EgoVLPv2: Egocentric Video-Language Pre-training with Fusion in the Backbone

Shraman Pramanick[1,2†] Yale Song[2] Sayan Nag[3] Kevin Qinghong Lin[4] Hardik Shah[2]
Mike Zheng Shou[4] Rama Chellappa[1] Pengchuan Zhang[2]

[1]Johns Hopkins University, [2]Meta AI, [3]University of Toronto, [4]National University of Singapore

Abstract

Video-language pre-training (VLP) has become increasingly important due to its ability to generalize to various vision and language tasks. However, existing egocentric VLP frameworks utilize separate video and language encoders and learn task-specific cross-modal information only during fine-tuning, limiting the development of a unified system. In this work, we introduce the second generation of egocentric video-language pre-training (EgoVLPv2), a significant improvement from the previous generation, by incorporating cross-modal fusion directly into the video and language backbones. EgoVLPv2 learns strong video-text representation during pre-training and reuses the cross-modal attention modules to support different downstream tasks in a flexible and efficient manner, reducing fine-tuning costs. Moreover, our proposed fusion in the backbone strategy is more lightweight and compute-efficient than stacking additional fusion-specific layers. Extensive experiments on a wide range of VL tasks demonstrate the effectiveness of EgoVLPv2 by achieving consistent state-of-the-art performance over strong baselines across all downstream. Our project page can be found at https://shramanpramanick.github.io/EgoVLPv2/.

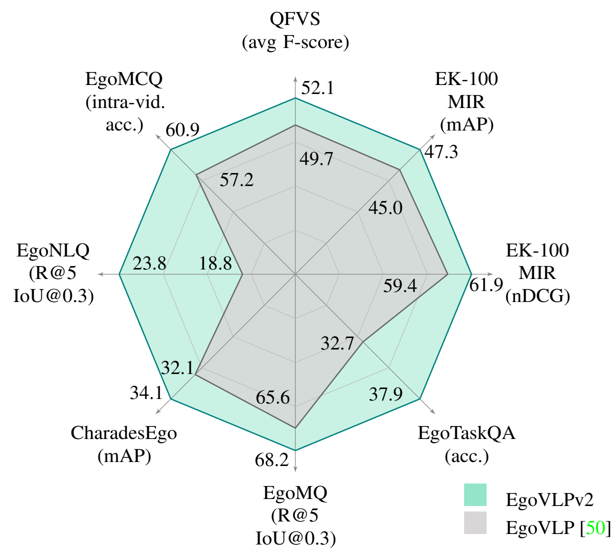


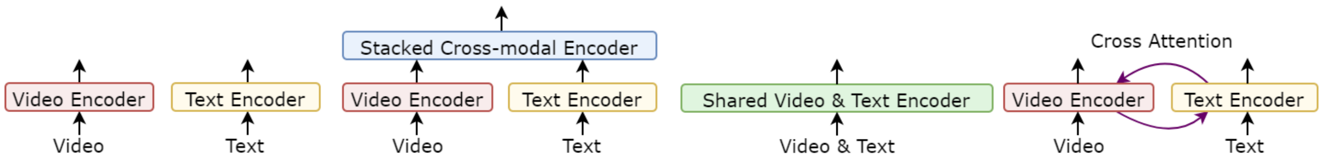
Figure 1: **EgoVLPv2 achieves the state-of-the-art** performance across a broad range of egocentric video understanding tasks (see Table 1 for details) among similar-sized baselines by incorporating cross-modal attention in the transformer backbones to learn video-language representation.

1. Introduction

Video-Language Pre-training (VLP) has proven to be the *de-facto* solution for a variety of video-text tasks, e.g., video-text retrieval [98, 66, 4], VQA [95, 104, 112], zero-shot recognition, [7, 49, 32] and video-text grounding [61, 51]. This is fueled by recent advances in vision [15, 53, 6, 4, 2, 19, 54] and language [84, 14, 52, 102, 74, 12, 73], coupled with large-scale data [98, 111, 59, 4, 24, 13]. Existing video-language datasets generally fall under two categories: third-person view and first-person view (egocentric). The noticeable domain gap between them restricts VLP frameworks pre-trained on third-person videos from performing well on egocentric benchmarks [50]. However, the recent introduction of a massive-scale egocentric dataset Ego4D [24] helps unlock the full potential of egocentric VLP.

Existing egocentric VLP approaches [50, 110, 60, 3] pre-train separate (*dual*) video and language encoders and learn task-specific cross-modal information only during fine-tuning, limiting the development of unified egocentric VL frameworks. Moreover, they lack strong zero-shot inference ability on multi-modal downstream tasks. This issue is commonly addressed by stacking dedicated fusion layers on top of the dual video and text encoders [57, 37, 96, 82, 99, 100, 105], or with a shared video-language architecture [41, 1, 35, 83, 86]. However, these approaches introduce a large number of fusion-specific pa-

†Part of this work was done during an internship at Meta AI.



(a) Dual Encoders (b) Stacked Fusion Layers (c) Shared Encoders (d) Fusion in the Backbone (Ours)

Figure 2: **Four categories of VLP frameworks.** (a) use separate (*dual*) video and text backbones, with InfoNCE [64] as the common pretraining objective [50, 110, 3, 60] (b) use cross-modal fusion layers on top of dual encoders, with MLM, VTM, etc. as common pretraining tasks [57, 37, 96, 82] (c) use a single encoder for different modalities, with similar learning objectives as (b) [41, 1, 35] (d) Fusion in the Backbone (Ours).

rameters, and the resulting encoder cannot be directly applied to uni-modal (video-only) tasks.

In this work, we present the second generation of egocentric VLP (EgoVLPv2), a significant improvement over the previous generation [50] by incorporating cross-modal fusion directly into the video and language backbones. Our approach improves over existing VLP frameworks by: (i) fewer fusion parameters compared to stacked fusion-specific transformer layers or shared encoders, requiring less GPU memory, compute resources, and training time; (ii) the flexibility to switch between dual and fusion encoders, by turning on and off cross-attention fusion using a gating mechanism; (iii) being applicable to both uni- and multi-modal tasks.

Inserting cross-modal fusion directly into the backbone helps unify a wide range of dual- and fusion-encoder-based downstream tasks. Specifically, the "switching" ability of EgoVLPv2 enables us to utilize the same pre-trained encoders for fast retrieval and grounding tasks, which require dual and fusion encoders, respectively. Moreover, in contrast to existing egocentric VLP frameworks that learn task-specific fusion parameters during fine-tuning, EgoVLPv2 reuses the pre-trained cross-attention modules across different tasks, significantly reducing the fine-tuning cost. This enables us to introduce query-focused video summarization as a downstream task, which has recently gained attention in the community [62, 91, 92, 30, 93, 63]. The scarcity of annotated data has been a bottleneck to training decent-sized models end-to-end on this task, with the only available egocentric dataset, QFVS [77], providing merely 135 video-query training samples. EgoVLPv2 achieves new state-of-the-art results on QFVS with a decent margin over the baselines.

In summary, our contributions are: (i) We advance a step forward in egocentric VLP by proposing EgoVLPv2, the second generation of EgoVLP [50] with cross-modal fusion in the backbone. Our proposed framework can switch between dual and fusion encoders and requires 45% lesser compute (GMACs) than learning additional fusion-specific transformer layers. (ii) The switching capability of EgoVLPv2 allows us to unify a wide range of dual- and fusion-encoder-based downstream tasks under the same VLP framework and reduce the task-specific fine-tuning cost by employing the same pre-trained cross-attention modules across different video-language tasks. (iii) We demonstrate the effectiveness of EgoVLPv2 on eight egocentric benchmarks and achieve state-of-the-art performance among comparable-sized backbones. We summarize these results in Figure 1.

2. Related Works

2.1. VLP Frameworks

Video-language pre-training (VLP) has attracted increasing attention in recent years, following the success of image-language pre-training [71, 39, 29, 16, 5, 10, 56, 45, 17, 106, 101, 103, 69, 46, 87, 89, 27, 88, 65, 38] and their applications [9, 21, 26, 43, 70]. There are three broad categories of VLP frameworks (see Figure 2):

Dual Encoders: Many existing egocentric VLP frameworks [50, 110, 3, 60] falls into this category. They use separate video and language backbones and learn task-specific cross-modal fusion during fine-tuning [4, 58, 97, 85]. They are commonly trained using InfoNCE [64] or MIL-NCE [58] objectives, and have been successful in video-text retrieval.

Shared Encoder: Approaches that learn a combined encoder for video and text fall under this category [41, 1, 35, 83, 86]. They are modality independent and can be applied to an image, video, text, audio, time-series, and single-view 3D data. Common learning objectives include masked language modeling [14, 112], masked frame modeling [81, 112], masked token modeling [96], masked modal modeling [57, 96], sentence ordering modeling [36], frame ordering modeling [36, 40], and video-text matching [36].

Encoders with Stacked Fusion Layers: This line of work uses dedicated cross-modal fusion layers on top of dual encoders [57, 37, 96, 82, 99, 100, 105], trained using similar objectives as shared encoders.

The latter two categories introduce a large number parameters for cross-modal fusion. In this work, we propose a fourth category (Figure 2 (d)) by inserting cross-modal fusion in uni-modal backbones using a gating mechanism. Our framework is flexible to act as either dual or shared encoders by switching cross-attention modules off and on.

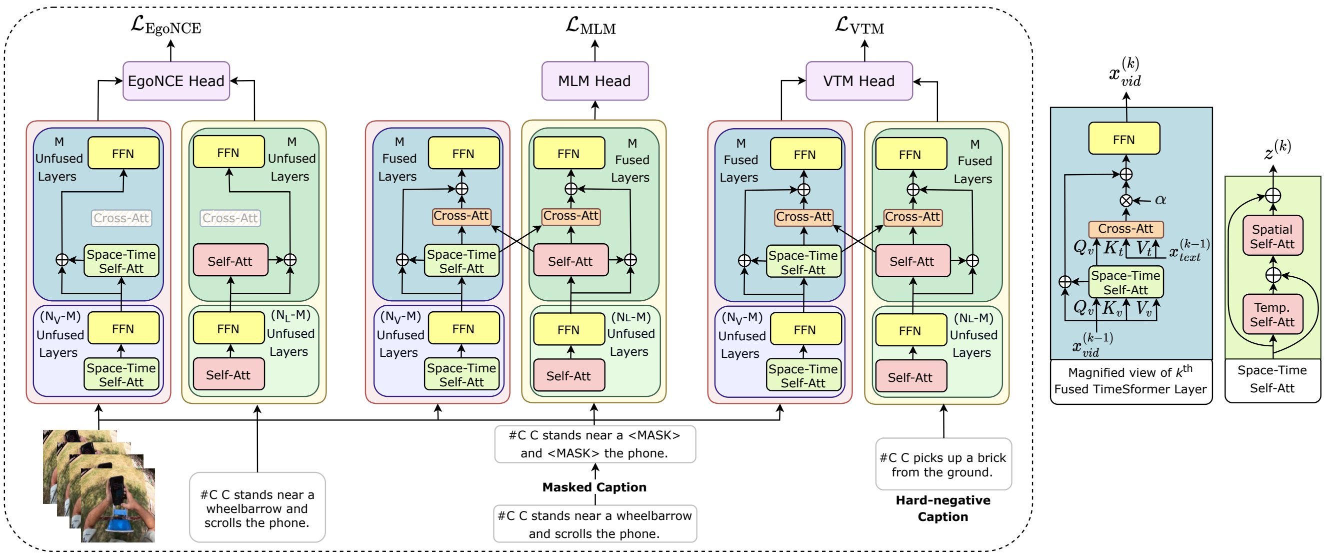


Figure 3: **Computation of three objectives, $\mathcal{L}_{\text{EgoNCE}}$, $\mathcal{L}_{\text{MLM}}$, and $\mathcal{L}_{\text{VTM}}$.** We insert cross-modal fusion into uni-modal backbones with a gating mechanism. During pre-training, every forward iteration contains three steps: (i) cross-attention modules are switched off, EgoVLPv2 acts as dual encoder, $\mathcal{L}_{\text{EgoNCE}}$ is computed. (ii) cross-attention is switched on, EgoVLPv2 acts as fusion encoder, and video-masked narration pair is fed into EgoVLPv2 to compute $\mathcal{L}_{\text{MLM}}$ (iii) cross-attention is kept on, hard-negative video-narration pairs are fed into EgoVLPv2 to compute $\mathcal{L}_{\text{VTM}}$. This *fusion in the backbone* strategy results in a lightweight and flexible model compared to using fusion-specific transformer layers.

2.2. Video-Language Datasets

The success of VLP can be partially attributed to the availability of large-scale open-world video-text datasets such as ActivityNet [33], WebVid-2M [4], and HowTo100M [59]. These datasets comprise videos sourced from the Web, and are paired with the corresponding ASR captions, making them popular for VLP pre-training. Despite their impressive size, these existing video-text pretraining datasets typically feature 3rd-person views. On the other hand, egocentric videos has received increasing interests from the community. Previous egocentric datasets [13, 79, 48, 75, 67] were small-scale and domain-specific. The recently released Ego4D [24] is the first massive-scale egocentric dataset consisting of 3670 hours of videos collected by 931 people from 74 locations across 9 different countries world-wide. Recently, EgoClip [50] offered a filtered version of Ego4D with variable-length clip intervals instead of single timestamps. We train our proposed framework, EgoVLPv2, on the EgoClip version of Ego4D.

3. EgoVLPv2

3.1. Fusion in the Backbone

We use TimeSformer [6, 4] and RoBERTa [52] as our video and language backbones. However, such separate (*dual*) uni-modal encoder design does not capture cross-modality interaction and, thus, fails to produce fine-grained multi-modal representation. Existing VLP frameworks achieve cross-modal fusion by: (i) learning a shared architecture [41, 1, 35, 83, 86] or stack fusion layers on top of dual encoders [57, 37, 96, 82, 99, 100, 105], or (ii) learning cross-modal fusion during fine-tuning [50, 110, 3, 60, 4, 58, 97, 85]. While the former offers superior cross-modal representation and zero-shot inference ability on multi-modal downstream tasks, they introduce a large number of fusion parameters than the latter. In this work, we insert cross-modal fusion into the top few layers of uni-modal backbones to strike a balance between the two ideas.

Figure 3 shows the architecture of EgoVLPv2. Each TimeSformer encoder layer has a divided space-time attention module containing temporal and spatial self-attentions with residual connections. The output of space-time attention at k^{th} encoder layer, $z^{(k)}$, can be expressed as:

$$
\begin{aligned}
\hat{x}_{vid}^{(k)} &= x_{vid}^{(k-1)} + \textsc{Temp-SA}(x_{vid}^{(k-1)}) \\
z^{(k)} &= x_{vid}^{(k-1)} + \textsc{Spa-SA}(\hat{x}_{vid}^{(k)}) \\
&= \textsc{Space-Time}(x_{vid}^{(k-1)})
\end{aligned} \tag{1}
$$

where $x_{vid}^{(k-1)}$ is the output of the $(k-1)^{th}$ encoder layer, Temp-SA and Spa-SA represent temporal and spatial self-attention blocks, respectively. We insert multi-modal fusion inside the backbone by introducing gated cross-attention after the space-time attention module. Hence, the output of k^{th} fused TimeSformer layer, $x_{vid}^{(k)}$, can be expressed as:

$$
\begin{aligned}
z^{(k)} &= \text{SPACE-TIME}(x_{vid}^{(k-1)}) \\
x_{vid}^{(k)} &= x_{vid}^{(k-1)} + z^{(k)} + \alpha * \text{CA}(z^{(k)}, x_{text}^{(k-1)}) \\
x_{vid}^{(k)} &= x_{vid}^{(k)} + \text{FFN}(x_{vid}^{(k)})
\end{aligned} \tag{2}
$$

where $x_{text}^{(k-1)}$ is the output from the $(k-1)^{th}$ RoBERTa layer, CA, FFN denote cross-attention block and feed-forward network, respectively, and α is a learnable gating parameter initialized from 0. Each RoBERTa layer contains multi-head self-attention [84] followed by feed-forward layers. Similar to the fused TimeSformer module, we insert cross-attention into the RoBERTa backbone:

$$
\begin{aligned}
\hat{x}_{text}^{(k)} &= \text{SA}(x_{text}^{(k-1)}) \\
x_{text}^{(k)} &= x_{text}^{(k-1)} + \hat{x}_{text}^{(k)} + \alpha * \text{CA}(\hat{x}_{text}^{(k)}, x_{vid}^{(k)}) \\
x_{text}^{(k)} &= x_{text}^{(k)} + \text{FFN}(x_{text}^{(k)})
\end{aligned} \tag{3}
$$

where SA is the traditional self-attention module. For simplicity, we insert cross-attention into the same number of layers in both backbones. Notably, such *fusion in the backbone* strategy is not only limited to TimeSformer and RoBERTa; but can also be applied to any transformer-based video [54, 19, 2] and text [14, 74, 102] encoders.

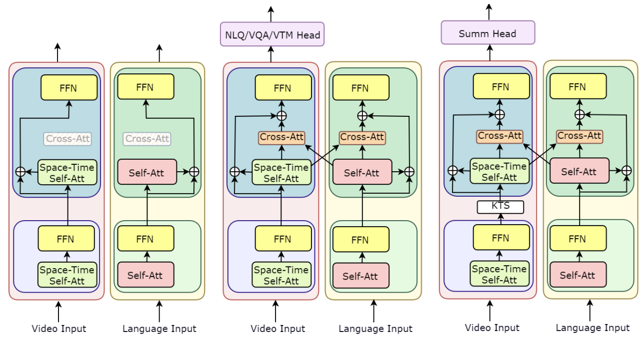
Fusion in the backbone with gated cross-attention has the following advantages: (i) Cross-attention parameters can easily be switched off by setting the gating scalar α to 0; thus, the model behaves as a dual encoder, which is helpful for scenarios that require "unfused" video and textual features; (ii) Our fusion approach is more lightweight and compute-efficient than adding fusion-specific transformer layers, which is demonstrated in detail in Section 4.5.

3.2. Pre-training Objectives

We use three pre-training objectives: (1) Egocentric noise contrastive estimation (EgoNCE), (2) masked language modeling (MLM), and (3) video-text matching (VTM).

EgoNCE: Lin et al. [50] proposed EgoNCE for dual-encoder-based egocentric VLP. It makes two modifications over InfoNCE [64]: (i) Besides the matched video-text samples, all pairs that share at least one noun or one verb are treated as positives. (ii) Every batch of N video-text samples is augmented with another N visually similar videos, which are treated as additional negatives. Overall, video-to-text EgoNCE objective, $\mathcal{L}_{\text{v2t}}^{\text{ego}}$, can be expressed as:

$$
\mathcal{L}_{\text{v2t}}^{\text{ego}} = \frac{1}{|\widetilde{\mathcal{B}}|} \sum_{i \in \widetilde{\mathcal{B}}} \log \frac{\sum_{k \in \mathcal{P}_{\text{i}}} \exp\left(\frac{\mathbf{v}_{\text{i}}^T \mathbf{t}_k}{\tau}\right)}{\sum_{j \in \mathcal{B}} \left(\exp\left(\frac{\mathbf{v}_{\text{i}}^T \mathbf{t}_{\text{j}}}{\tau}\right) + \exp\left(\frac{\mathbf{v}_{\text{i}}^T \mathbf{t}_{\text{j}'}}{\tau}\right)\right)} \tag{4}
$$



(a) Retrieval w/ Dual Encoder. (b) VQA/retrieval w/ Fusion Encoder. (c) QFVS w/ Fusion Encoder.

Figure 4: **EgoVLPv2 can be adapted to various dual- and fusion-encoder-based video-language tasks,** ranging from retrieval, video question-answering, and video grounding to query-focused video summarization.

where the i^{th} video embedding v_i and j^{th} text embedding t_j are L_2 normalized features, and τ is a temperature factor. $\widetilde{B}$ is the augmented batch with $2N$ samples. The term in brown are the modified positive samples, and the term in blue are the modified negative samples. The text-to-video EgoNCE objective, $\mathcal{L}_{\text{t2v}}^{\text{ego}}$, can be defined symmetrically. The total EgoNCE loss is: $\mathcal{L}_{\text{EgoNCE}} = \mathcal{L}_{\text{v2t}}^{\text{ego}} + \mathcal{L}_{\text{t2v}}^{\text{ego}}$.

We compute EgoNCE in a dual-encoder setting. Specifically, we set $\alpha = 0$, and thus, the cross-attention modules are switched off to calculate the EgoNCE loss.

MLM: Masked language modeling and video-text matching are proven helpful in fusion-encoder-based VLP literature [14, 112]. For MLM, we randomly mask 15% text tokens,[1] and the loss, $\mathcal{L}_{\text{MLM}}$, aims to reconstruct the masked tokens based on surrounding words and video patches by minimizing the negative log-likelihood.

VTM: For the VTM objective, the model is given a video-text sample, and the output is a binary label $y \in \{0, 1\}$ indicating if the input pair is matched. $\mathcal{L}_{\text{VTM}}$ is constructed as a binary cross-entropy loss over the predicted scores. Following [5, 16], we sample the global hard-negative video-text pairs using the similarities computed by EgoNCE.

We compute $\mathcal{L}_{\text{MLM}}$ and $\mathcal{L}_{\text{VTM}}$ in a fusion-encoder setting. In this case, $\alpha \neq 0$ and the cross-attention modules are switched on. Overall, our EgoVLPv2 pre-training pipeline can be summarized in the following three steps:

- **EgoNCE** requires unfused video and text features, so we switch off cross-attention ($\alpha = 0$). Thus, $\mathcal{L}_{\text{EgoNCE}}$ is computed with EgoVLPv2 acting as a dual encoder.
- **MLM & VTM** requires multi-modal representation. We switch on cross-attention modules and compute $\mathcal{L}_{\text{MLM}}$

[1]Following BERT, we decompose this 15% into 10% random words, 10% unchanged, and 80% with a special token [MASK].

and $\mathcal{L}_{\text{VTM}}$ with EgoVLPv2 acting as a fusion encoder.

- For **back-propagation**, the three losses are added, resulting in $\mathcal{L}_{\text{total}} = (1-\gamma-\delta)\mathcal{L}_{\text{EgoNCE}} + \gamma\mathcal{L}_{\text{MLM}} + \delta\mathcal{L}_{\text{VTM}}$, and back-propagated into the model end-to-end. γ and δ are hyper-parameters that control the contribution of different terms on $\mathcal{L}_{\text{total}}$. An ablation on different pre-training objectives of EgoVLPv2 is provided in Section 4.5. The pseudo-code for pre-training EgoVLPv2 can be found in the supplementary.

3.3. Adaptation to Downstream Tasks

We now describe how we adapt EgoVLPv2 to different downstream tasks as shown in Figure 4.

Video-Text Retrieval: We perform retrieval in two settings: (i) *dual encoders:* we switch off cross-attention and use EgoVLPv2 as a dual encoder, and compute the cosine similarity between video clips and text narrations. (ii) *fusion encoders:* we switch on cross-attention. The top M layers of the video and language backbones interact and produce multi-modal representations, which are fed into the pre-trained VTM head to compute matching scores. We also compute an ensemble of the two to further boost the performance, discussed in Section 4.5.

Video Grounding and Question Answering: We perform both uni- (video-only) and multi-modal (text-guided) video grounding. We switch off cross-attention for uni-modal grounding and use only the video encoder. We use EgoVLPv2 as a fusion encoder for text-guided grounding and video question answering.

Query-focused Video Summarization: The input videos are very long (3-5 hours) for this task. We first use the unfused $N-M$ layers[2] of our video and text encoders to extract uni-modal features from 5 second clips and the text query. Next, we apply the KTS shot boundary detector [68] to segment the long video. After this, the query and segment-wise clip features are fed into the top M fused layers of EgoVLPv2 to compute the multi-modal representation. Finally, we learn an additional single-layer transformer to design the interrelation across all 5 second long clips in every segment. We present additional details for the query-focused video summarization framework in the supplementary.

4. Experiments

4.1. Pre-training & Downstream Datasets

We pre-train EgoVLPv2 on the EgoClip [50] version of Ego4D [24], the largest publicly available egocentric video dataset. EgoClip sources untrimmed egocentric videos from Ego4D and offers filtered video-narration samples with variable-length clip intervals instead of single timestamps of Ego4D. Moreover, EgoClip excludes the videos appearing in the validation and test sets of the Ego4D benchmark [24], resulting in around 3.8M pre-training samples covering over 2927 hours of video from 129 different scenarios.

[2]For simplicity, we keep the number of unfused and fused layers the same in the video and text encoder.

Dataset	Task	Multi-modal	Fusion	Metrics (%)	Eval.
Ego4D [24]	MCQ w/ dual	✓	✗	Inter- & Intra Acc.	ZS
	MCQ w/ fusion	✓	✓	Inter- & Intra Acc.	ZS
	NLQ	✓	✓	Recall @ N	HT
	MQ	✗	–	mAP, Recall @ N	HT
QFVS [77]	Video Summ.	✓	✓	F-1	HT
EgoTaskQA [28]	Video QA	✓	✓	Mean Acc.	HT, FT
CharadesEgo [79]	CLS†	✓	✗	Video-level mAP	ZS, FT
EK-100 [13]	MIR w/ dual	✓	✗	mAP, nDCG	ZS, FT

Table 1: **Egocentric downstream datasets, metrics, and evaluation protocols.** We evaluate EgoVLPv2 on a wide variety of benchmarks: video-text retrieval (EgoMCQ, CharadesEgo, EK-100), uni-modal and text-guided video grounding (EgoMQ, EgoNLQ), video question answering (EgoTaskQA) and query-focused video summarization (QFVS). The evaluation protocols include zero-shot (ZS), task-specific head-tuning (HT), and end-to-end fine-tuning (FT). †ChardesEgo is a multi-class classification problem, but we convert this to a retrieval task. Please find more details in Section 4.1 and in supplementary.

We evaluate EgoVLPv2 across multiple benchmarks on five egocentric datasets, summarized in Table 1:

- On Ego4D [24] benchmarks: Multiple-Choice Questions (**EgoMCQ**) is a text-to-video (T $\rightarrow$ V) retrieval task with five video clips for every query text. Natural Language Query (**EgoNLQ**) is a natural language grounding [25, 22, 80] task that aims to localize a single time interval within a video given a text query. Moment Query (**EgoMQ**) is a video-only temporal action localization [8] task.

- Query-focused video summarization (**QFVS**) [77] aims to generate a concise version of a long (3-5 hours) egocentric video based on a natural language query.

- Video question-answering on **EgoTaskQA** [28] provides four question types (descriptive, predictive, explanatory, and counterfactual) with direct and indirect references, and evaluates the prediction over spatial, temporal, and causal domains of goal-oriented task understanding. Notably, to the best of our knowledge, we are the first to unify QFVS and EgoTaskQA as two downstream tasks of a VLP framework.

- Action Recognition on **CharadesEgo** [79]: a multi-class classification of daily indoor activities, with class names being short natural language phrases like '*Putting something on a shelf*.' Hence, leveraging text representations with class names, we treat this task as a retrieval problem.

Method	# Pre-train Dataset	EgoMCQ Accuracy (%) Inter	Intra	EgoNLQ validation set mIOU@0.3 R@1	R@5	mIOU@0.5 R@1	R@5
SlowFast [20]	–	–	–	5.45	10.74	3.12	6.63
EgoVLP [50]	3.8M	90.6	57.2	10.84	18.84	6.81	13.45
HierVL-Avg [3]	3.8M	90.3	53.1	–	–	–	–
HierVL-SA [3]	3.8M	90.5	52.4	–	–	–	–
LaViLa-B [110]	56M	93.8	59.9	10.53	19.13	6.69	13.68
EgoVLPv2	3.8M	**91.0**	**60.9**	**12.95**	**23.80**	**7.91**	**16.11**
$\Delta_{\text{Ours - EgoVLP}}$	–	0.4 ↑	3.7 ↑	2.11 ↑	4.96 ↑	1.10 ↑	2.66 ↑

Table 2: **Performance on EgoMCQ and EgoNLQ's validation set.** EgoVLPv2 yields significant gains over existing baselines on both tasks. LaViLa is pre-trained on 15× more narrations generated by GPT-2 [72], and is colored gray. On EgoMCQ, reported results are achieved by directly ensembling dual- and fusion-encoder-based inference.

Method	IoU=0.3 R@1	R@5	IoU=0.5 R@1	R@5	IoU=0.7 R@1	R@5	mAP (%) @ IoU 0.1	0.3	0.5	Avg.
SlowFast [20]	33.45	58.43	25.16	46.18	15.36	25.81	9.10	5.76	3.41	6.03
Frozen [4]	40.06	63.71	29.59	48.32	17.41	26.33	15.90	10.54	6.19	10.69
EgoVLP [50]	40.43	65.67	30.14	51.98	19.06	29.77	16.63	11.45	6.57	11.39
EgoVLPv2	**41.97**	**68.24**	**31.08**	**54.15**	**20.96**	**31.10**	**17.58**	**11.92**	**6.90**	**12.23**
$\Delta_{\text{Ours - EgoVLP}}$	1.54 ↑	2.57 ↑	0.94 ↑	2.17 ↑	1.90 ↑	1.33 ↑	0.95 ↑	0.47 ↑	0.33 ↑	0.84 ↑

Table 3: **Performance on EgoMQ's validation set.** EgoVLPv2 sets a new state-of-the-art across all baselines using VSGN [109] as grounding head.

- Multi-instance retrieval on Epic-Kitchens-100 [13] (**EK-100 MIR**): this is a text-to-video (T → V) and video-to-text (V → T) retrieval task, with a significant semantic overlap between different narrations. Detailed statistics of pre-training and downstream datasets and evaluation metrics are given in the supplementary.

4.2. Evaluation Protocol

We evaluate EgoVLPv2 using three evaluation protocols:

- **Zero-Shot (ZS)**. The pre-trained backbones are directly applied for V ↔ T retrieval without fine-tuning on downstream datasets. We perform zero-shot retrieval via: (i) *dual encoders,* computing the cosine similarity between video clips and textual narrations, and (ii) *fusion encoder*, incorporating the pre-trained VTM head to compute the video-text matching score.

- Task-specific **Head-tune (HT)**. We extract features using the frozen encoder and train task-specific heads[3] using the training split of downstream datasets.

- **Fine-tune (FT)**. We fine-tune the entire pre-trained video-text model end-to-end using the training split of downstream datasets.

[3]VSLNet [107] for EgoNLQ, VSGN [109] for EgoMQ, single-layer transformer encoder [84] for summarization, and linear layers for video QA.

Method	Video-1	Video-2	Video-3	Video-4	Average
SeqDPP [23]	36.59	43.67	25.26	18.15	30.92
SH-DPP [76]	35.67	42.72	36.51	18.62	33.38
QC-DPP [77]	48.68	41.66	36.51	29.96	44.19
TPAN [108]	48.74	45.30	56.51	33.64	46.05
CHAN [93]	49.14	46.53	58.65	33.42	46.94
HVN [30]	51.45	47.49	61.08	35.47	48.87
QSAN [92]	48.52	46.64	56.93	34.25	46.59
WHM [62]	50.96	48.28	58.41	**39.18**	49.20
IntentVizor [91]	51.27	53.48	61.58	37.25	50.90
EgoVLP† [50]	49.64	53.60	59.87	35.76	49.72
EgoVLPv2	**53.30**	**54.13**	**62.64**	38.25	**52.08**
$\Delta_{\text{Ours - EgoVLP}}$	3.66 ↑	0.53 ↑	2.77 ↑	2.49 ↑	2.36 ↑

Table 4: **Performance on query-focused video summarization (QFVS).** Existing baselines are trained end-to-end, whereas EgoVLPv2 only learns a tiny head on top of pre-trained encoders. †EgoVLP denotes the performance achieved by the officially released checkpoint.

4.3. Implementation Details

We use TimeSformer-B [6, 4] and RoBERTa-B [52] as our video and language backbones. The video encoder has 12 layers and 12 heads, and is configured with the patch size of 16×16 and the hidden dimension of 768. The spatial attention modules are initialized from a ViT [15]. We resize videos to 224×224 and sample 4 frames per video for pre-training and 16 for fine-tuning on downstream tasks. We use RoBERTa-B pre-trained on English Wikipedia and Toronto Book Corpus. For our best model,[4] we fuse the top 6 layers of the two encoders. We pre-train our model for 20 epochs with a batch size of 256, using AdamW [55] with a peak learning rate of 3e-5 for the backbones and 12e-5 for the cross-modal parameters. We use linear warmup over the first 2 epochs and use linear decay. Pre-training takes five days on 32 A100 GPUs. Other necessary pre-training and downstream details are given in the supplementary.

4.4. Main Results

We use **boldface** and underline for the best and second-best performing methods in every table and indicate the performance improvements over the state-of-the-art with Δ.

Ego4D: Table 2 and 3 present the performance of EgoVLPv2 on three different Ego4D benchmarks: EgoMCQ, EgoNLQ and EgoMQ. On EgoMCQ, our model achieves 91.0% inter-video and 60.9% intra-video accuracy, significantly improving over the baselines. Note that EgoVLPv2 achieves 1% absolute gain on the challenging intra-video MCQ task over LaViLa, which is trained using 15× more narrations generated by a pre-trained large language model, GPT-2 [72]. On EgoNLQ, EgoVLPv2 yields an impressive gain of 2.11% R@1 for IoU = 0.3 over EgoVLP. Moreover, using a

[4]An ablation on the number of fusion layers is provided in Section 4.5.

Method	Eval.	Direct Open	Direct Binary	Direct All	Indirect Open	Indirect Binary	Indirect All
VisualBERT [42]	FT	24.62	68.08	37.93	21.05	57.61	37.01
PSAC [44]	FT	26.97	65.95	38.90	15.31	57.75	32.72
HME [18]	FT	27.66	68.60	40.16	18.27	52.55	33.06
HGA [31]	FT	22.75	68.53	36.77	8.66	53.72	28.36
HCRN [34]	FT	30.23	69.42	42.40	27.82	59.29	41.56
ClipBERT [37]	FT	27.70	67.52	39.87	11.17	40.71	24.08
EgoVLP† [50]	FT	31.69	71.26	42.51	27.04	55.28	38.69
EgoVLPv2	FT	**35.56**	**75.60**	**46.26**	**29.14**	**59.68**	**42.28**
$\Delta_{\text{Ours - EgoVLP}}$	FT	3.87 ↑	4.34 ↑	3.75 ↑	2.10 ↑	4.40 ↑	3.59 ↑
EgoVLP† [50]	HT	20.52	64.63	32.76	16.87	48.40	29.19
EgoVLPv2	HT	**26.59**	**69.10**	**37.87**	**22.11**	**57.19**	**35.20**
$\Delta_{\text{Ours - EgoVLP}}$	HT	6.07 ↑	4.47 ↑	5.11 ↑	5.24 ↑	8.79 ↑	6.01 ↑

Table 5: **Performance on EgoTaskQA *direct* and *indirect* splits.** EgoVLPv2 outperforms prior work across all settings, metrics, and data splits. †EgoVLP denotes the performance achieved by the officially released checkpoint.

Method	Eval.	CharadesEgo mAP	Method	Eval.	EK-100 MIR mAP	EK-100 MIR nDCG
Actor [78]	FT	20.0	S3D [94]	FT	29.2	44.7
SSDA [11]	FT	23.1	MME [90]	FT	38.5	48.5
Ego-Exo [47]	FT	30.1	JPoSE [90]	FT	44.0	53.5
EgoVLP [50]	FT	32.1	EgoVLP [50]	FT	45.0	59.4
HierVL-Avg [3]	FT	32.6	HierVL-Avg [3]	FT	44.9	59.8
HierVL-SA [3]	FT	33.8	HierVL-SA [3]	FT	46.7	61.1
EgoVLPv2	FT	**34.1**	EgoVLPv2	FT	**47.3**	**61.9**
$\Delta_{\text{Ours - EgoVLP}}$	FT	2.0 ↑	$\Delta_{\text{Ours - EgoVLP}}$	FT	2.3 ↑	2.5 ↑
$\Delta_{\text{Ours - HierVL-SA}}$	FT	0.3 ↑	$\Delta_{\text{Ours - HierVL-SA}}$	FT	0.6 ↑	0.8 ↑
EgoVLP [50]	ZS	25.0	EgoVLP [50]	ZS	16.6	23.1
HierVL-Avg [3]	ZS	25.2	HierVL-Avg [3]	ZS	16.7	23.5
HierVL-SA [3]	ZS	26.0	HierVL-SA [3]	ZS	18.9	24.7
EgoVLPv2	ZS	**26.2**	EgoVLPv2	ZS	**26.7**	**29.1**
$\Delta_{\text{Ours - EgoVLP}}$	ZS	1.2 ↑	$\Delta_{\text{Ours - EgoVLP}}$	ZS	10.1 ↑	6.0 ↑
$\Delta_{\text{Ours - HierVL-SA}}$	ZS	0.2 ↑	$\Delta_{\text{Ours - HierVL-SA}}$	ZS	7.8 ↑	4.4 ↑

Table 6: **Performance on CharadesEgo and EK-100 MIR.** EgoVLPv2 achieves significant gains in fine-tuning and zero-shot settings for both tasks. Results are achieved by dual-encoder-based inference.

smaller task-specific head and fewer epochs of head-tuning, EgoVLPv2 outperforms existing baselines, which indicates the importance of learning cross-modal information during pre-training.[5] On the uni-modal grounding task, EgoMQ, our framework also sets a new state-of-the-art, outperforming EgoVLP by 1.54% R@1 for IoU = 0.3, implying the flexibility of *fusion in the backbone* over dual and shared encoder-based pre-training.

QFVS: We evaluate EgoVLPv2 on query-focused video summarization task. The QFVS dataset contains only 135 video-query training samples with long (3-5 hours) videos, and all existing baselines are trained end-to-end. In contrast, we learn a tiny head (single-layer transformer) on top of the pre-trained encoders. As shown in Table 4, our model persistently attains the state-of-the-art F-1 score across all four videos in this dataset. The pre-trained video-language representation helps EgoVLPv2 to achieve strong performance, whereas the baselines struggle to learn good cross-modal features due to the small training set.

[5] Additional details are provided in supplementary.

Fusion Strategy	# Fusion Layers	#Trainable Params.	GMACs per instance	EgoMCQ Inter	EgoMCQ Intra
Fusion in the Backbone	3	374.5M	288.62	90.5	60.0
	6	381.6M	300.16	**91.0**	**60.9**
	9	388.7M	311.71	**91.0**	**60.9**
	12	395.8M	323.26	**91.0**	**60.9**
Additional Fusion Layers	3	396.9M	402.88	90.5	60.3
	6	414.6M	437.90	90.5	60.8
	9	432.4M	472.91	90.6	60.8
	12	450.1M	507.92	90.6	**60.9**

Table 7: **Ablation study on fusion strategies.** Our proposed *fusion in the backbone* strategy performs slightly better than using fusion-specific transformer layers, but with less parameters and less compute .

EgoTaskQA: Table 5 shows the results on the egocentric video question-answering tasks on the EgoTaskQA dataset. Our model achieves significant gains across various baselines in the fine-tuning regime. Notably, EgoVLPv2 performs consistently well in the challenging *indirect* split, which demonstrates its ability to solve complicated reference tasks. In the head-tuning regime, we only learn a linear layer on top of frozen encoders, where EgoVLPv2 beats EgoVLP by a strong margin, which proves the efficacy of cross-modal pre-trained representation.

CharadesEgo: This is a multi-class action recognition task, with class names as short text phrases. We convert this to a video-to-text (V $\rightarrow$ T) retrieval problem as in CLIP [71], and perform dual-encoder-based retrieval. As shown in Table 6, EgoVLPv2 obtains a new state-of-the-art in both fine-tuning and zero-shot regimes. Since CharadesEgo videos are significantly different from Ego4D, being captured by crowd-sourced workers using mobile cameras, these results demonstrate the generalizability of EgoVLPv2.

EK-100: Table 6 shows our results on EK-100 MIR. In the fine-tuning regime, EgoVLPv2 achieves noticeable improvements over the supervised approaches (S3D, MME, JPoSE) and VLP methods (EgoVLP, HierVL). In the zero-shot setup, EgoVLPv2 beats EgoVLP and HierVL by 7.8% mAP and 4.4% nDCG scores. The consistent performance gains again show the quality of pre-trained encoders.

4.5. Ablation Study

Fusion in the Backbone: We compare our fusion module to the conventional practice of using fusion-specific transformer layers, which we implement following ALBEF [39].[6]

[6] https://github.com/salesforce/ALBEF/

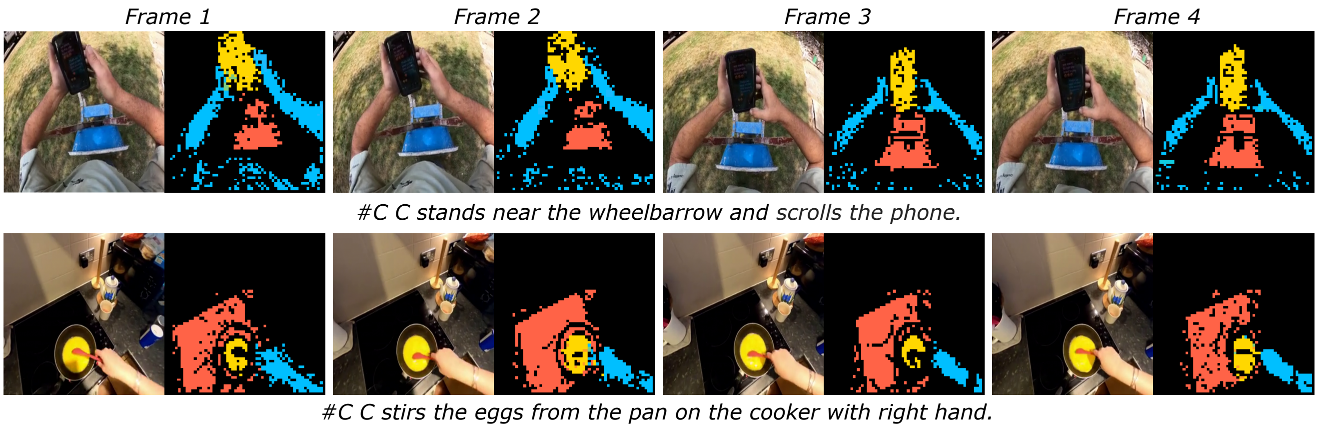


Figure 5: **Text-to-video cross-attention from multiple heads in the last layer of EgoVLPv2 with** 16×16 **patches.** We look at the attention maps of the `[CLS]` token from the text encoder on input video frames. Different heads, depicted in different colors, focus on different objects or parts. These maps show the strong cross-modal representation learned by EgoVLPv2 during pre-training, which helps to enhance performance on video-language downstream tasks.

Pre-training Objectives				**EgoMCQ (%)**					
				Dual Enc.		**Fusion Enc.**		**Ensemble**	
EgoNCE	MLM	VTM	VTM-Hard	Inter	Intra	Inter	Intra	Inter	Intra
✓	–	–	–	89.5	52.6	–	–	–	–
✓	✓	–	–	89.6	52.4	–	–	–	–
✓	–	–	✓	89.6	53.4	**90.6**	59.1	**91.0**	60.0
✓	✓	✓	–	89.5	53.6	89.1	51.5	90.2	56.8
✓	✓	–	✓	**89.8**	**56.7**	**90.6**	**59.6**	**91.0**	**60.9**

Table 8: **Ablation study on different pre-training objectives of EgoVLPv2.** We evaluate on EgoMCQ using our model either as a dual encoder, as a fusion encoder, or an ensemble of both. Removing any objective leads to a performance drop. The flexibility of the proposed fusion in the backbone module helps us boost retrieval performance using an ensembling strategy.

Table 7 shows that the proposed fusion strategy performs slightly better than stacked fusion layers. For both methods, increasing the number of fusion layers to 6 results in a non-trivial performance gain. However, our proposed architecture is significantly more parameter- and compute-efficient. For instance, with 6 fusion layers, the proposed architecture contains 33M fewer parameters and requires 45% lesser computing cost, which shows the efficacy of our method.

Pre-training Objectives: We ablate different pre-training objectives and evaluate the pre-trained models on EgoMCQ using EgoVLPv2 as a *dual* encoder, as a *fusion* encoder, and an ensemble of the two by summing their similarity scores for each video-text pair. As shown in Table 8, removing any pre-training objective lead to a performance drop. Specifically, VTM with hard-negative mining is largely beneficial across all three evaluation strategies. Fusion encoder-based evaluation brings significant improvements over dual-encoders; moreover, as EgoMCQ contains only 5 sentences for every video, both evaluation methods offer similar latency. Ensembling the two yields further $1-2\%$ performance gain for both inter- and intra-video accuracy metrics.

4.6. Attention Visualization & Error Analysis

In Figure 5, we show that different heads in the cross-modal attention can attend to different semantic regions of the video frames, guided by the narration. We observe that the pre-trained model learns well to recognize a wide variety of objects appearing in egocentric actions, such as indoor furniture, cooking appliances, phones, tablets, car steering, bicycle handles, etc. Such strong cross-modal information learned during pre-training helps EgoVLPv2 in multi-modal downstream tasks. The visualizations in Figure 5 are obtained with 960p video frames, resulting in sequences of 3601 tokens for 16×16 patches. However, vastly hindered objects in cluttered environments, especially in low-light conditions, are occasionally not focused. We show such error cases in the supplementary.

5. Conclusion

This work introduces EgoVLPv2, the second generation of egocentric video-language pre-training and a significant improvement over the previous generation [50] by incorporating cross-modal fusion directly into the video and language backbones. Our proposed *fusion in the backbone* strategy is lightweight, compute-efficient, and allows us to unify various VL tasks in a flexible and efficient manner. We conduct extensive experiments to demonstrate the effectiveness of EgoVLPv2 on a wide range of downstream tasks, consistently achieving state-of-the-art performance. Moreover, we visually demonstrate the effectiveness of the learned cross-attention representation.

Acknowledgement

The codebase for this work is built on the EgoVLP [50], LaViLa[110], FIBER [16], and VSLNet [107] repository. We would like to thank the respective authors for their contribution, and the Meta AI team for discussions and feedback. Shraman Pramanick and Rama Chellappa were partially supported by a MURI program from the Army Research Office under the grant W911NF17-1-0304.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 1, 2, 3

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1, 4

[3] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 6, 7

[4] 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. 1, 2, 3, 6

[5] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, 2022. 2, 4

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, pages 813–824. PMLR, 2021. 1, 3, 6

[7] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4613–4623, 2020. 1

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 5

[9] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. Neural storyboard artist: Visualizing stories with coherent image sequences. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2236–2244, 2019. 2

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 2

[11] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020. 7

[12] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019. 1

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 1, 3, 5, 6

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1, 2, 4

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 6

[16] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 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

[18] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 7

[19] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 1, 4

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6

[21] Jianlong Fu, Tao Mei, Kuiyuan Yang, Hanqing Lu, and Yong Rui. Tagging personal photos with transfer deep learning. In *Proceedings of the 24th International Conference on World Wide Web*, pages 344–354, 2015. 2

[22] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 5

[23] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27, 2014. 6

[24] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 3, 5

[25] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 5

[26] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1138–1147, 2021. 2

[27] Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. Unifying vision-language representation space with single-tower transformer. In *AAAI*, 2023. 2

[28] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 5

[29] 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

[30] Pin Jiang and Yahong Han. Hierarchical variational network for user-diversified & query-focused video summarization. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 202–206, 2019. 2, 6

[31] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116, 2020. 7

[32] Alec Kerrigan, Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Reformulating zero-shot action recognition for multi-label actions. *Advances in Neural Information Processing Systems*, 34:25566–25577, 2021. 1

[33] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[34] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 7

[35] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[36] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2567–2576, 2021. 2

[37] 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, 2, 3, 7

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2

[39] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 7

[40] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020. 2

[41] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022. 1, 2, 3

[42] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 7

[43] Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. Emotion reinforced visual storytelling. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 297–305, 2019. 2

[44] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019. 7

[45] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 2

[46] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023. 2

[47] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 7

[48] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 287–295, 2015. 3

[49] Chung-Ching Lin, Kevin Lin, Lijuan Wang, Zicheng Liu, and Linjie Li. Cross-modal representation learning for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19978–19988, 2022. 1

[50] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Denial Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9

[51] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding, 2023. 1

[52] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1, 3, 6

[53] 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. 1

[54] 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. 1, 4

[55] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[56] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[57] 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, 2, 3

[58] 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. 2, 3

[59] 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. 1, 3

[60] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395*, 2022. 1, 2, 3

[61] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 1

[62] Saiteja Nalla, Mohit Agrawal, Vishal Kaushal, Ganesh Ramakrishnan, and Rishabh Iyer. Watch hours in minutes: Summarizing videos with user intent. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 714–730. Springer, 2020. 2, 6

[63] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34:13988–14000, 2021. 2

[64] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

[65] Jaeyoo Park and Bohyung Han. Multi-modal representation learning with text-driven soft masks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2798–2807, 2023. 2

[66] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*, 2020. 1

[67] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *arXiv preprint arXiv:2308.07123*, 2023. 3

[68] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 540–555. Springer, 2014. 5

[69] Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik Shah, Yann LeCun, and Rama Chellappa. Volta: Vision-language transformer with weakly-supervised local-feature alignment. *arXiv preprint arXiv:2210.04135*, 2022. 2

[70] Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940, 2022. 2

[71] 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, 7

[72] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 6

[73] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1

[74] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 1, 4

[75] Anshul Shah, Benjamin Lundell, Harpreet Sawhney, and Rama Chellappa. Steps: Self-supervised key step extraction from unlabeled procedural videos. *arXiv preprint arXiv:2301.00794*, 2023. 3

[76] Aidean Sharghi, Boqing Gong, and Mubarak Shah. Query-focused extractive video summarization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 3–19. Springer, 2016. 6

[77] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4788–4797, 2017. 2, 5, 6

[78] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018. 7

[79] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 3, 5

[80] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3224–3234, 2021. 5

[81] 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. 2

[82] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3

[83] Zineng Tang, Jaemin Cho, Jie Lei, and Mohit Bansal. Perceiver-vl: Efficient vision-and-language modeling with iterative latent attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4410–4420, 2023. 1, 2, 3

[84] 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. 1, 4, 6

[85] 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. 2, 3

[86] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 1, 2, 3

[87] 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. *Transactions of Machine Learning Research*. 2

[88] Jinpeng Wang, Pan Zhou, Mike Zheng Shou, and Shuicheng Yan. Position-guided text prompt for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23242–23251, 2023. 2

[89] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 2

[90] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 450–459, 2019. 7

[91] Guande Wu, Jianzhe Lin, and Claudio T Silva. Intentvizor: Towards generic query guided interactive video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10503–10512, 2022. 2, 6

[92] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Ziyu Guan, and Deng Cai. Query-biased self-attentive network for query-

focused video summarization. *IEEE Transactions on Image Processing*, 29:5889–5899, 2020. 2, 6

[93] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. Convolutional hierarchical attention network for query-focused video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12426–12433, 2020. 2, 6

[94] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 7

[95] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 1

[96] 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. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, 2021. 1, 2, 3

[97] 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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021. 2, 3

[98] 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

[99] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 1, 2, 3

[100] 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. 1, 2, 3

[101] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 2

[102] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 1, 4

[103] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 2022*, pages 521–539. Springer, 2022. 2

[104] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 1

[105] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 1, 2, 3

[106] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2

[107] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, 2020. 6, 9

[108] Yujia Zhang, Michael Kampffmeyer, Xiaodan Liang, Min Tan, and Eric P Xing. Query-conditioned three-player adversarial network for video summarization. *British Machine Vision Conference (BMVC)*, 2018. 6

[109] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021. 6

[110] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 1, 2, 3, 6, 9

[111] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1

[112] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 1, 2, 4