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/.

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 frame-

\(^*\)Part of this work was done during an internship at Meta AI.
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.
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 where hard-negative video-narration pairs are fed into EgoVLPv2 to compute \( \mathcal{L}_{\text{VLM}} \) (iii) cross-attention is kept on, hard-negative video-narration pairs are fed into EgoVLPv2 to compute \( \mathcal{L}_{\text{VLM}} \). This fusion in the backbone strategy results in a lightweight and flexible model compared to using fusion-specific transformer layers.

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:

\[
\begin{align*}
\hat{x}^{(k)}_{\text{vid}} &= x^{(k-1)}_{\text{vid}} + \text{TEMP-SA}(x^{(k-1)}_{\text{vid}}) \\
z^{(k)} &= x^{(k-1)}_{\text{vid}} + \text{SPA-SA}(\hat{x}^{(k)}_{\text{vid}}) \\
&= \text{SPACE-TIME}(x^{(k-1)}_{\text{vid}})
\end{align*}
\]

where \( x^{(k-1)}_{\text{vid}} \) 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, \( x^{(k)}_{\text{vid}} \), can be expressed as:
where SA is the traditional self-attention module. For simplicity, we insert cross-attention into the RoBERTa backbone:

$$
\begin{align*}
    z^{(k)} &= \text{SPACE-TIME}(x^{(k-1)}_{vid}) \\
    x^{(k)}_{vid} &= x^{(k-1)}_{vid} + z^{(k)} + \alpha \cdot \text{CA}(z^{(k)}, x^{(k-1)}_{vid}) \\
    x^{(k)}_{text} &= x^{(k)}_{text} + \text{FFN}(x^{(k)}_{text})
\end{align*}
$$

where $$x^{(k-1)}_{text}$$ is the output from the $$(k-1)^{th}$$ RoBERTa layer, CA, FFN denote cross-attention block and feed-forward network, respectively, and $$\alpha$$ 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:

$$
\begin{align*}
    x^{(k)}_{text} &= \text{SA}(x^{(k-1)}_{text}) \\
    x^{(k)}_{text} &= x^{(k-1)}_{text} + z^{(k)} + \alpha \cdot \text{CA}(z^{(k)}_{text}, x^{(k)}_{text}) \\
    x^{(k)}_{text} &= x^{(k)}_{text} + \text{FFN}(x^{(k)}_{text})
\end{align*}
$$

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 $$\alpha$$ 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, $$\mathcal{L}_{\text{ego}}^{\text{v2t}}$$, can be expressed as:

$$
\mathcal{L}_{\text{ego}}^{\text{v2t}} = \frac{1}{|B|} \sum_{i \in B} \log \frac{\exp \left( \frac{\gamma^t v_i - \gamma^v t_i}{\tau} \right)}{\sum_{j \in B} \left( \exp \left( \frac{\gamma^t v_i}{\tau} \right) \right) + \sum_{j \in B} \left( \exp \left( \frac{\gamma^v t_i}{\tau} \right) \right)}
$$

where $$\gamma^t$$ and $$\gamma^v$$ are L2 normalized features, and $$\tau$$ is a temperature factor. $$B$$ 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, $$\mathcal{L}_{\text{ego}}^{\text{t2v}}$$, can be defined symmetrically. The total EgoNCE loss is: $$\mathcal{L}_{\text{EgoNCE}} = \mathcal{L}_{\text{ego}}^{\text{v2t}} + \mathcal{L}_{\text{ego}}^{\text{t2v}}$$.

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

- For **back-propagation**, the three losses are added, resulting in $L_{total} = (1 - \gamma - \delta)\mathcal{L}_{EgoNCE} + \gamma \mathcal{L}_{MLM} + \delta \mathcal{L}_{VTM}$, and back-propagated into the model end-to-end. $\gamma$ and $\delta$ are hyper-parameters that control the contribution of different terms on $L_{total}$. 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 unimodal 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 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 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 various-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.

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.

\[
\begin{array}{|l|l|l|l|l|}
\hline
\text{Dataset} & \text{Task} & \text{Multi-modal} & \text{Fusion} & \text{Metrics ($\%$)} & \text{Eval.} \\
\hline
\text{Ego4D} [24] & \text{MCQ w/ dual} & \checkmark & \checkmark & \text{Inter- & Intra Acc} & \text{ZS} \\
\text{MCQ w/ fusion} & \checkmark & \checkmark & \text{Inter- & Intra Acc} & \text{ZS} \\
\text{NLQ} & \checkmark & \checkmark & \text{Recall@N} & \text{HT} \\
\text{MQ} & \checkmark & \checkmark & \text{mAP, Recall@N} & \text{HT} \\
\text{QFVS} [77] & \text{Video Summ.} & \checkmark & \checkmark & \text{F-1} & \text{HT} \\
\text{EgoTaskQA} [28] & \text{Video QA} & \checkmark & \checkmark & \text{Mean Acc.} & \text{HT, FT} \\
\text{CharadesEgo} [79] & \text{CLS}\text{\dagger} & \checkmark & \checkmark & \text{Video-level mAP} & \text{ZS, FT} \\
\text{EK-100} [13] & \text{MIR w/ dual} & \checkmark & \checkmark & \text{mAP, nDCG} & \text{ZS, FT} \\
\hline
\end{array}
\]

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). CharadesEgo 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 using three evaluation protocols:

- **Performance on EgoMQ's validation set.**
- **Table 2:** Performance on EgoMCQ and EgoNLQ's validation set.

We fine-tune the entire pre-trained video model end-to-end using the training split of 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.

### 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 1e-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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Video-1</th>
<th>Video-2</th>
<th>Video-3</th>
<th>Video-4</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqDPP [23]</td>
<td>36.59</td>
<td>43.67</td>
<td>25.26</td>
<td>18.15</td>
<td>30.92</td>
</tr>
<tr>
<td>SH-DPP [76]</td>
<td>35.67</td>
<td>42.72</td>
<td>36.51</td>
<td>18.62</td>
<td>33.38</td>
</tr>
<tr>
<td>QC-DPP [77]</td>
<td>48.68</td>
<td>41.66</td>
<td>36.51</td>
<td>29.96</td>
<td>44.19</td>
</tr>
<tr>
<td>TPAN [108]</td>
<td>48.74</td>
<td>45.30</td>
<td>56.51</td>
<td>33.64</td>
<td>46.05</td>
</tr>
<tr>
<td>CHAN [93]</td>
<td>49.14</td>
<td>46.53</td>
<td>56.65</td>
<td>33.42</td>
<td>46.94</td>
</tr>
<tr>
<td>HVN [30]</td>
<td>51.45</td>
<td>47.49</td>
<td>61.08</td>
<td>35.47</td>
<td>48.87</td>
</tr>
<tr>
<td>QSAN [92]</td>
<td>48.52</td>
<td>46.64</td>
<td>56.93</td>
<td>34.25</td>
<td>46.59</td>
</tr>
<tr>
<td>WHM [62]</td>
<td>50.96</td>
<td>48.28</td>
<td>58.41</td>
<td>39.18</td>
<td>49.20</td>
</tr>
<tr>
<td>IntentVizor [91]</td>
<td>51.27</td>
<td>53.48</td>
<td>61.58</td>
<td>37.25</td>
<td>50.90</td>
</tr>
</tbody>
</table>

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

---

**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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-train</th>
<th>EgoMCQ</th>
<th>EgoNLQ validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Accuracy (%)</td>
<td>mIOU@0.3</td>
<td>mIOU@0.5</td>
</tr>
<tr>
<td>Inter</td>
<td>Intra</td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>EgoVLP [50]</td>
<td>3.8M</td>
<td>90.6</td>
<td>57.2</td>
</tr>
<tr>
<td>HierVL-Avg [1]</td>
<td>3.8M</td>
<td>90.3</td>
<td>53.1</td>
</tr>
<tr>
<td>HierVL-SA [1]</td>
<td>3.8M</td>
<td>90.5</td>
<td>52.4</td>
</tr>
<tr>
<td>EgoVLPv2</td>
<td>3.8M</td>
<td>91.0</td>
<td>60.9</td>
</tr>
</tbody>
</table>

| Δ | — | 0.4↑ | 3.7↑ | 2.11↑ | 4.96↑ | 1.10↑ | 2.66↑ |

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

- 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 heads3 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.

3VSLNet [107] for EgoNLQ, VSGN [109] for EgoMQ, single-layer transformer encoder [84] for summarization, and linear layers for video QA.
Table 5: Performance on EgoTaskQA and indirect splits. EgoVLPv2 outperforms prior work across all settings, metrics, and data splits. \(^\dagger\)EgoVLP denotes the performance achieved by the officially released checkpoint.

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.

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.

\[^{5}\]Additional details are provided in supplementary.

\[^{6}\]https://github.com/salesforce/ALBEF/
Figure 5: **Text-to-video cross-attention from multiple heads in the last layer of EgoVLPv2 with $16 \times 16$ patches.** We look at the attention maps of the $[\text{CLS}]$ 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.

### Table 7: 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 8 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 $16 \times 16$ 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


[91] Kun He, Zhao Wu, Zijian Zeng, Ziyu Guan, and Deng Cai. Query-biased self-attentive network for query-


