$\mathcal{E}$-ViLM : Efficient Video-Language Model via Masked Video Modeling with Semantic Vector-Quantized Tokenizer

Jacob Zhiyuan Fang[†], Skyler Zheng[†], Vasu Sharma[‡], Robinson Piramuthu[†]
[†]Amazon, [‡]Meta
{zyfang, nzhengji,robinpir}@amazon.com vasusharma@meta.com

Abstract

*To build scalable models for the challenging real-world tasks, it is important to learn from diverse, multi-modal data in various forms (*e.g.*, videos, text, images). Amongst the existing works, a plethora of them have been focusing on leveraging large but cumbersome cross-modal architectures. Regardless of their effectiveness, larger architectures unavoidably prevent the models from being extended to real-world applications, so building a lightweight VL architecture and an efficient learning schema is of great practical value. In this paper, we propose an $\mathcal{E}$fficient **V**ideo-**L**anguage **M**odel (dubbed as $\mathcal{E}$-ViLM) and a masked video modeling (MVM) schema, assisted with a semantic vector-quantized tokenizer. In particular, our $\mathcal{E}$-ViLM learns to reconstruct the semantic labels of masked video regions, produced by the pre-trained vector-quantized tokenizer which discretizes the continuous visual signals into labels. We show that with our simple MVM task and regular VL pre-training modelings, our $\mathcal{E}$-ViLM , despite its compactness, is able to learn expressive representations from Video-Language corpus and generalize well to extensive Video-Language tasks including video question answering, text-to-video retrieval, etc. In particular, our $\mathcal{E}$-ViLM obtains obvious efficiency improvements by reaching competing performances with faster inference speed:* i.e.*, our model reaches* 39.3% *Top-*1 *accuracy on the MSRVTT benchmark, retaining* 91.4% *of the accuracy of state-of-the-art larger VL architecture with only* 15% *parameters and* 94.8% *fewer GFLOPs. We also provide extensive ablative studies that validate the effectiveness of our proposed learning schema for $\mathcal{E}$-ViLM .*

1. Introduction

The task of video (**V**) and language (**L**) pre-training aims to learn joint and robust cross-modal representations from video-text pairs. Recent advancements of VL pre-training have obtained great development and are primarily reflected in the aspects of leveraging more video-text pairs for scaling pre-training [55, 91, 53]; superior visual encoder for expressive video representations [44, 42, 78]; and unified VL architectures [23, 85, 79, 81, 77] etc. Despite this, the success of these prior arts unanimously relies on escalating parameters that impede their real-world applications due to the high latency and large memory footprint during inference.

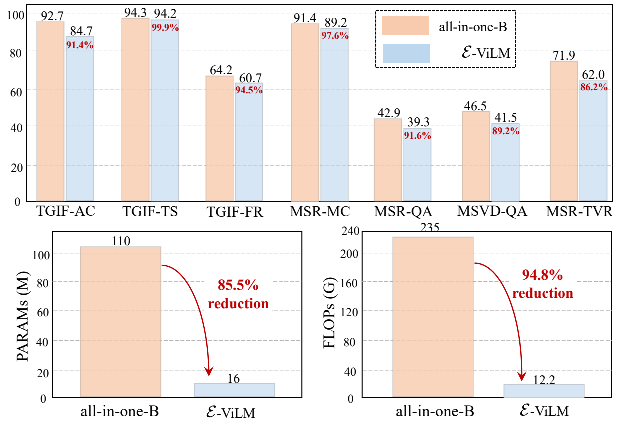

Figure 1. Comparisons of performances, parameters (PARAMs in M), and inference speed (FLOPs in G) between state-of-the-art video-language model [36] as opposed to $\mathcal{E}$-ViLM . We propose a small video-language architecture as well as a novel pre-training schema. $\mathcal{E}$-ViLM is lightweight: it takes less than 15% parameters of all-in-one-B and 94.8% fewer GFLOPs while retaining competing performances on extensive video-language tasks. We show that our model retains prominent performances for video question answering task on TGIF [32], MSRVTT [87], MSVD [12], and text-to-video retrieval task on MSRVTT (R@10).

Existing efforts to develop small VL architectures [80, 20, 83, 84] or end-to-end cross-modal architectures [38, 81] for image captioning [19], VQA [11], are restricted to image-text domains only. The challenge then becomes how to train these small VL architectures with increased accuracy. The majority of existing VL pre-training models stand on a common plateau of modeling, i.e., a cross-modal con-

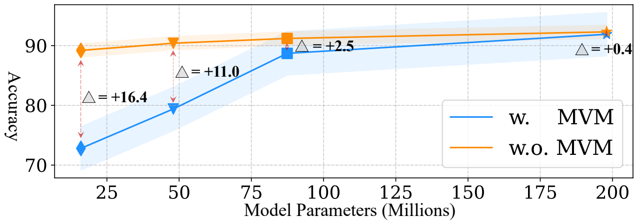

Figure 2. Masked visual token learning exhibits obvious advantages on smaller VL models, which prefer more discretized visual tokens with highly condensed semantics as supervisions. Baseline denotes the VL models w.o. masked visual token modeling. Experimental details are discussed in Appendix.

trastive modeling [60, 33] in which visual representations are optimized to be aligned with the matched text while contrasting with the unmatched text; and a masked language modeling [14] schema in which the model learns to predict randomly masked textual tokens. The modeling of visual representations for VL learning, especially on small VL architectures, are typically less effective and still needs more investigation.

Recent studies also witness significant developments in masked token prediction for unsupervised representation learning in both natural language processing (e.g., BERT [14], GPT [61, 9, 60]) and vision domain (e.g., BEiT [6] and MAE [28]) from which the learned representations generalize well to diverse tasks. Despite remarkable headway made, in the regime of cross-modal tasks dealing with image/video and language, this *"mask-and-predict"* modeling is primarily limited to just the language modality [69, 43, 46, 92, 20, 80, 17] (reconstructing solely masked language tokens). Hereby, we conjecture that a critical reason that imposes the masked visual token prediction on the VL domain is the inconsistent information density during joint VL pre-training, especially on smaller architectures where visual representations lack expressiveness. In particular: a unified VL architecture (e.g., ViLT [38] and [77]) encodes both modalities where *continuous* visual embedding contains redundant information yet embeddings from language are highly condensed with latent semantics. Such inconsistency prevents the model from inferring contextualized & aligned cross-modal relations, and it becomes even more visible on smaller VL architectures for highly redundant visual data like videos. Figure 2 shows the Multiple-Choice accuracy of VL models, pre-trained w./w.o Masked Video Token Modeling with different number of learnable parameters: we clearly observe that smaller VL architecture benefits more from discretized masked video token prediction. This clearly indicates that MVM can be the cure of lesion in the furtherance of a more tiny and powerful VL architecture.

In this paper, we investigate how to set up masked video modeling (MVM) with the goal of obtaining more robust and general VL representations on small VL architectures for diverse video and language tasks. We propose to use a semantic Vector-Quantized (VQ) module to discretize the visual embedding into a set of labels in order to address the aforementioned problem. To accomplish this, we choose the nearest code from a parameterized and learnable codebook as the target tokens for reconstruction. Different from previous efforts like VIOLET [23] and [24], we train the video tokenizer without using pixel space, but by reconstructing the words from the video-level captions. As a result, the density gap is lessened because the input visuals have been converted into a smaller number of discrete labels. In order to encourage the learning of inferring the latent semantics of masked visual regions from contextualized visual contexts and languages, our proposed MVM is then formulated as predicting the semantic labels of a set of randomly masked video regions. Masked video/language modeling and widely used video-text matching modeling constitute our overall loss. We carry out extensive tests to verify the performance of our proposed MVM. We particularly experimented with VL pre-training on the WebVID [5] dataset and then assessed the learned representations on a multitude of benchmarks related to video and video-language, including generalization ability, video question answering, and video retrieval by text (linear probe of activity recognition). Our proposed $\mathcal{E}$-ViLM achieves competing performances consistently across all benchmarks with less than 15% learnable parameters and 40 times fewer GFLOPs than the state-of-the-art (see Figure 1). The learned representations from the $\mathcal{E}$-ViLM also exhibit excellent generalization abilities when applied to an activity recognition task.
To summarize our contribution:

- We propose a small and novel video-language architecture, $\mathcal{E}$-ViLM, which achieves prominent accuracy & inference speed trade-off. To the best of our knowledge, this is the very initial attempt that leverages a small video-language model for VL representation learning.
- To counter the pain point of training small VL architecture caused by inconsistent cross-modal information density, we propose a novel semantic reconstruction based video tokenizer for discretized masked video modeling. This facilitates our small VL model to efficiently learn from large-scale video-text pairs for expressive representations.
- Extensive quantitative experiments validate the effectiveness of our proposed model and learning schema. Our proposed $\mathcal{E}$-ViLM achieves competing results across extensive benchmarks with obvious inference advantages.

2. Related Work

Efficient Vision-Language Models. The remarkable performance of most VL models [58, 27, 29, 68, 31, 30, 25,

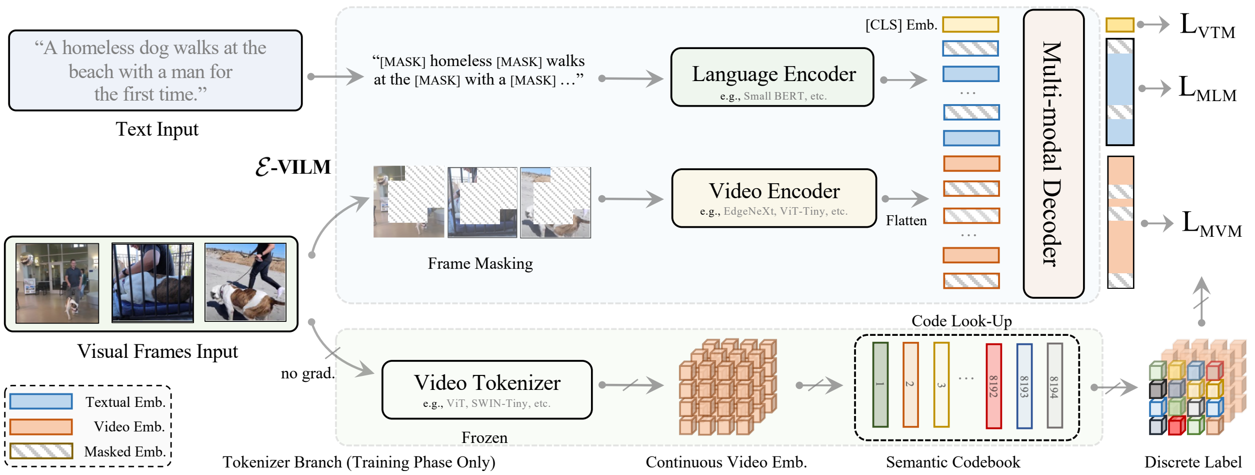

Figure 3. Overall training pipeline of our proposed $\mathcal{E}$-ViLM . We leverage efficient design in pursuit of a lightweight Video-Language architecture with fast inference speed. To achieve this, we utilize EdgeNext [54] as the video backbone that encodes the masked video frames as video embedding, and a tiny variation of BERT [14], as the language encoder. A multi-modal decoder module is composed of a stack of self-attention blocks [76] with decreased hidden dimensions for video/language embedding fusion. For competing performances, we propose to employ a video tokenizer (composed of an encoder with ViT [60] or Video SWIN [50, 49], and a vector-quantized tokenizer) to map the *unmasked* frames into continuous video embedding, which are later discretized into semantic tokens by nearest neighbor look-up. Our masked video modeling ($\mathcal{L}_{\text{MVM}}$) is formulated as predicting the categorical labels of the masked video embedding. Our overall training schema also includes the video-text matching modeling ($\mathcal{L}_{\text{VTM}}$) and masked language modeling ($\mathcal{L}_{\text{MLM}}$). Note that the tokenizer branch (in the green rectangle) is included only during the training phase providing semantic labels. (Best viewed in color).

42, 89, 86, 70, 86] is closely dependent on the great visual representations from expressive visual encoders like object detector [2, 46]. Zhang et al. [92] develop improved Faster-RCNN [64] for the sake of better visual representations that substantially uplifts the results on various VL tasks. Despite so, cumbersome visual encoder like object decoder arguably hinders the usefulness of VL models in real-world applications due to their demands of high computational resources and are de facto the major computational burden in VL architecture. To resolve this challenge, many recent efforts propose to optimize the detector [80, 20], utilize grid visual representations [34, 38, 19, 17, 79] with no regional operations in the detector. There also emerge many efforts to build efficient Vision-Language models. For instance, [80] proposes to utilize an EfficientNet variations [71] as the detector for visual feature extraction, which largely minimizes the inference bottleneck. Other follow-ups exploit knowledge distillation technique [20] to improve the accuracy of VL models. [34, 53, 38, 20] explore the feasibility of using end-to-end models to reduce the inference time on object detector than two-stage VL architectures.

Masked Auto-encoder. Recent studies also witness significant development in masked token prediction for unsupervised representation learning in both natural language processing, e.g., BERT [14] and GPT [61, 9, 60]. This *mask-and-predict* pre-training manner extends to the visual representation learning: BEiT [6] for the first time proposes to tokenize the continuous visual embedding into discrete tokens and then reconstruct the randomly masked token. MAE [28] further simplifies this by directly predicting the masked patches and using their pixel values as the reconstruction target. Follow-up works further advance this to different architectures [45], various visual data (*i.e.*, videos [21, 57, 72], depth image [88] etc). In particular, concurrent works from Tong et al. [73] and Feichtenhofer et al. [22] both show that video data, due to its high spatio-temporal redundancy, requires a significantly high masking ratio. Chen et al. [13] adopts a masked region modeling to reconstruct the masked regional representations of the image for VL pre-training but with limited improvement. Bachmann et al. [4] also [26] further extend this to a multi-modal-multi task setup where they show that additional modalities can be incorporated into the training objective by multi-modal masking. In particular, VIOLET [23] also proposes to leverage the video tokenizer to provide discretized video patch labels for masked video modeling. Our work differentiates with [23] majorly from two perspectives: 1. we focus on developing small and efficient architectures. 2. [23] leverages tokenizer of DALL-E [62] as video tagger, which is obtained from a pixel-level re-construction task while ours is trained from semantic-level re-construction.

3. Efficient Video-Language Modeling

In this section, we describe how we build and train our efficient Video-Language Model ($\mathcal{E}$-ViLM) that is smaller

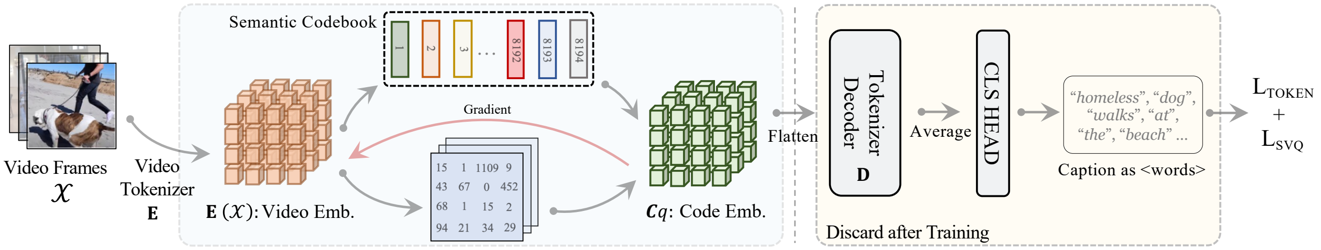

Figure 4. Overall training pipeline of our proposed video tokenizer. Our tokenizer is an Encoder-Decoder architecture where video frames are encoded into continuous video embedding, and discretized into code via nearest neighbor search. We optimize the tokenizer by reconstructing the words from captions via a multi-label classification task.

and faster. This section first discusses in detail about the architectural design of the model, followed by our proposed Masked Video Modeling aiming for more efficient VL pre-training and ends with our overall training schema.

3.1. Model Architecture

Figure 3 illustrates the overall architecture of our $\mathcal{E}$-ViLM, which is an end-to-end architecture consisting of : a Language Encoder ($\mathbf{ENC}_{\mathrm{L}}$) that encodes the text ($\mathcal{T}$) into a sequence of continuous embedding, Video Encoder ($\mathbf{ENC}_{\mathrm{V}}$) that encodes stack of sparsely sampled video frames ($\mathcal{I}$), and a transformer decoder ($\mathbf{DEC}_{\mathrm{VL}}$) that allows cross-modal interaction across Visual and Language modalities:

$$\mathbf{e}^V = \mathbf{ENC}_{\mathrm{V}}(\mathcal{I}),\ \ \mathbf{e}^L = \mathbf{ENC}_{\mathrm{L}}(\mathcal{T}),\ \ \mathbf{h} = \mathbf{DEC}_{\mathrm{VL}}(\mathbf{e}^V, \mathbf{e}^L),\ \ (1)$$

where $\mathbf{e}^{V/L}$ denotes the visual/language embedding. In quest of a lightweight VL model, we purposefully constrain the size of each module in $\mathcal{E}$-ViLM by leveraging a few efficient but well-performing deep architectures across both Vision and Language benchmarks.

Video Encoder. In our work, we resort to a lightweight 2D visual architecture as the visual encoder to obtain frame-wise grid representations as visual embedding. To be specific, given the sparsely sampled T video frames $\mathcal{I} = \{\mathcal{I}_0 \dots \mathcal{I}_T\}$, $\mathcal{I}_i \in \mathbb{R}^{H \times W \times 3}$, each frame is represented as continuous embedding $\mathbf{e}_i^V \in \mathbb{R}^{H/S \times W/S \times D}$, where S and D denote the down-sample ratio of the network and the dimension of embedding, respectively. We then flatten and concatenate all embeddings, followed by a linear transformation for dimension adjustment to produce the video-level visual representation $\mathbf{e}^V$.

Language Encoder. Transformer architecture [76] and its variations [14, 61] achieve exceptional performance on multiple natural language understanding tasks. Yet, transformer like architectures suffer from slow inference speeds primarily because of the quadratic complexity of the self-attention mechanism and a large numbers of parameters. Although improved works [74, 67] take advantage of efficient/local self-attention mechanism or knowledge distillation [35] technique to address this, their performances still fall short of expectation. Instead, we propose to utilize a small BERT architecture with decreased hidden dimension and squeezed vocabulary where only frequent words are retained in favor of a more tiny tokenizer layer. A linear projection layer maps the language embedding to a consistent dimension at the end. We introduce detailed architectural settings in the experiment section.

Multi-modal Decoder. Our multi-modal decoder is composed of a stack of transformer blocks that takes the concatenation of visual and language embeddings as input, enabling the cross-modal interaction. To incorporate information for distinguishing between the two modalities, we add token type embeddings $\mathbf{t}^L$ and $\mathbf{t}^V$ to language embedding and video embedding, respectively. We also attach a special classification token `[CLS]` (as shown in Figure 3):

$$[\mathbf{h}^{\mathrm{CLS}}, \mathbf{h}^{\mathrm{L}}, \mathbf{h}^V] = \mathbf{DEC}_{\mathrm{VL}}([\mathbf{e}^{\mathrm{CLS}}, \mathbf{e}^L + \mathbf{t}^L, \mathbf{e}^V + \mathbf{t}^V]).\ (2)$$

Similar to the language encoder, we decrease the hidden dimension in each block.

3.2. MVM with Semantic Video Tokenizer

We propose a novel masked video modeling (MVM) based on a semantic video tokenizer to assist VL representation learning. Masked language modeling (MLM) obtains great success [14] in a series of natural language comprehension tasks. Recent efforts on visual masked auto-encoder [28, 72] also show prominent effectiveness on self-supervised representation learning. In essence, the visual model learns to restore the masked raw pixels based on the observable pixels, which are continuous and high-dimensional, while the prediction target in MLM is the language token with high semantics and dense information. In particular, VL architecture encodes both modalities where continuous visual embedding contains redundant information yet semantics from language are highly condensed. Such inconsistency prevents the model from inferring out contextualized & aligned cross-modal relations, which play an essential role in VL pre-training [10], especially on highly redundant visual data like videos. For this reason, we propose to leverage an independent Tokenizer branch (described in a later section) as a video tagger, providing one-hot semantic labels $\mathbf{y}_i^V$ for the i-th masked video region. To be specific, we randomly mask out video frame patches by replacing them with zero-value pixels. Our proposed MVM aims to reconstruct the semantic labels of the masked regions via a classification task. And we use cross-entropy to measure this error:

$$\mathcal{L}_{\text{MVM}} = -\sum_{i=1}^{N} \frac{1}{N} \text{CE}(\hat{\mathbf{y}_i}^V, \mathbf{y}_i^V), \tag{3}$$

where N represents the number of masked regions and $\hat{\mathbf{y}_i}^V$ is the predicted semantic probability of the i-th masked video region, obtained by a semantic classification head: $\hat{\mathbf{y}_i}^V = \mathbf{HEAD}^V(\mathbf{h}_i^V)$. In the later section, we explain how we construct our video tokenizer and discuss how it differentiates from previous efforts.

3.3. Semantic Vector-Quantized Video Tokenizer

A recent work leverages a visual tokenizer to discretize the visual regions for self-supervised representation learning [59]. [52, 23] further extend this manner to Vision-Language learning and utilize a pre-trained visual tokenizer (i.e., tokenizer of DALL-E [63] and VQ-GAN [18]) as a tagger. However, a notable defect of such a tokenizer comes from its lack of semantics, as they are unanimously obtained from the pre-task of reconstructing raw image pixels in an auto-encoder fashion. To overcome this, we propose to construct our video tokenizer by reconstructing the semantics of the videos.

Following [75, 18], we first learn an asymmetric Encoder-Decoder model consisting of a ViT [60] or VideoSWIN [50] based encoder ($\mathbf{E}$) and a decoder ($\mathbf{D}$) that is composed of stacks of self-attention blocks. It learns to represent the video frames with codes from a learnable and discrete codebook with M latent code embedding $\mathcal{C} = \{\mathbf{c}_1 \dots \mathbf{c}_M\}$, with $\mathbf{c}_j \in \mathbb{R}^D$ (see Figure 3). Our vector-quantization module $\mathbf{q}(\cdot)$ approximates a set of latent codes $\mathcal{C}_q$ for input video frames by a nearest neighbor look-up using the shared embedding space such that:

$$\mathcal{C}_q = \mathbf{q}\Big(\mathbf{E}(\mathcal{X})\Big) := \Bigg(\underset{\mathbf{c}\in\mathcal{C},\mathbf{x}\in\mathcal{X}}{\text{argmin}} \Big|\Big|\mathbf{E}(\mathbf{x}_i) - \mathbf{c}_j\Big|\Big|\Bigg), \tag{4}$$

with $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{T\times H/h\times W/w}$, and $\mathbf{x}_i \in \mathbb{R}^{h\times w\times 3}$ denotes i-th the image patch obtained from T sampled image frames via non-overlapping partitioning.

We introduce the semantic reconstruction task by predicting the discrete words $\mathcal{T} = \{\mathbf{w}_k\}$ of the video caption with latent codes by the decoder. We compute the multi-label probability using a classification head ($\mathbf{HEAD}^S$) given the averaged output embedding from the decoder ($\mathbf{D}$):

$$\hat{\mathbf{y}}^S = \mathbf{HEAD}^S\Big(\frac{1}{|\mathcal{X}|}\sum \mathbf{D}(\mathcal{C}_q)\Big). \tag{5}$$

Considering the extremely imbalanced word distribution and the rarity of positive samples, we adopt focal loss for tokenizer training [7, 20], which shows outstanding results on a multi-label classification task:

$$\mathcal{L}_{\text{TOKEN}} = \frac{1}{K}\sum_{k=1}^{K} \begin{cases} \log(\hat{\mathbf{y}_k}^S), & \mathbf{y}_k^S = 1, \\ \hat{\mathbf{y}_k}^S \cdot \log(1-\hat{\mathbf{y}_k}^S), & \mathbf{y}_k^S = 0, \end{cases} \tag{6}$$

where $\hat{\mathbf{y}_k}^S \in [0,1]$ is the predicted probability for the k-th category, and K denotes the size of the vocabulary. Intuitively, our tokenizer is trained to represent visual frames as a distribution over latent constituents, and these latent codes contain vivid semantics that are essential to Vision-Language pre-training task. Because back-propagation is not possible due to the non-differentiable quantization step, we copy the gradients over $\mathcal{C}_q$ directly to the $\mathbf{E}(\mathcal{X})$ for encoder (straight-through gradient estimating [8]). Our Semantic Vector-Quantized Tokenizer is optimized by the objective:

$$\mathcal{L}_{\text{SVQ}} = \mathcal{L}_{\text{TOKEN}} + \big|\big|sg[\mathbf{E}(\mathcal{X})] - \mathcal{C}_q\big|\big|_2^2 + \big|\big|\mathbf{E}(\mathcal{X}) - sg[\mathcal{C}_q]\big|\big|_2^2, \tag{7}$$

and $sg[\cdot]$ denotes the stop-gradient operation [75], which ensures the encoder commits to an embedding and its output does not grow arbitrarily. After the training, we retain only the encoder and the vector-quantized module as our tokenizer. Note that during VL pre-training, the tokenizer is capable of predicting the discrete semantic labels either online or off-the-shelf.

3.4. Video-Language Pre-Training

The training of $\mathcal{E}$-ViLM consists of two stages, where we first obtain the video tokenizer by optimizing it jointly with $\mathcal{L}_{\text{SVQ}}$ and $\mathcal{L}_{\text{TOKEN}}$. For the overall Video-Language pre-training, we use the common VL pre-training modeling objectives: masked language modeling loss ($\mathcal{L}_{\text{MLM}}$), where a certain ratio of the textual words are masked and the model predicts their label; also Video-Text matching loss ($\mathcal{L}_{\text{VTM}}$) where the model is expected to predict whether the video-text pair matches. Our overall loss at the pre-training stage is the combination of the aforementioned items:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{MVM}}. \tag{8}$$

4. Experiments

We now introduce the detailed implementation of $\mathcal{E}$-ViLM and also discuss the empirical results that validate the effectiveness of our proposed training schema. To highlight the generalizability of $\mathcal{E}$-ViLM, we benchmark performances of $\mathcal{E}$-ViLM and compare them with prior arts on

Methods	**# P**	**FP**	**# F**	**P.T.**	TGIF			MSRVTT		MSVD
					Action	Transition	Frame	MC	QA	QA
SSML [1]	-	-	-	HowTo100M	-	-	-	-	35.0	35.1
CoMVT [65]	-	-	30	HowTo100M	-	-	-	-	39.5	42.6
ClipBERT [42]	137M	3.8K	16	COCO [48] + VG [39]	82.8	87.8	60.3	88.2	37.4	-
JustAsk [89]	157M	4.4K	-	HowToVQA69M [89]	-	-	-	-	41.5	46.3
all-in-one-S [36]	33M	134	3	Web2.5M+HT100M	91.2	92.7	64.0	-	39.5	41.7
all-in-one-B [36]	110M	235	3	Web2.5M+HT100M	**92.7**	94.3	64.2	91.4	42.9	46.5
VIOLET [23]	198M	4.9K	5	Web2.5M+CC	92.5	**95.7**	**68.9**	**91.9**	**43.9**	**47.9**
all-in-one-T [36]	12M	18.1	3	Web2.5M+HT100M	80.6	83.5	53.9	64.8*	36.8*	32.2*
$\mathcal{E}$-ViLM	16M	12.2	4	Web2.5M	**84.7**	**94.2**	**60.7**	**89.2**	**39.3**	**41.5**

Table 1. Comparison of $\mathcal{E}$-ViLM with existing methods with VL pre-training on video question answering and multiple-choice (MC) tasks. **P.T.** denotes the pre-trained data. **# P** and **# F** represent the number of parameters and number of frames. The bottom part of the table shows the results of small VL architectures ($\leq$ 20M parameters). * denotes the results are re-produced using official released pre-trained checkpoint[36]. **FP** denotes the FLOPS of models (in G). Best results are underlined.

multiple video-language benchmarks. We then empirically study the benefits of different VL models and conduct ablative studies for various settings. To this end, we discuss in-depth about the efficiency of $\mathcal{E}$-ViLM and outline potential future works of efficient Vision-Language learning.

4.1. Datasets

Pre-training Dataset. In our experiment, we pre-train our model on WebVid-2.5M [5], which contains 2.5M video-text pairs scraped from the web with captions describing the holistic semantics of the video. Videos from WebVid are averagely around 18 seconds with 20 words per caption.

Downstream Datasets. To assess the representations learned by $\mathcal{E}$-ViLM, we conduct extensive evaluations across multiple video-language downstream tasks with more than 10 benchmarks: spanning over video question answering task on MSRVTT [87], MSVD [12], TGIF-QA (including TGIF-Action, TGIF-Transition and TGIF-Frame) [47, 32] datasets; video retrieval by text task on MSRVTT, DiDeMo [3]. Meanwhile, in order to further evaluate the generalizability of $\mathcal{E}$-ViLM, we continue to transfer the learned VL representations for linear-probe action recognition task on HMDB51 [40] and Kinetics-400 [37] datasets, as well as the zero-shot multiple-choice task on MSRVTT. We report the performance of Recall at K ($R@K, K = 1, 5, 10$) for text-to-video retrieval task; and report the accuracy on VQA task, action recognition and zero-shot multiple choice tasks.

Architectural Details. We build our $\mathcal{E}$-ViLM based on an efficient visual architecture, EdgeNext [54], that involves both convolution layers and efficient local attention operation but with a light computational burden. Specifically, our visual encoder (EdgeNext_small) only consumes 1.26G FLOPs for one 256×256 frame with 5.59M parameters. Note that our $\mathcal{E}$-ViLM is not architecture-specific and we choose EdgeNext [54] in our implementation due to their well-organized open-source code and good accuracy-speed trade-off. For the language encoder, we adopt a tiny BERT variation with decreased hidden size (128) and fewer attention heads per layer (4). In order to further compress the language branch, we truncate the vocabulary of the tokenizer by retaining only the top 30% most frequent words in COCO dataset [48] to avoid dataset bias. This additionally squeezes the language encoder with only 2.4M learnable parameters. Our cross-modal decoder is composed of 4 transformer blocks with decreased hidden size (384). In total, our $\mathcal{E}$-ViLM accommodates 16M parameters (without a task specific head) and can encode video (4 frames)-caption pairs with 12.2 FLOPs(G), which is obviously advantageous compared with previous state-of-the-art. Table 1 also provides the specific number of parameters and FLOPs comparisons with other VL architectures in regard to efficiency.

We instantiate our video tokenizer with different visual backbones and compare their performances: namely, a ViT [16] architecture pre-trained by [60]. Our tokenizer also includes a decoder which consists of 8 layers of transformer architecture, yet this module is dropped after the training phase of the video tokenizer. Our semantic codebook contains $9,420$ learnable prototype embedding in the dimension of 32 initialized randomly. To compare the effects of various tokenizers, we also explore to utilize other architectures as the tokenizer, *e.g.*, discrete variational auto-encoder (dVAE) from DALL-E [63] which is trained by reconstructing the RGB pixels, and a Video Swin Transformer (VideoSwin-Tiny) [50] pre-trained on Kinetics-400 [37] using our learning schema. Our video tokenizer "*escorts*" the pre-training of $\mathcal{E}$-ViLM in the evaluation mode, so it barely takes up resources during VL pre-training.

Methods	P.T.	MSRVTT	DiDeMo
		text-to-video	text-to-video
HT100M [56]	HT100M	14.9 / 40.2 / 52.8	-
HERO [44]	HT100M	16.8 / 43.4 / 57.7	2.1 / 11.4 / 36.1
ClipBERT [42]	COCO [48] + VG [39]	22.0 / 46.8 / 59.9	20.4 / 48.0 / 60.8
TACo [90]	HT100M	28.4 / 57.8 / 71.2	-
VideoCLIP [86]	HT100M	30.9 / 55.4 / 66.8	16.6 / 48.0 / -
Frozen [5]	CC [66] + Web2.5M	32.5 / 61.5 / 71.2	31.0 / 59.8 / 72.4
VIOLET [23]	CC [66] +Web2.5M	34.2 / 63.5 / 73.6	32.9 / 63.0 / 74.7
all-in-one-B [77]	HT100M +Web2.5M	39.5 / 63.3 / 71.9	-
all-in-one-T [77]	HT100M +Web2.5M	16.3 / 37.4 / 53.1*	12.5 / 26.4 / 38.9*
$\mathcal{E}$-ViLM	Web2.5M	27.0 / 51.7 / 64.4	23.8 / 44.9 / 53.2

Table 2. Comparison of $\mathcal{E}$-ViLM with state-of-the-art VL models on **text-to-video-retrieval** task. Results are reported on R@1 / R@5 / R@10. **P.T** denotes the VL pre-training corpus.

Methods	HMDB51	K400
	Top-1	Top-1
Frozen [5]	54.3	50.5
T-AVG. [77]	43.1	44.3
all-in-one-B	52.9	50.8
all-in-one-T	18.4*	16.7*
$\mathcal{E}$-ViLM	39.6	39.7
$\mathcal{E}$-ViLM $^{\text{MVM}}$	41.6	41.9

Table 3. Activity recognition via linear probe on Kinetics-400 [37] and HMDB51 [40] datasets.

Methods	MSRVTT		DiDeMo
	MC	text-to-video	text-to-video
HT100M	52.9	7.5 / 21.2 / 29.6	-
TACo	-	9.8 / 25.0 / 33.4	-
VideoCLIP	-	10.4 / 22.2 / 30.0	16.6 / 46.9 / -
all-in-one-B	78.1	-	-
all-in-one-T	59.0	3.8 / 9.7 / 16.0*	5.6 / 13.2 / 20.5*
$\mathcal{E}$-ViLM	65.2	6.6 / 18.2 / 26.9	8.8 / 23.0 / 32.6

Table 4. Zero-shot performances of $\mathcal{E}$-ViLM on text-to-video retrieval and multiple choice (MC) tasks. We report Top-1 Acc. on MC task and recall scores for retrieval task.

Training. For VL pre-training, we sparsely sample 4 video frames per video and resize them into 256×256 after random augmentation. Different masking strategies and masking ratios, as previous works indicate [73, 72, 24, 23], may affect the representation learning. We adopt block-wise masking [6] with masking ratio 40%. We use optimizer adamW [51] with an initial learning rate of $1e-4$, weight decay of 0.01 during pre-training and train $\mathcal{E}$-ViLM on WebViD for 20 epochs with $\mathcal{L}_{\text{MVM}} + \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{MLM}}$ jointly. For MLM, we randomly mask 15% of the words during training. For downstream tasks, we stick to the sparsely sampled 4 frames and use task-specific learning rates/epochs. Same as the pre-training phase, we trained the video tokenizer on WebViD for 10 epochs with initial learning rate $2e-4$. The embedding before/after look-up have all been $l2$-normalized for Euclidean distance computation. We further discuss the effect of tokenizer in the supplementary materials.

4.2. Results

Video Question Answering. We compare $\mathcal{E}$-ViLM with previous state-of-the-art methods on Video Question Answering (VQA) task on multiple datasets and present the results in Table 1. We clearly observe that $\mathcal{E}$-ViLM reach on par performances or even surpasses most prior arts, where they unanimously employ large architectures. In particular, $\mathcal{E}$-ViLM achieves 39.3% Top-1 Acc. on MSRVTT-QA task, with a significant gain of 2.5% Acc. over all-in-one-T [77], which is a tiny VL architecture. Remarkable improvements are observed on T-GIF and MSVD benchmarks consistently: $\mathcal{E}$-ViLM exceeds all-in-one-T for $+9\%$ Top-1 Acc. on MSVD, $+4.1\%$ on TGIF-Action, $+10.7\%$ on TGIF-Transition and $+6.8\%$ higher on TGIF-Frame. MC denotes the multiple choice task and $\mathcal{E}$-ViLM also performs prominently better than its counterpart. Notably, we note that $\mathcal{E}$-ViLM even exceed most large VL architectures, *i.e.*, ClipBERT [42], CoMVT [93] and SSML [1], indicating the superiority of our efficient architecture which has an obvious advantage regarding the efficiency and inference speed. Also, $\mathcal{E}$-ViLM is surprisingly more data efficient than the competitors and only exploits 2.5M video-text pairs, whereas the baseline approaches use much larger scale pre-training: for instance, JustAsk [89] adopts 69M video-question-answer triplets auto-extracted from narrated videos for training.; All-in-one-T has been pre-trained on much larger VL corpus, WebViD2.5M and HowTo100M [56], which consists of over 100 million video clip-text pairs.

Text-to-Video Retrieval. We study to extend $\mathcal{E}$-ViLM for video retrieval from text task. Table 2 summarizes results on MSRVTT [87] and DiDeMo [3] for comparisons with previous methods. We follow [77, 79] to speed up retrieval efficiency by replacing the one-to-one image-text matching schema with encoding unimodal inputs during fine-tuning and inference. This additionally accelerates the notoriously slow retrieval process. For instance, our model only takes up 2 minutes for one round of text-to-video retrieval on MSRVTT testing split with one V100 GPU device, while regular VL models, however, usually take up to several hours [77] or more. We explain more about how we conduct

VTM	MLM	MVM	text-to-video	QA	MC
-	-	-	14.6 / 34.8 / 51.0	32.8	62.0
✓	-	-	22.9 / 45.9 / 58.9	34.5	69.4
✓	✓	-	23.6 / 47.6 / 60.1	35.6	72.8
✓	✓	✓	27.0 / 51.7 / 64.4	39.3	89.2

Table 5. Ablative studies for effects of different modeling items on text-to-video retrieval task. First row shows the results obtained w.o. any pre-training (downstream task fine-tuning only).

Toeknizer	VOC. RATE	HMDB51	MSRVTT-QA
Baseline	-	33.9	37.9
$\mathcal{E}$-ViLM + $\text{T}_{\text{DALL-E}}$	-	34.2	38.1
$\mathcal{E}$-ViLM + $\text{T}_{\text{BEVT/PECO}}$	-	34.5	38.3
$\mathcal{E}$-ViLM + $\text{T}\_\mathcal{L}_{\text{BCE}}$	64.2%	35.9	38.7
$\mathcal{E}$-ViLM + $\text{T}\_\mathcal{L}_{\text{TOKEN}}$	92.0%	41.6	39.3

Table 6. Performances of $\mathcal{E}$-ViLM with different tokenizers from [63, 15, 82] or obtained by different strategies. VOC. RATE denotes code-book utilization rate and higher value indicates a broader semantic space coverage.

video retrieval in the Appendix. $\mathcal{E}$-ViLM achieves significant performance gain over a tiny VL variation of all-in-one [77], pre-trained across the aggregation of HowTo100M and WebViD2.5M datasets. Although $\mathcal{E}$-ViLM leverages only 16M parameters, it still surpasses some prevailing retrieval architectures, say, HT100M [56], HERO [44], ClipBert [42], and further reaches performance close to TACo [90]. To be specific, $\mathcal{E}$-ViLM achieves R@1/ R@5/ R@10 recall 27.0 / 51.7 / 64.4 on MSRVTT dataset, and 23.8 / 44.9 / 53.2 on DiDeMo. Admittedly, $\mathcal{E}$-ViLM still shows an obvious performance gap with state-of-the-art architectures, which majorly benefited from dozens times more parameters [77, 23, 5]; specific optimization just for video retrieval task (Frozen [5]); and larger VL corpus for pre-training. Compared with them, $\mathcal{E}$-ViLM achieves great speed & performance trade-off and is also notably more efficient.

Transferring to Activity Recognition. In order to comprehensively study the transferability of $\mathcal{E}$-ViLM, we extend it to adapt $\mathcal{E}$-ViLM for activity recognition tasks in a linear probing manner (only the classification layer is trainable). We choose HMDB51 [40] and Kinetics-400 datasets as our testbeds. HMDB51 contains $6,849$ video clips extracted from commercial movies, divided into 51 human action categories, each containing a minimum of 101 clips. Kinetics-400 [37] dataset is a collection of $650,000$ video clips that cover 400 human action classes. We conduct the classification training in a linear probing manner and report the Top-1 Acc. in Table 3. We observe that $\mathcal{E}$-ViLM evidently outperforms all-in-one-T by a large margin: 41.6% *vs.* 18.4%. $\mathcal{E}$-ViLM $^{\text{MVM}}$ denotes the pre-trained $\mathcal{E}$-ViLM optimized with our masked video modeling. We observe that adopting MVM further improves the Top-1 Acc. by additional 2.0%. This verifies that the VL representations learned have great generalization ability, and the proposed $\mathcal{L}_{\text{MVM}}$ further strengthens that.

Zero-shot Multiple Choice and Video Retrieval. We now examine $\mathcal{E}$-ViLM from the perspective of zero-shot recognition and retrieval. We report performances $\mathcal{E}$-ViLM on both the MSRVTT and DiDeMo datasets, where the model is pre-trained and evaluated without any further fine-tuning. Table 4 shows the results of $\mathcal{E}$-ViLM with previous SOTA methods. $\mathcal{E}$-ViLM reaches 65.2% Top-1 accuracy on MSRVTT and $6.6/18.2/26.9$ (R@1/R@5/R@10) for text-to-retrieval task, which is close to such larger VL architectures as VideoCLIP and TaCo. In comparison with all-in-one-T [77], our method noticeably exceeds it. This clearly verifies the great generalization ability of our proposed method.

Effect of MVM in VL Pre-training. In the furtherance of gaining more understanding of the impact of MVM throughout the different stages, we illustrate their performances on MSRVTT retrieval task in Table 5 and then compare the results w/w.o. MVM on HMDB and MSRVTT-MC tasks. We observe from the table that $\mathcal{E}$-ViLM with our proposed MVM clearly outperforms the counterpart (without MVM) and the combination of all modeling items leads to the best result. Also, we observe that the proposed MVM modeling is complementary to MLM and VTM across multiple tasks consistently. Tab 7 shows our comparison results leveraging different video tokenizers. We compare our optimized tokenizer to [49] (motion-level reconstruction), DALL-E (pixel-level reconstruction as [23]), and tokenizer optimized with BCE as reconstruction loss. Our focal-loss optimized tokenizer outperforms others on QA task (39.3 v.s. 38.5). $\mathcal{E}$-ViLM outperforms DALL-E's pixel-level tokenizer to a large extent, confirming that semantically discretized labels are important for VL learning on small VL architectures. The table also shows the performance of [15, 82]'s official tokenizer and we observe a big performance gap (34.5 v.s. 41.6) with $\mathcal{E}$-ViLM mainly because of the lack of semantic optimization as $\mathcal{E}$-ViLM. Different MVM target ($\mathcal{L}_{\text{BCE}}$) suffers from imbalanced words distribution and yields a less vocabulary utilization rate and performances.

4.3. Choice of Video Tokenizer

Performances of $\mathcal{E}$-ViLM can be dependent on the choice of different video tokenizers. In this section, we continue to discuss the effects that Video Tokenizer has on VL pre-training. We instantiate three variations of Video Tokenizer with different visual backbones for $\mathcal{E}$-ViLM and compare the effect of them for VL pre-training: 1). The Tokenizer from DALL-E [63], a text-to-image generation model, 2). A video SWIN based tokenizer pre-trained from the Kinetic Activity Dataset [37]; and 3). A ViT [16] based tokenizer was pre-trained on a image-text contrastive pre-training [60]. We introduce the details of them:

Pixel-reconstruction Tokenizer. Some previous works [6, 23] leverage Tokenizer in DALL-E [63] as the image tokenizer learned by discrete variational autoencoder (dVAE). A recent study proposes to train an advanced version of tokenizer [59] providing better discrete labels for self-supervised representation learning. Overall, there are two modules during visual token learning, namely, a tokenizer and a decoder (disposed after learning). The tokenizer maps image pixels into discrete tokens via nearest neighbor search according to a learnable visual codebook (or vocabulary). And the decoder learns to reconstruct the input raw image based on the visual tokens. The major aspect that differentiates the tokenizer of $\mathcal{E}$-ViLM is that it tries to reconstruct the semantic words instead of pixels. Tokenizer of DALL-E [63] is utilized in concurrent works for self-supervised visual representation learning like [59], also Vision-Language representation learning [24].

Semantic-reconstruction Tokenizer. We propose to learn the tokenizer by only reconstructing the semantic information of the image without resorting to the RGB pixels. In our experiments, this is achieved by our proposed caption word classification, as previously described. We construct two baselines here for comparison: 1). a Video SWIN [49] backbone-based tokenizer; and 2). a ViT [16] based tokenizer that has been pre-trained by [60], and leverages 400M image-text pairs for contrastive pre-training.

For both options, we adopt a similar set of hyper-parameters and settings. In specific, we use $9,420$ prototype codes, with each in 32 dimensions. For decoders, in tokenizer's training, a 8 layer transformer blocks are used. Another challenge that prevents Video SWIN for our L_{MVM} comes from its inconsistent number of output tokens due to the pooling operation along the temporal axis. And following experiences from Fu et al., [24, 23], we remove the 3D pooling layer in the architecture, and this makes the output tokens consistent with our produced tokens. Another notable disadvantage of leveraging Video SWIN comes from its large computation consumption, which largely slows the VL pre-training more than vanilla ViT. In Table 7, we report the performances of leveraging different Tokenizers, as well as the vocabulary usage rate. We observe that the ViT-based tokenizer reaches the highest performance on the VQA task, and the ViT-based method reaches a much higher vocabulary usage rate, which validates the diversity of code embedding, all contributing to the semantic reconstruction collaboratively. Interestingly, we observe that the performance gains from Video SWIN are extremely high for activity recognition tasks, yet the improvement is not that obvious for general VL tasks like VQA. We conjecture that this is caused by the pre-training of Video SWIN on Kinetics datasets [37], which is a classification task for activity recognition. Comparing with the tokenizer of DALL-E, our semantic tokenizers bring obvious better performances, though we leverage much fewer video-text pairs for pre-training. It is worth mentioning that the vocabulary rate for Video SWIN is particularly lower than our method, indicating that it may not generalize well to diverse scenes.

		HMDB51	MSRVTT-QA
	VOC. RATE	Acc.	Acc.
Baseline	-	33.9	37.9
$\mathcal{E}$-ViLM $_{\text{+DALL-E}}$	-	34.2	38.1
$\mathcal{E}$-ViLM $_{\text{+VideoSWIN}}$	38.9%	51.3	38.5
$\mathcal{E}$-ViLM $_{\text{+ViT}}$	92.0%	41.6	39.3

Table 7. Performances of $\mathcal{E}$-ViLM with different tokenizers. VOC. RATE denotes vocabulary usage rate. The vocabulary rate for DALL-E is not officially reported, so we neglect it.

4.4. About Fast Video-Text Retrieval

As discussed in [77, 41], cross-modal architecture takes paired image/video-text pairs for matching scores, which are used for rankings in text-to-image/video retrieval. Yet, such joint retrieval can largely impede the retrieval evaluation (e.g., for 1K video-caption pairs, it yields 1K$\times$1K=1M iterations of forward). To circumvent this, we follow fast retrieval, as in [77], where during downstream fine-tuning we feed in either only texts or videos as input. This, according to our empirical results, reduces the performance to some extent but largely speeds up the evaluation. Though we did not apply these text/video-only strategies during pre-training as in [79], we believe our retrieval results can be further improved by sacrificing the inference efficiency or adopting a more consistent pre-training strategy, which is, however, beyond the scope of this work. This also somehow explains the limited performance of $\mathcal{E}$-ViLM on zero-shot text-to-video retrieval tasks, where no fine-tuning has been conducted and the model naturally does not generalize well to text-only or video-only input. For the retrieval task, we fine-tune the model for 200 epochs with an initial learning rate of $3e-4$.

4.5. Parameters of the Model

We provide the detailed breakdown of parameters in $\mathcal{E}$-ViLM , and present it in Table 8. In Figure 5, we compare the performances & parameters of $\mathcal{E}$-ViLM with state-of-the-art VL architectures on the MSRVTT QA task. We see from the picture that $\mathcal{E}$-ViLM achieve great performance & inference speed trade off over other competent competitors like all-in-one-T, CLIPBert.

5. Discussions

We then proceed to assess the generalization ability of intermediate-learned representations during the pre-training.

Layer Name	Params
Textual Embedding Layer	1.2M
Textual Encoder	2.4M
Textual MLP Layer	49.5K
Token Type Emb. Layer	1.2K
Visual Encoder	5.3M
Visual MLP Layer	117K
Multi-Modal Decoder	7.1M
Multi-modal Pooling Layer	147K
CLS HEAD (VQA)	770
TOTAL	16.3M

Table 8. Parameter breakdown of $\mathcal{E}$-ViLM for VQA task on MSRVTT-QA.

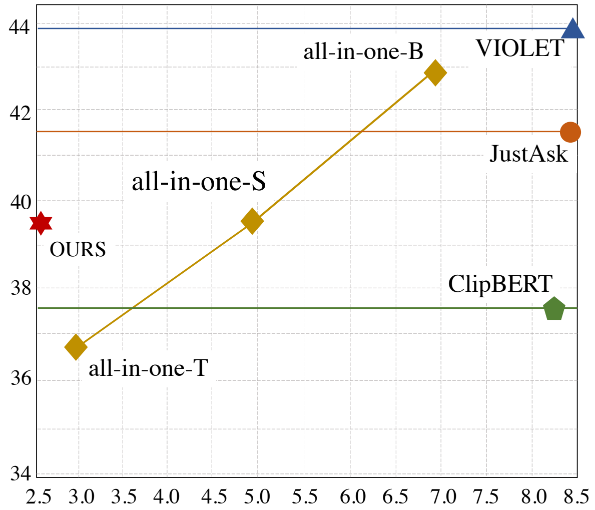

Figure 5. Comparisons of $\mathcal{E}$-ViLM with state-of-the-art VL architectures on MSRVTT-QA. The X axis shows the values of `log(parameters)` to adapt scales.

As shown in Figure 6, we see that from training steps <3K, 6K, 9K, 12K, 15K>, MVM consistently outperforms the regular VL modelings to a non-trivial gap throughout the VL pre-training. This proves that our MVM brings continuous benefits during the VL pre-training and is also more data-efficient. We provide more comparisons, analytical results, and further discussion of the Video Tokenizer in the Appendix.

6. Conclusion

In this paper, we propose the $\mathcal{E}$-ViLM, a small Video-Language model that learns cross-modal representations efficiently and generalizes well to a series of VL downstream tasks. Compared with existing Video-Language models, $\mathcal{E}$-ViLM vastly increases the inference speed by over 40 times with only 15% of the parameters. With our proposed semantic vector-quantized video tokenizer, our VL pre-training is formulated as reconstructing the masked video region's discrete labels as well as the masked language and video-text matching modeling. Extensive experiments indicate that $\mathcal{E}$-ViLM achieves competing performances, approaching most large Video-Language models and significantly outperforming its small counterpart. We anticipate that $\mathcal{E}$-ViLM will lead to more future works in building efficient Vision and Language models for broader real-world impact.

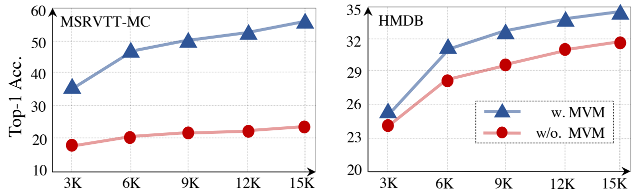

Figure 6. Performances of $\mathcal{E}$-ViLM with or without MVM throughout the 15K-step VL pre-training. Left: zero-shot multiple-choice task on MSRVTT. Right: Activity recognition task on HMDB51 in the linear probing manner.

References

[1] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6644–6652, 2021. 6, 7

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 6, 7

[4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders, 2022. 3

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 6, 7, 8

[6] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 3, 7, 9

[7] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor.

Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020. 5

[8] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 5

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3

[10] Jize. Cao. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision.*, 2020. 4

[11] Qingqing Cao, Prerna Khanna, Nicholas D Lane, and Aruna Balasubramanian. Mobivqa: Efficient on-device visual question answering. 2022. 1

[12] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 1, 6

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 3

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3, 4

[15] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 8

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6, 8, 9

[17] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 2, 3

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 5

[19] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18009–18019, 2022. 1, 3

[20] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1428–1438, 2021. 1, 2, 3, 5

[21] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 3

[22] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners, 2022. 3

[23] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 1, 2, 3, 5, 6, 7, 8, 9

[24] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. *arXiv preprint arXiv:2209.01540*, 2022. 2, 7, 9

[25] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077, 2022. 3

[26] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. 3

[27] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022. 3

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3, 4

[29] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021. 3

[30] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs amp; outputs, 2021. 3

[31] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention, 2021. 3

[32] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 1, 6

[33] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom

Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[34] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020. 3

[35] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019. 4

[36] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv e-prints*, pages arXiv–2203, 2022. 1, 6

[37] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6, 7, 8, 9

[38] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1, 2, 3

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6, 7

[40] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 6, 7, 8

[41] Jie Lei, Xinlei Chen, Ning Zhang, Mengjiao Wang, Mohit Bansal, Tamara L Berg, and Licheng Yu. Loopitr: Combining dual and cross encoder architectures for image-text retrieval. *arXiv preprint arXiv:2203.05465*, 2022. 9

[42] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 3, 6, 7, 8

[43] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 2

[44] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 1, 7, 8

[45] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv preprint arXiv:2205.10063*, 2022. 3

[46] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2, 3

[47] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 6

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 7

[49] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3, 8, 9

[50] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3, 5, 6

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[52] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 5

[53] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 1, 3

[54] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. *arXiv preprint arXiv:2206.10589*, 2022. 3, 6

[55] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 1

[56] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 7, 8

[57] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022. 3

[58] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 3

[59] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 5, 9

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 6, 8, 9

[61] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2, 3, 4

[62] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[63] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 5, 6, 8, 9

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[65] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 6

[66] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 7

[67] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021. 4

[68] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2021. 3

[69] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[70] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 3

[71] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[72] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 3, 4, 7

[73] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. 3, 7

[74] Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. Small and practical bert models for sequence labeling. *arXiv preprint arXiv:1909.00100*, 2019. 4

[75] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 5

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[77] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 1, 2, 7, 8, 9

[78] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2022. 1

[79] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021. 1, 3, 7, 9

[80] Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Minivlm: A smaller and faster vision-language model. *arXiv preprint arXiv:2012.06946*, 2020. 1, 2, 3

[81] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1

[82] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14733–14743, 2022. 8

[83] Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. *arXiv preprint arXiv:2210.07795*, 2022. 1

[84] Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. Distilled dual-encoder model for vision-language understanding. *arXiv preprint arXiv:2112.08723*, 2021. 1

[85] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 1

[86] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3, 7

[87] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 6, 7

[88] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. *arXiv preprint arXiv:2203.09855*, 2022. 3

[89] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 3, 6, 7

[90] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 7, 8

[91] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[92] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2, 3

[93] Peng Zhiliang, Dong Li, Bao Hangbo, Ye Qixiang, and Wei Furu. A unified view of masked image modeling. *arXiv preprint arXiv:2210.10615*, 2022. 7