budget is trimmed to approximately one-third of BLIP-2’s, specifically 150K optimization steps.
- Our experiments show that EVL<sub>Gen</sub> can produce satisfactory results with a significantly reduced

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 <sup>†</sup>	110M+ <sup>‡</sup>	-	80k/250k*	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>
BLIP-2	104M	110M+	250k	80k	44.6	30.6	26.0	137.7	234 hrs
EVL <sub>Gen</sub>	104M	55M	-	90k	45.9	30.6	25.8	134.0	47 hrs
EVL <sub>Gen</sub>	11M <sup>†</sup>	110M	-	150k	46.3	30.0	23.0	135.1	80 hrs
EVL <sub>Gen</sub>	104M	110M	-	150k	46.9	30.8	24.8	137.0	80 hrs
EVL <sub>Gen</sub>	104M	110M	-	250k	<b>48.4</b>	<b>30.9</b>	<b>27.2</b>	<b>139.1</b>	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<sub>Gen</sub> 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 (110M) and 78% of CCS-14M (11M) from the CapFilt dataset. ‡: BLIP-2 incorporates an additional set of 32 learnable queries, each with a dimension of 768.

training dataset of 11 million image-text pairs, while still undergoing 150K optimization steps.

- EVL<sub>Gen</sub> retains its efficacy even when the training budget is restricted to as few as 90K steps, showing the model’s efficiency and robustness.

We further evaluate BLIP-2 and EVL<sub>Gen</sub> on zero-shot NoCaps and Flickr30K datasets. Shown in Table 2, EVL<sub>Gen</sub> consistently outperforms BLIP-2 in both datasets using different LLMs.

**Training Time** In the Stage-1 pre-training phase, BLIP-2 requires considerable time, necessitating multiple forward passes to optimize three separate loss functions. We document the training durations for both BLIP-2 and EVL<sub>Gen</sub> when utilizing eight A100-80G GPUs in the last column of Table 1.

Although BLIP-2 significantly reduces training time relative to predecessors like CoCa, it still mandates an extended training duration, approximately ten days (8 days for stage 1 and 2 days for stage 2). This extensive time commitment limits the feasibility of researchers to investigate various ViT configurations. Most subsequent works based on BLIP-2 continue to use the pre-trained Q-former in conjunction with the *eva-vit-g* model, thereby narrowing the scope of ViT exploration. In contrast, EVL<sub>Gen</sub> significantly trims the training time while maintaining satisfactory performances, thus providing researchers with the latitude to explore a wider array of advanced ViTs in future investigations.

Furthermore, MACs (FLOPs) in Q-Former and TomeFormer is discussed in Section 5.5.

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

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

## 4.2 Evaluation of EVL<sub>Gen</sub>-Video

We proceed to evaluate the performance of fine-tuned EVL<sub>Gen</sub>-Video models in video captioning tasks, utilizing OPT-2.7b as the language model decoder. Our investigation includes two specific variants of EVL<sub>Gen</sub>-Video: the first is exclusively pre-trained on image data, while the second is further *enhanced by pre-training on a corpus of 2 million video-text pairs sourced from the WebVid (Bain et al., 2021) dataset*. To provide a comprehensive evaluation, we benchmark EVL<sub>Gen</sub>-Video against five distinct models, described as follows:

- **Baseline (concat)**: This model processes each frame of a video individually and concatenates their visual features to generate a single prompt for the LLM. This method is analogous to the strategy employed in InstructBLIP.
- **Baseline (mean)**: Similar to the concat baseline, this model processes each video frame individually but averages the visual features to create a

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
EVL <sub>Gen</sub> -image	68.4	47.6	32.4	65.3
EVL <sub>Gen</sub> -video	69.8	48.3	32.6	65.8
EVL <sub>Gen</sub> -video-scst	<b>74.0</b>	<b>49.2</b>	<b>33.0</b>	<b>66.5</b>
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 EVL<sub>Gen</sub>-image.

single prompt for the LLM.

- **Video-LLaMA**: This variant incorporates the BLIP-2 framework and enhances it with an additional temporal Q-former layer. For this evaluation, we focus solely on the vision-language component of Video-LLaMA.
- **VideoChat**: This model extends BLIP-2 by integrating additional Uniformer modules within the ViT architecture and also incorporates learnable temporal queries in its Q-former component.
- **VideoCoCa**: In this model, we adapt the OpenCoCa framework by mlfoundations and augment the existing CoCa architecture with a learnable attentional pooler, resulting in VideoCoCa.

**Evaluation on MSR-VTT** As detailed in Table 3, EVL<sub>Gen</sub>-Video demonstrates superior performance relative to the baseline models, even without the aid of video-text pre-training. This result highlights the effectiveness of our proposed *Temporal Attentive Soft Token Contextualizing* in capturing temporal dynamics. Additionally, we observe an enhancement in performance when incorporating video-text pre-training along with Self-Critical Sequence Training (SCST) (Rennie et al., 2017).

*Temporal Attentive Soft Token Contextualizing* has the distinct advantage of maintaining the integration of the well-pretrained image-text model (i.e., EVL<sub>Gen</sub>-Image). This contrasts with models such as Video-LLaMA and VideoChat, where the original BLIP-2 architecture is altered, necessitating a complex re-alignment process using video-text pairs. Our empirical analysis indicates that such re-alignment is a non-trivial endeavor (as shown in Table 3, Video-LLaMA and VideoChat struggle to re-align with 2M WebVid video-text pairs). It is worth noting that our VideoCoCa model is at a disadvantage when benchmarked against

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
EVL <sub>Gen</sub> -video	<b>158.2</b>	<b>68.4</b>	<b>46.8</b>	<b>83.1</b>

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

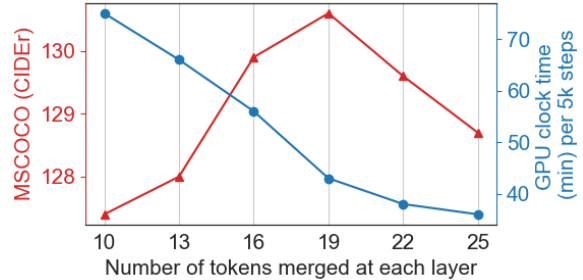


Figure 3: Trade-off between MSCOCO captioning scores (depicted in red) and GPU training time (depicted in blue) as a function of the number of tokens merged ( $r$ ) in TomeFormer.

Google’s reported results, which benefit from extensive training on a much larger billion-scale dataset.

**Evaluation on MSVD** Similarly, we evaluate EVL<sub>Gen</sub>’s performance against Video-LLaMA, VideoChat, and VideoCoCa using the MSVD caption dataset (presented in Table 4). Our results corroborate that EVL<sub>Gen</sub> consistently surpasses these competing models, further attesting to its robust performance across different video caption tasks.

## 5 Ablations and Analysis

### 5.1 Ablations on TomeFormer

Within the TomeFormer, the vision-to-language connector in EVL<sub>Gen</sub>, we introduce a hyperparameter  $r$  that regulates the number of spatial tokens merged at each layer. Increasing  $r$  substantially reduces the token count, but runs the risk of eliminating important visual details. On the other hand, a smaller  $r$  produces two main effects: (1) a more diffuse representation of visual features, complicating the optimization landscape, and (2) elongated soft prompts for the LLM, leading to increased computational cost during training, such as memory overflow and extended training duration.

To study the effects of  $r$ , we conduct an ablation experiment using  $8 \times$  RTX-A6000 and the CCS-14M dataset for pre-training. The models are trained for 60,000 steps, and their performance is evaluated using CIDEr scores on MSCOCO cap-



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

tioning. In Figure 3, we observe that a smaller  $r$  (e.g., 10) places a higher computational load on both TomeFormer and the LLM, extending training time and compromising optimization, as evidenced by lower CIDEr scores. In contrast, a larger  $r$  value (e.g., 25) expedites training but at the expense of model performance, likely due to excessive feature compression and consequent information loss. Additional ablation results on VQAv2, GQA and OKVQA are provided in Appendix G.

## 5.2 Ablations on ViT

Experimental results on  $EVL_{Gen}$  with different visual encoders (ViT) are provided in Appendix F.  $EVL_{Gen}$  is robust to different visual encoders, and the stronger ViT generally leads to better results. This implies that while  $EVL_{Gen}$  also requires re-training for different ViTs, but the single-stage training and quick convergence allow it to benefit from a future release of the latest ViTs, given its capability of fast adaptation.

## 5.3 Vicuna-7b as the LLM

In Table 5, we provide experimental results of BLIP-2 and ours using Vicuna-7b as the LLM decoder, on zero-shot VQAv2, GQA, OKVQA and MSCOCO captioning (CIDEr) tasks without additional fine-tuning.

$EVL_{Gen}$  achieves better performance in GQA, OKVQA, and MSCOCO captioning using considerably less computing, though our model under-

Models	Data	steps	VQA	GQA	OK	COCO
BLIP-2	104M	330k	<b>57.8</b>	35.7	27.8	138.0
$EVL_{Gen}$	104M	90k	53.4	34.7	30.6	137.8
$EVL_{Gen}$	11M	150k	54.6	34.0	27.3	138.0
$EVL_{Gen}$	104M	150k	55.5	<b>36.3</b>	<b>30.6</b>	137.9
$EVL_{Gen}$	104M	250k	54.8	35.6	30.4	<b>139.1</b>

Table 5: Comparison of different models' performance on zero-shot VQA and MSCOCO captioning (CIDEr) tasks without additional fine-tuning. Both BLIP-2 and  $EVL_{Gen}$  use Vicuna-7b as the LLM decoder.

performs in VQAv2. As we discussed in Limitations, the inability of  $EVL_{Gen}$  to extract question-conditioned visual features may lead to inferior results on VQAv2.

## 5.4 Token Merging Visualization in $EVL_{Gen}$

One notable advantage of  $EVL_{Gen}$  over BLIP-2 is the absence of a requisite Stage-1 pre-training for the vision-to-language connector. This simplifies the training pipeline by removing the need to train the model to extract text-informative visual features. We posit that the token merging process in TomeFormer naturally aggregates tokens associated with visually similar elements, thereby yielding concise yet semantically rich visual features from the onset of training. This inherent capability allows  $EVL_{Gen}$  to benefit from a more streamlined, single-stage training regimen with just one learning objective.

Essentially, our token merging strategy serves as an efficient approximation of Q-Former's functionality, compressing visual features in a semantically meaningful manner. Figure 4 illustrates this, displaying the visual tokens before and after training with our TomeFormer. The figure shows that the compressed visual features obtained via token merging are semantically informative and offer basic object segmentation within the image. Furthermore, the semantic coherence of these merged tokens improves as training advances. Additional visualization examples are shown in Appendix E.

## 5.5 MACs (FLOPs) in Q-Former and TomeFormer

In this section, we compute **multiply-accumulate operations** (MACs) in Q-Former and TomeFormer. MACs performs  $a \leftarrow a + (b \times c)$ . Whereas, FLOPs is **floating operations** which includes  $\times / + / \div \dots$  etc. One MACs has one  $\times$  and one  $+$ . And thus, roughly speaking, FLOPs is two times as MACs.

In our experiments, BLIP-2 and  $EVL_{Gen}$  have identical ViTs and LM decoders. Thus, we only



Models	Stage 1 (MACs)	Stage 1 steps	Stage 2 (MACs)	Stage 2 steps
BLIP-2	36.7G	250k	6.28G	80k
EVL <sub>Gen</sub>	-	-	11.9G	250k
EVL <sub>Gen</sub>	-	-	11.9G	150k
EVL <sub>Gen55M</sub>	-	-	5.6G	90k

Table 6: **Multiply-accumulate operations** (MACs) comparison of Q-Former (of BLIP-2) and TomeFormer (of EVL<sub>Gen</sub>) when utilizing OPT-2.7b as the LLM.

compare the MACs in VL Connector in BLIP-2 and EVL<sub>Gen</sub> (i.e., Q-Former and TomeFormer).

There’s a large MACs in BLIP-2 stage-1 due to three forward passes using Q-Former, where the last forward-pass used for caption loss dominates (27.0G). In contrast, EVL<sub>Gen</sub> does not require such a representation training stage (stage-1) at all.

Another reason why BLIP-2 stage-1 is slow is that the computation of Image-Text Contrastive and Image-Text Matching losses needs `concat_all_gather` operations that require GPU communications. Further Image-Text Matching requires binomial sampling of hard negatives. In comparison, our EVL<sub>Gen</sub> circumvents such computations/communications.

## 6 Discussion and Conclusion

This paper introduces EVL<sub>Gen</sub>, an efficient and streamlined pre-training framework for vision-language generative models. Like BLIP-2, EVL<sub>Gen</sub> employs frozen ViT and LLM. It further leverages a conventional Transformer architecture with token-merging capabilities, known as TomeFormer, to act as the vision-to-language connector. Compared to BLIP-2, EVL<sub>Gen</sub> offers the distinct advantage of one-stage training. This reduces computational overhead and maintains competitive performance even with only 1/3 to 1/6 of the computational budget required by BLIP-2.

We have also extended EVL<sub>Gen</sub>’s applicability to video captioning tasks by incorporating the *Temporal Attentive Soft Token Contextualizing* into its ViT. This enhances the model’s temporal modeling capabilities, culminating in the creation of EVL<sub>Gen</sub>-Video. This extension has proven efficacious, delivering commendable performance even without specialized video-text pre-training. Our investigation underscores that a temporal module, which does not disrupt the integration of the well-pretrained image-text model (e.g., BLIP-2 and EVL<sub>Gen</sub>), is a key factor contributing to this success.

EVL<sub>Gen</sub> demonstrates the possibility of achieving state-of-the-art performance in vision-language tasks without the need for complex training regimens or high computational budgets. This work thus makes a significant contribution to the ongoing efforts to develop more accessible, efficient, and powerful models for understanding and generating visual and textual information.

## Limitations

While E2VLGen has showcased its capacity for rapid convergence in VLM pre-training and has demonstrated notable proficiency in zero-shot image/video captioning, certain limitations warrant consideration.

- Our approach is guided by a straightforward design aimed at facilitating the efficient and effective training of VLMs. To maintain simplicity in our methodology, we adopt a fixed value of  $r$  (19) within TomeFormer to compress visual information (i.e., a fixed length of visual soft-prompt). However, it is worth acknowledging that various images or videos might benefit from distinct optimal compression rates ( $r$ ). Consequently, the incorporation of a variable  $r$  (i.e., variable length of soft-prompts for language models) may be deemed more desirable (A similar concern is present in BLIP-2, where the length of soft-prompts is consistently set to 32.).
- One trade-off associated with the simplistic design of TomeFormer is its inability to enable text-specific selection of visual features. In applications like VQA, extracting visual features conditioned on the accompanying questions is considered beneficial. However, the current configuration of TomeFormer lacks the provision for this text-conditioned property within VLMs. A prospective redesign of TomeFormer that incorporates text-conditioned visual feature selection holds the potential to enhance VQA performance.

## Ethics Statement

This research aims to enhance both the efficiency and applicability of vision-language generative models via EVL<sub>Gen</sub>. Although our research does not involve human subjects directly, it is important to acknowledge and discuss the broader ethical implications.

**Data Bias and Fairness:** Our model is trained on publicly available datasets, namely CapFilt, MSCOCO, MSR-VTT, MSVD, and WebVid.

While these datasets are widely used, we acknowledge that we cannot fully ascertain the extent to which they may contain discriminatory, biased, or sensitive material. Given that our model inherits the biases present in these training datasets, there exists the risk of perpetuating or even amplifying existing societal biases. Despite the broad acceptance of these datasets, caution should be exercised.

### Real-world Deployment and Responsible Usage:

Like all generative models,  $EVL_{Gen}$  could be misappropriated for creating misleading or harmful content. Thus, it is imperative to implement safety mechanisms to counter such misuse when deploying the model in real-world applications. Special attention should also be paid to ensure that the model does not inadvertently produce outputs that could disclose sensitive or personal information. Finally, while  $EVL_{Gen}$  is intended as a general-purpose model, its application in contexts that could worsen societal biases or spread misinformation is a pressing concern. Developers and researchers employing  $EVL_{Gen}$  are advised to be cognizant of these risks and consider incorporating fairness-aware or truth-aware components into their systems.

## References

- Emanuele Aiello, LILI YU, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. 2024. Jointly training large autoregressive multimodal models. In *The Twelfth International Conference on Learning Representations*.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.
- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE/CVF international conference on computer vision*.
- Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*.
- Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token merging: Your ViT but faster. In *International Conference on Learning Representations*.
- David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Annual meeting of the association for computational linguistics: human language technologies*.
- Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023a. *X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages*.
- Qi Chen, Chaorui Deng, and Qi Wu. 2022a. Learning distinct and representative modes for image captioning. In *Advances in Neural Information Processing Systems*.
- Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023b. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems*.
- Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. 2022b. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality.
- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023a. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023b. *Instructblip: Towards general-purpose vision-language models with instruction tuning*.
- Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. 2022a. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*.

- Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022b. An empirical study of training end-to-end vision-and-language transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Matthew Dutson, Yin Li, and Mohit Gupta. 2023. Eventful transformers: Leveraging temporal redundancy in vision transformers. In *IEEE/CVF International Conference on Computer Vision*.
- Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE conference on computer vision and pattern recognition*.
- Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF conference on computer vision and pattern recognition*.
- Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. In *IEEE/CVF International Conference on Computer Vision*.
- Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *IEEE/CVF conference on computer vision and pattern recognition*.
- Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF conference on computer vision and pattern recognition*.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.
- Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2023. Bootstrapping vision-language learning with decoupled language pre-training. In *Advances in Neural Information Processing Systems*.
- Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *IEEE/CVF international conference on computer vision*.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in neural information processing systems*.
- Kunchang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.
- Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong

- Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantic aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in neural information processing systems*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in neural information processing systems*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.
- Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF conference on computer vision and pattern recognition*.
- Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*.
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*.
- Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. 2021. Iared<sup>2</sup>: Interpretability-aware redundancy reduction for vision transformers. In *Advances in Neural Information Processing Systems*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *IEEE conference on computer vision and pattern recognition*.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.
- Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *IEEE/CVF conference on computer vision and pattern recognition*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernamed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*.
- Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*.
- Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.
- Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*.
- Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*.
- Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023. Image as a foreign

- language: BEiT pretraining for vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022c. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.
- Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msrvt: A large video description dataset for bridging video and language. In *IEEE conference on computer vision and pattern recognition*.
- Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023a. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*.
- Zhiyang Xu, Ying Shen, and Lifu Huang. 2023b. Multi-Instruct: Improving multi-modal zero-shot learning via instruction tuning. In *Annual Meeting of the Association for Computational Linguistics*.
- Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. 2021. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. In *Advances in Neural Information Processing Systems*.
- Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2022. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners.
- Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022a. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*.
- Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022b. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.
- Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*.
- Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2023a. VPGTrans: Transfer visual prompt generator across LLMs. In *Advances in Neural Information Processing Systems*.
- Chunhui Zhang, Chao Huang, Youhuan Li, Xiangliang Zhang, Yanfang Ye, and Chuxu Zhang. 2022a. Look twice as much as you say: Scene graph contrastive learning for self-supervised image caption generation. In *ACM International Conference on Information & Knowledge Management*.
- Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Empirical Methods in Natural Language Processing: System Demonstrations*.
- Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## A Additional Related Works

**Image-Language Models** Vision-language models generally fall into two categories: dual-encoder models and fusion-encoder models. Pioneering works like CLIP (Radford et al., 2021) and

ALIGN (Jia et al., 2021) serve as exemplary dual-encoder models, demonstrating exceptional performance in zero-shot classification tasks. These architectures also excel in image-text retrieval, as their features can be pre-computed and stored, allowing for *efficient similarity score computation via dot-product operations*. Fusion-encoder models (Lu et al., 2019; Tan and Bansal, 2019; Alayrac et al., 2022; Dou et al., 2022b; Li et al., 2022; Dou et al., 2022a; Xu et al., 2023a), such as ALBEF (Li et al., 2021a), mPLUG (Ye et al., 2023), X-VLM (Zeng et al., 2022), and VLMo (Bao et al., 2022), employ cross-attention mechanisms to enable deep interactions between visual and linguistic features. Other designs include concatenating features of each modality before feeding them into a Transformer (Chen et al., 2020; Li et al., 2020; Zhang et al., 2021; Gan et al., 2020; Li et al., 2021b; Cho et al., 2021; Huang et al., 2020, 2021; Shen et al., 2022; Kamath et al., 2021; Yang et al., 2022a; Wang et al., 2022b; Kim et al., 2021; Xue et al., 2021; Wang et al., 2022a, 2023; Xu et al., 2021). These models excel in complex tasks like closed-set Visual Question Answering (VQA) and visual entailment.

BLIP-2 (Li et al., 2023a) is proposed to leverage pre-trained frozen ViTs and LLMs to alleviate the computation demands in the full end-to-end training. Under this learning paradigm, Zhang et al. (2023a) introduce visual-prompt transfer learning to mitigate visual-language re-alignment cost when using different LLMs. Jian et al. (2023) propose decoupled language pre-training to alleviate the intensive data requirement in BLIP-2.

More recently, VLM research (Yin et al., 2023; Fu et al., 2023) also explores visual instruction tuning (Xu et al., 2023b; Liu et al., 2023; Ye et al., 2023; Dai et al., 2023a), multi-modal in-context learning (Chen et al., 2022a; He et al., 2023; Shao et al., 2023; Gupta and Kembhavi, 2023; Yang et al., 2022b) and Chain-of-Thought prompting (Zheng et al., 2023; Lu et al., 2022), interleaved image-text generation (Aiello et al., 2024), and hallucination (Li et al., 2023c).

**Visual Redundancy** The concept of visual redundancy plays a pivotal role in the field of computer vision. It pertains to the phenomenon where semantic information is conveyed through multiple channels, often involving the use of various visual elements like shape and color to represent complex symbols. Recognizing the impact of this

Phase	Bsz	init LR	min LR	steps	warm-up
Stage-1	1600	$1e^{-4}$	$1e^{-5}$	250k	5k
Stage-2	1600	$1e^{-4}$	$1e^{-5}$	80k	5k

Table 7: Training configurations for our re-produced BLIP-2.

redundancy on deep learning algorithms, there has been a shift towards devising methods to minimize it, thereby enhancing efficiency. For example, IA-RED2 (Pan et al., 2021) has an interpretable design to dynamically and gracefully remove redundant tokens. In Chen et al. (2022b), the redundancy of ViT is discussed on embedding, attention, and weight levels. Eventful Transformer (Dutson et al., 2023) discusses the temporal redundancy in the temporal dimension.

## B BLIP-2 Training Configurations

We re-train BLIP-2 from scratch using LAION-115M, CCS-14M, MSCOCO from CapFilt dataset (Li et al., 2022). The models are trained on eight A100-80G GPUs. The training configurations are shown in Table 7.

## C Technical Details of Token Merging

In this section, we briefly summarize the technical designs of Token Merging (ToMe) (Bolya et al., 2023). Token Merging was initially proposed in Bolya et al. (2023) for accelerating ViTs without training. Whereas we re-purpose ToMe to condense the visual features used as language prompts in the LLM. Please refer to Sections 3 of Bolya et al. (2023) for full details.

**Strategy.** The token merging operations take place in between the attention and MLP blocks of each Transformer layer. ToMe reduces  $r$  tokens per layer. And over the  $L$  layers of a Transformer, it reduces a total of  $r \times L$  tokens. In our experiments, we set  $r = 19$  and our TomeFormer has 12 layers.

**Token Similarity.** The similarities of tokens are defined by the cosine similarity (dot product) of keys of tokens.

**Bipartite Matching.** The bipartite soft matching algorithm is summarized as follows:

- Tokens are randomly partitioned into two sets  $\mathbb{A}$  and  $\mathbb{B}$ .
- Each token in set  $\mathbb{A}$  is linked to the most similar token in set  $\mathbb{B}$ .

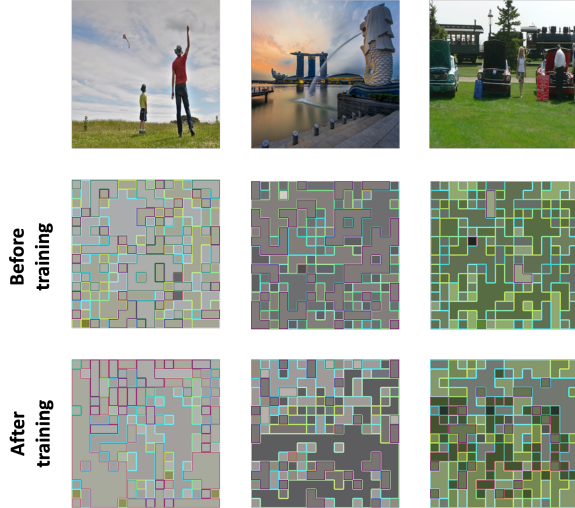


Figure 5: Additional pre- and post-training visualization of merged tokens in  $EVL_{Gen}$ .

- Keep links with top  $r$  similarities.
- Merge tokens with top  $r$  links.
- Concatenate set  $\mathbb{A}$  and  $\mathbb{B}$  back into a single set.

## D Details on Our Implementations of BLIP-2 and VideoCoCa

Our reported results of our re-trained BLIP-2 are slightly worse than what was reported in Li et al. (2023a). There are mainly three reasons:

- We are only able to download 104M image-text pairs from the original 129M CapFlit dataset.
- We intentionally exclude the VG dataset from our pre-training procedure, as it mainly consists of localized captions. Thus, our re-trained BLIP-2 is more challenging when evaluated on GQA, which is built on VG dataset.
- The exact dataset weighting is unknown from the LAVIS project, we use a weighting that is based on the size of each pre-training dataset, i.e., CSS14M, LAION115M, MSCOCO.

For video captioning in Table 3 and Table 4, because VideoCoCa is not open-sourced, we use a pre-trained model OpenCoCa released by mlfoundations.

## E Additional Token Merging Visualization in $EVL_{Gen}$

In this section, we provide additional examples of token merging visualization (before and after training) in  $EVL_{Gen}$  in Figure 5.

The visual features compressed via token merging in the TomeFormer exhibit semantic informativeness even prior to training. This inherent char-

LLM	ViT	VQA	GQA	OK	COCO
OPT	CLIP <sub>L</sub>	44.7	<b>30.9</b>	22.7	123.9
	EVA-ViT <sub>G</sub>	<b>45.2</b>	30.6	<b>22.8</b>	<b>130.6</b>
Vicuna	CLIP <sub>L</sub>	49.0	33.0	23.6	125.2
	EVA-ViT <sub>G</sub>	<b>52.5</b>	<b>34.6</b>	<b>27.9</b>	<b>132.4</b>

Table 8: Ablation studies on different visual encoders of  $EVL_{Gen}$ . VQA $\rightarrow$ VQAv2, OK $\rightarrow$ OKVQA, COCO $\rightarrow$ MSCOCO (CIDEr).

$r$	VQA	GQA	OK	COCO
10	45.7	31.3	23.6	127.5
13	46.2	31.4	24.5	128.0
16	46.3	30.9	24.3	129.9
19	45.2	30.7	22.8	130.6
22	45.5	31.5	21.8	129.7
25	44.7	31.1	21.5	128.7

Table 9: Ablation studies on  $r$  in TomeFormer.

acteristic facilitates  $EVL_{Gen}$ 's ability to converge quickly in an end-to-end training setup.

## F Ablation on Visual Encoders

One of the limitations of BLIP-2 is that it requires an extensive stage-1 pre-training for every different vision encoder. This prohibits practitioners from exploring stronger ViTs when they are available.  $EVL_{Gen}$  offers fast training of models, allowing for exploration of different ViTs as visual encoders.

We conduct an ablation experiment on two ViTs (CLIP<sub>L</sub> and EVA-ViT<sub>G</sub>) using  $8\times$  RTX-A6000 and the CCS-14M dataset for pre-training. The models are trained for 60,000 steps.

Shown in Table 8,  $EVL_{Gen}$  is robust to different visual encoders, and the stronger ViT leads to better results. This implies that while  $EVL_{Gen}$  also requires retraining for different ViTs, but the single-stage training and quick convergence allow it to benefit from a future release of the latest ViTs, given its capability of fast adaptation.

## G Ablations on TomeFormer

In this section, we provide experimental results in VQAv2, GQA, and OKVQA of  $EVL_{Gen}$ , by varying hyper-parameter  $r$  in TomeFormer. As we can see from Table 9,  $EVL_{Gen}$  is robust to the choice of  $r$ .