EgoVLPv2: Egocentric Video-Language Pre-training with Fusion in the Backbone (Supplementary Material)

A. Radar Chart Figure 1 Details

Here, we explain the details of the radar chart in Figure 1, which summarizes the comparative performance of EgoVLPv2 with EgoVLP [16]. First, for illustrative purposes, we normalize each axis by the score achieved by EgoVLPv2, which turns the axes in the range $(0, 1]$. Next, we keep the origin of each axis at 0.7 normalized value, which reasonably separates the inner and outer frames for better readability. Finally, we annotate each vertex with absolute performance metric scores. Notably, in most previous radar chats in the vision-language literature [26, 30], the axes have different scales and shifts, which may cause misinterpretations and fallacies. However, our illustration is uniform and accurate to scale.

B. Algorithm

The algorithm for pre-training EgoVLPv2 is given in Algorithm 1. Section 3.2 provides details of different pre-training objectives.

Algorithm 1 Pre-training EgoVLPv2

Require: Batch $\mathcal{B}_N : \{x_{vid}, x_{text}\}$
Learnable gating parameter: α
EgoVLPv2 Encoder: $\mathcal{F} : \begin{cases} \mathcal{F}_{\text{dual}} \text{ if } \alpha = 0 \\ \mathcal{F}_{\text{fused}} \text{ if } \alpha \neq 0 \end{cases}$
for $(x_{vid}, x_{text}) \in \mathcal{B}_N$ **do**
  $\mathcal{L}_{\text{EgoNCE}} \leftarrow \text{EgoNCE}(\mathcal{F}_{\text{dual}}(x_{vid}, x_{text}))$ ▷ EgoNCE
  $x^{\text{MLM}}_{text} \leftarrow Mask(x_{text})$
  $\mathcal{L}_{\text{MLM}} \leftarrow \text{MLM}(\mathcal{F}_{\text{fused}}(x_{vid}, x^{\text{MLM}}_{text}))$ ▷ MLM
  $x^{\text{VTM}}_{text} \leftarrow HardNeg(x_{text})$
  $\mathcal{L}_{\text{VTM}} \leftarrow \text{VTM}(\mathcal{F}_{\text{fused}}(x_{vid}, x^{\text{VTM}}_{text}))$ ▷ VTM
  $\mathcal{L}_{\text{total}} \leftarrow (1 - \gamma - \delta)\mathcal{L}_{\text{EgoNCE}} + \gamma\mathcal{L}_{\text{MLM}} + \delta\mathcal{L}_{\text{VTM}}$
end for
Back-prop into $\mathcal{F}$ end-to-end with $\mathcal{L}_{\text{total}}$.

C. Dataset Details

This section provides additional details of our pre-training and downstream datasets.

Ego4D & EgoClip: Ego4D [10] is the first-of-its-kind massive-scale egocentric video-language dataset and benchmark suite. It offers 3670 hours of daily life activity videos captured by 931 unique camera wearers from 74 worldwide locations and 9 different countries. The videos in Ego4D span hundreds of scenarios (kitchen, laboratory, workshop, porch, shopping, driving, leisure, etc.) with various daytime and weather conditions. A portion of the dataset is accompanied by audio, 3D meshes of the environment, eye gaze, stereo, and synchronized videos from multiple egocentric cameras at the same event. Each narration in Ego4D is a free-form sentence and has a single timestamp. For example, the narration "`#C C walks towards a laundry machine`" is associated with the video content, which occurs at 28.3s of a particular video. However, an activity occurs for a certain duration, and such a single timestamp can not reflect the start and end points where the particular activity takes place. EgoClip [16] offers a filtered version of Ego4D and designs a contextual variable-length clip pairing strategy to assign every narration with start and end timestamps. Moreover, EgoClip excludes videos that belong to the validation and test sets of the Ego4D benchmark challenges and retains textual annotation from multiple narrators, allowing us to have narration diversity during pre-training. Overall, EgoClip contains 2927 hours of videos which form 3.8M clip-text pairs, with an average clip length of 1.0s and a standard deviation of 0.9s. We use this EgoClip version of Ego4D for pre-training. We evaluate EgoVLPv2 on three different downstream benchmarks of Ego4D: multiple-choice questions (EgoMCQ), natural language query (EgoNLQ), and moment query (EgoMQ).

QFVS: The query-focused video summarization (QFVS) [23] dataset builds upon previously existing UT egocentric (UTE) [15] dataset, which contains four 3-5 hours long videos captured in uncontrolled everyday scenarios. QFVS curates 46 queries for every video, where each query contains two distinct concepts (nouns) [29, 22, 4]. For example, a query can be {HAT, PHONE}, or {FOOD, DRINK}. These 46 queries cover four distinct scenarios: (i) both the concepts appear in the same video shot (15 such queries),[1]

[1] QFVS defines every consecutive 5s video clip as a shot.

(ii) the concepts appear in the video, but not in a single shot (15 such queries), (iii) only one concept appears in the video (15 such queries), and (iv) none of the concepts in the query are present in the video (1 such query). We use prompt engineering to generate natural language using the concepts in the query and feed the sentence in our model. For instance, a given query {HAT, PHONE} is converted as "*All scenes containing hats and phones*". We use 10 different prompts during head-tuning. The QFVS dataset also annotates concepts for every video shot. It proposes a robust evaluation strategy: find the similarity between the concepts in the generated and ground truth summary by maximum weight matching of a bipartite graph, and compute precision, recall, and F1 score from the number of matched concepts. This evaluation strategy helps to capture how well a system summary can retain semantic information instead of visual quantities, as used in previously existing evaluation methods, such as a system-generated summary has to consist of the same key units (frame or shot) as in the user summary [5, 25, 28] or comparing pixels and low-level features [9, 13, 14, 32, 34].

EgoTaskQA: The EgoTaskQA [12] benchmark uses the same egocentric videos as the LEMMA dataset [11], which contains goal-oriented and multi-tasked human activities with rich human-object interactions and action dependencies in both single-agent and two-agent collaboration scenarios. The videos are segmented into clips with an average duration of 25s. The questions in the EgoTaskQA dataset are machine-generated and aim to evaluate models' capabilities to describe, explain, anticipate, and make counterfactual predictions about goal-oriented events. The answers are of two types - open-answer queries and binary statement verifications. The EgoTaskQA dataset contains 40K balanced question-answer pairs selected from 368K programmatically generated questions from 2K egocentric videos. Moreover, this dataset offers two different benchmark splits (i) *normal* or *direct* split where the train, test, and validation sets are randomly sampled in a 3:1:1 ratio and (ii) *indirect* split where the actions and objects are strongly correlated and test the model's task understanding capability with challenging questions. We approach the video QA as a classification task and report accuracy for open queries and binary verification in the direct and indirect splits.

CharadesEgo: The CharadesEgo [24] dataset consists of 68.5K annotated samples from 7860 videos from both first and third-person views, covering 157 classes of daily indoor activities. We only use the first-person subset, which contains 3085 videos for training and 846 videos for testing. ChardesEgo is originally a multi-class classification problem, with class labels being short phrases like '*Putting something on the shelf.*' We treat this problem to a video-to-text (V $\rightarrow$ T) retrieval task as in CLIP [21] by leveraging the text encoder to extract features from class names. We directly evaluate the model on the validation set in the zero-shot setting. In the fine-tuning setting, we leverage the 33.1K training samples to perform an end-to-end fine-tuning of EgoVLPv2. Following the previous literature [16, 36, 1], we report video-level mAP as the evaluation metric.

EK-100: The Epic-Kitchens-100 [6] dataset contains 100 hours of egocentric cooking videos. The training set consists of 67.2K video samples, whereas the validation and test set has 9.6K and 13.1K samples, respectively. Each sample is associated with text narration. We perform multi-instance retrieval (V $\leftrightarrow$ T) on the EK-100 dataset, which is challenging due to the significant semantic overlap between different narrations. The evaluation metrics are mean Average Precision (mAP) and the normalized Discounted Cumulative Gain (nDCG).

D. Implementation Details

D.1. Pre-training on EgoClip

Table D.1 presents the hyper-parameters used during pre-training. We use TimeSformer-B [3, 2] and RoBERTa-B [17] as our video and language backbones. We chose the best learning rate using a grid search. We ablate our other design choices in Section E. We use PyTorch's native FP16 mixed precision training and gradient checkpoint during pre-training.

After every epoch, we validate the pre-trained checkpoint on EgoMCQ and select the model with the best EgoMCQ intra-video score for other downstream tasks. We extract 4 frames for every video sample during pre-training and reshape those to 224 $\times$ 224. We also apply standard `RandomResizedCrop`, `RandomHorizontalFlip`, `ColorJitter` and normalization to every frame. We tokenize the text using RoBERTa tokenizer and pad/truncate every narration to a maximum length of 30. Pre-training takes five days on 32 A100 GPUs.

D.2. Downstream Settings

This section presents our fine-tuning and head-tuning strategy for different downstream tasks. For a fair comparison with the baselines [16, 36, 1], we follow the same downstream configuration as the baselines when possible. The downstream is performed with 16 frames per video sample.

EgoNLQ: This task is a video-text localization problem, with each video clip longing up to 1200s. Hence, performing end-to-end fine-tuning can be hard on EgoNLQ. Following [16, 36], we pre-extract features from the video-text samples using our pre-trained model and train VSLNet [31] for 100 epochs, with a learning rate of 1e$-$3 and batch size of 32. We keep all other configurations the same as [16].[2] However, we

[2] https://github.com/showlab/EgoVLP

Hyper-parameters	Notation	Value
Model		
Video encoder	–	TimeSFormer-B [3, 2]
Text encoder	–	`roberta-base` [17]
Video & text embedding	–	768
Video encoder patch size	–	16×16
Video & text projector	–	4096-4096-4096
# Fusion layers	–	6
Pre-training		
Batch size	–	256
Epochs	–	20
Number of frames	–	4
Frame resolution	–	224×224
Vocab size	–	50265
MLM prob.	–	0.15
Max. length of text	–	30
Temp. in Equation 4	τ	0.05
MLM & VTM loss weights	γ, δ	0.25, 0.5
Optimizer	–	AdamW [18]
Peak LR for backbones	–	3e−5
Peak LR for cross-att	–	12e−5
Peak LR for loss heads	–	12e−5
Warmup	–	Linear (first 2 epochs)
LR decay	–	Linear
End LR	–	1e−7
Betas in AdamW	(β_1, β_2)	(0.9, 0.98)
Eps in AdamW	–	1e−8
Weight decay	–	1e−2

Table D.1: **Pre-training hyper-parameter details of EgoVLPv2.**

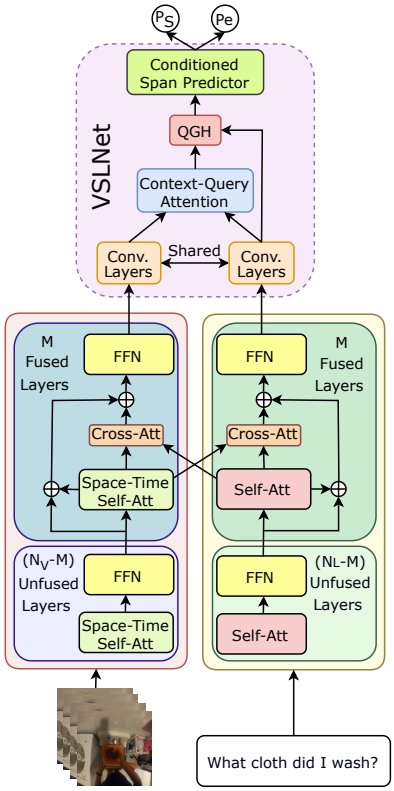


Figure D.1: **Entire pipeline for EgoNLQ.** Following EgoVLP [16] and LAVILA [36], we pre-extract video-text features using pre-trained EgoVLPv2, and train VSLNet [31] on top of frozen encoders.

observe that we can beat the baselines using even a smaller task head and fewer epochs of tuning, which we describe in Section F. We show the complete EgoNLQ pipeline in Figure D.1.

EgoMQ: This is a video-only localization problem, and similar to EgoNLQ, the input videos are very long. Hence, end-to-end fine-tuning is also hard to perform on EgoMQ. Following EgoVLP [16], we pre-extract video features using pre-trained EgoVLPv2 and train VSGN [35] for 100 epochs, with a learning rate of 1e−4 and batch size of 32. We keep all other configurations the same as [16]. We perform a grid search for other hyper-parameters of VSGN.

QFVS: Query-focused video summarization aims to generate an abridged version of input video guided by a natural language query. To the best of our knowledge, we are the first to unify QFVS as a downstream of a VLP framework. The input videos for this task are very long (3-5 hours). We first use the unfused $N - M$ layers[3] of our video and text encoders to extract uni-modal features from every 5-second clip and the text query. Next, we apply the KTS shot boundary detector [20] to segment the long video.[4] 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 train the single-layer transformer for 20 epochs, with a batch size of 20, a peak learning rate of 1e−5 using AdamW [18] optimizer, cosine scheduler, and a linear warmup for the first 2 epochs. We also perform an ablation on the single-layer transformer in Section F.

EgoTaskQA: We treat the video QA as a classification problem, where we train linear layers on top of the fused feature representation generated by the pre-trained EgoVLPv2. In the fine-tuning setting, we fine-tune the pre-trained model for 36 epochs with a batch size of 64, using the AdamW [18] optimizer. We use cosine annealing with 10% linear warmup steps, with the peak learning rate of 2e−4 for the direct split and 1e−4 for the indirect split. In the head-tuning setup, we only train the classifier head on top of frozen backbones with the same configuration.

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

[4]Segmentation helps in two ways: (i) TimeSformer can not process the whole 3-5 hours long video (containing tens of thousands of frames) at once. (ii) Segmentation is also used to convert frame-level prediction scores into key shots. For details, please refer to [23, 7, 33].

Pre-training Objectives	EgoNCE Sampling		EgoMCQ (%)	
	Pos.	Neg.	Inter	Intra
InfoNCE + MLM + VTM	–	–	90.0	55.2
EgoNCE + MLM + VTM	✓	✗	90.4	58.8
EgoNCE + MLM + VTM	✗	✓	90.5	59.1
EgoNCE + MLM + VTM	✓	✓	**91.0**	**60.9**

Table E.1: **Ablation on EgoNCE sampling strategy.** EgoNCE [16] helps in improving the performance significantly compared to InfoNCE [19]. We also observe that both the positive and negative sampling of EgoNCE is important, and removing any of those leads to a performance drop.

Cross-Att	EgoMCQ (%)	
	Inter	Intra
$\alpha = 0.1$	90.1	59.8
$\alpha = 0.25$	90.4	59.9
$\alpha = 0.5$	90.1	58.0
$\alpha = 1$	89.4	56.9
Learnable α	**91.0**	**60.9**

Table E.2: **Ablation on the gated cross-attention.** Learnable gating scaler α performs better than a fixed value.

CharadesEgo: Following [16, 36, 1], we convert CharadesEgo as a retrieval problem. In the zero-shot setup, we perform dual-encoder-based inference. In the fine-tuning setup, we use EgoNCE as our objective. We fine-tune the model for 10 epochs with a batch size of 128 using AdamW [18] optimizer with $(\beta_1, \beta_2) = (0.9, 0.98)$, and weight decay of 0.01. We use cosine annealing with warmup, with 10% linear warmup steps, peak learning rate of 1.5e$-$4 and end learning rate of 1e$-$7. Since this is a multi-class dataset, where each video can include multiple actions, we report mAP as the evaluation metric. For input, we sample 16 frames from each video clip, and reshape the frames into 224×224.

EK-100 MIR: Since a narration can jointly be associated with multiple videos for EK-100 multi-instance retrieval task, we use the adaptive multi-instance max-margin loss [27] for this task with a margin value of 0.2. We keep the zero-shot configuration the same as CharadesEgo. We fine-tune the model for 100 epochs with a batch size of 128 using AdamW [18] optimizer with $(\beta_1, \beta_2) = (0.9, 0.98)$, and weight decay of 0.01. We use cosine annealing with warmup, with 10% linear warmup steps, peak learning rate of 2e$-$4 and end learning rate of 1e$-$7.

E. Additional Ablations on Pre-training

We conduct additional ablation experiments in this section to validate our design choices. Reported results on EgoMCQ in Table E.1, E.2, E.3 and Figure E.1 are achieved by directly ensembling dual- and fusion-encoder-based inference.

Effect of EgoNCE: We study the effect of the EgoNCE loss [16] compared to the more popular InfoNCE objective [19]. Given a batch of N video-text pairs, InfoNCE treats the matched N pairs as positives and every other pair as negatives. However, egocentric videos pose two unique challenges: (i) Same actions in different scenarios appear to be visually different (*talking on the phone indoors* and *outdoors*). (ii) Different actions in same scenarios appear to be similar (*writing on a tablet* and *watching a movie on a tablet* are visually indistinguishable). To overcome these challenges, EgoNCE is built upon InfoNCE with two modifications: (i) Besides the matched video-text samples in every batch, all narration pairs which share at least one noun and one verb are treated as positives. (ii) Every batch of N video-text pairs is augmented with another N visually similar videos, often containing different actions in the same scenarios. These added videos with the same texts as in the original batch are treated as additional negatives.

Table E.1 shows the effect of the modified positive and negative sampling of EgoNCE on EgoVLPv2. First, we observe that replacing EgoNCE with InfoNCE leads to a performance drop of 5.7% accuracy on the challenging intra-video metric of EgoMCQ. Further, discarding either positive or negative sampling also drops the results by 2.1-1.8% intra-video accuracy. These results align with the findings in [16] and indicate the efficacy of the EgoNCE objective for egocentric video-language pre-training.

Effect of Gated Cross-attention: Next, we study the importance of gated cross-attention modules with learnable gating scalar, α. Table E.2 shows that a fixed value of α leads to a significant performance drop. In our best pre-trained model, we also find that the learned value of α varies in different layers, ranging from 0.05 to 0.4.

Effect of Projector: We compare different choices of projector dimensions used in the EgoNCE head in Figure E.1. We observe that a three-layer projector works better than single and two-layer projectors. For instance, a 4096-4096-4096 dimensional projector improves the EgoMCQ intra-video retrieval performance by 0.85% over a single 4096 dimensional projector. Moreover, an increase in the width of the projector also helps in performance. Hence, we use 4096-4096-4096 as our default projector. Notably, these results oppose the findings in Zhao et al. [36], where the authors observe that using 256-dimension achieves better performance than a 512 dimensional projector. The reason behind such results is, in contrast to Zhao et al., [36], who only use InfoNCE, a larger projector helps us both in EgoNCE and VTM objectives by offering a stronger hard-negative sampling.

Effect of Batch Size: Next, we study the effect of pre-training batch size in Table E.3a. The performance improves using a batch size of 256 over 128. However, the perfor-

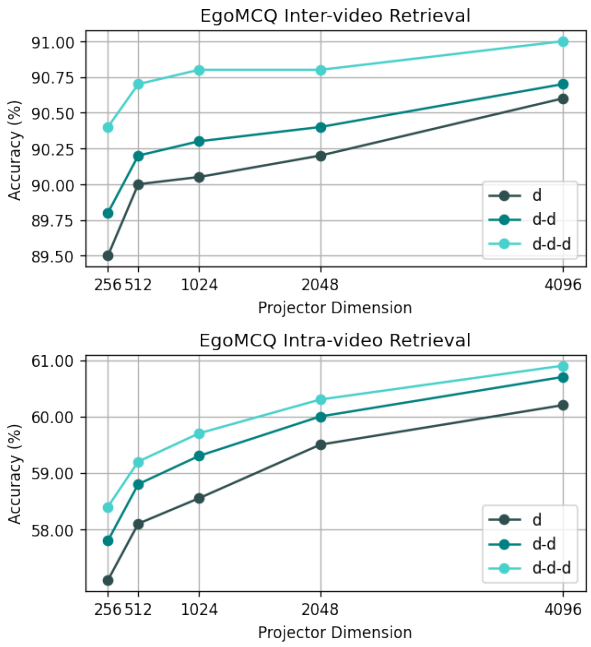


Figure E.1: **Ablation on the projector dimension used in the EgoNCE head.** A 3-layer projector works better than a single-layer projector. Moreover, an increase in the width of the projector also helps in performance.

Batch Size	EgoMCQ (%) Inter	Intra
128	90.6	59.8
256	**91.0**	**60.9**
512	**91.0**	60.6
1024	90.8	60.5

(a) **Ablation on batch size.** EgoMCQ performance is best with a batch size of 256.

# Frames (Pre-training)	EgoMCQ (%) Inter	Intra
2	90.1	56.7
4	91.0	60.9
5	91.2	61.2
6	91.4	61.5

(b) **Ablation on number of frames.** Increasing frames improve EgoMCQ performance.

Table E.3: **Ablation on pre-training batch size (a) and the number of frames (b).** A batch size of 256 produces the best results. Increasing the number of frames helps in a performance gain. For a fair comparison with the baselines [16, 36, 1], we keep 4 as our default frame number.

mance drops if we further increase the batch size to 512 or 1024. Therefore, we use 256 as our default batch size in all other experiments.

Effect of Number of Frames: Lastly, we ablate the number of frames per sample during pre-training in Table E.3b. We see a good improvement in the EgoMCQ performance when the number of frames is increased to 4. However, after 4, the performance improvement diminishes. We keep 4 as our default frame number for a fair comparison with the baselines [16, 36, 1], who also use 4 frames per sample during pre-training.

Model + Task head	EgoNLQ validation set mIOU@0.3 R@1	R@5	mIOU@0.5 R@1	R@5
SlowFast [8] + VSLNet [31]	5.45	10.74	3.12	6.63
EgoVLP [16] + VSLNet [31]	10.84	18.84	6.81	13.45
LAVILA[36] + VSLNet [31]	10.53	19.13	6.69	13.68
EgoVLPv2 + Span	11.08	21.27	7.05	14.29
EgoVLPv2 + QGH + Span	11.95	22.86	7.64	15.80
EgoVLPv2 + VSLNet [31]	**12.95**	**23.80**	**7.91**	**16.11**

Table F.1: **Ablation on task-head for EgoNLQ.** EgoVLPv2 beats existing models even using a smaller task-head.

Model + Task head	Video-1	Video-2	Video-3	Video-4	Average
EgoVLPv2 + Linear layers	50.17	50.95	59.38	34.58	48.77
EgoVLPv2 + 1-layer transformer	**54.97**	**55.74**	**64.10**	**40.83**	**53.91**
EgoVLPv2 + 2-layer transformer	52.78	51.98	66.80	34.10	51.42
EgoVLPv2 + 3-layer transformer	51.87	52.45	63.75	35.55	50.91

Table F.2: **Ablation on task-head for QFVS.** A single-layer transformer produces better performance than linear layers and multi-layer transformers.

F. Ablations on Downstream

This section presents an ablation on downstream task-specific heads for EgoNLQ and QFVS.

EgoNLQ: Following EgoVLP [16] and LAVILA [36], we use VSLNet [31] as the task-head for EgoNLQ. However, since our model learns cross-modal features during pre-training, we observe that we can beat the previous methods by a significant margin even using smaller task heads. As shown in Table F.1, when we only use the conditional span predictor module, which is just a linear layer, we can beat EgoVLP by 2.43% R@5 for IoU=0.3. Adding the QGH module further helps in improving the performance. Using the whole VSLNet can significantly beat EgoVLP and LAVILA across all metrics. Moreover, the previous methods train VSLNet for 200 epochs, whereas we achieve the best performance within 100 epochs. These results prove the efficacy of the cross-modal pre-trained representation of EgoVLPv2.

QFVS: Next, we compare different heads for QFVS in Table F.2. Notably, this dataset is very small, with only 135 training samples. We observe that a single-layer transformer head performs better than linear layers and multi-layer transformers. Linear layers can not model temporal relations across different video shots, which a transformer can efficiently do. However, multi-layer transformers overfit this dataset due to the small training set. Hence, we use a single-layer transformer for QFVS.

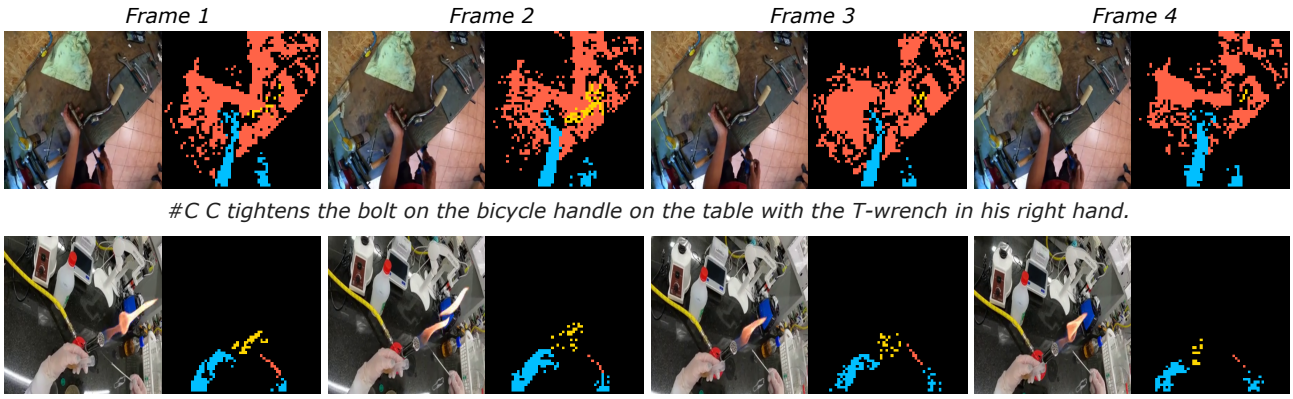


Figure G.1: **Limitations of our method:** tiny and hindered objects in cluttered environments are not distinctly attended by the pre-trained EgoVLPv2. We show the attention maps of the `[CLS]` token from the text encoder on input video frames in the text-to-video cross-attention module of the last layer of EgoVLPv2. Different heads, shown in different colors, focus on various semantic regions of the video frames. The visualizations are obtained with 960p video frames, resulting in sequences of 3601 tokens for 16×16 patches.

G. Error Analysis

Although EgoVLPv2 learns impressive cross-modal representation during pre-training, there are still some cases where the model fails to identify tiny and hindered objects, especially in cluttered environments. We show two such examples in Figure G.1. In the first video, the objects 'bicycle handle' and 'T-wrench' are barely visible even in human eyes, and thus, EgoVLPv2 can not consistently attend to these objects in all frames. However, it can recognize larger, more familiar things like tables and human hands. In the second video, we show an egocentric activity in a wet lab, where the camera wearer is wearing gloves, holding a test tube, and heating a wire using a bunsen burner. This is a complex scenario with multi-agent collaborative activities and fine-grained actions. Interestingly, EgoVLPv2 can correctly identify the human hands and track the motion of the thumb in different frames, even when wearing gloves. However, the test tube and the wire are hindered and are partially attended by the model. Since we pre-train EgoVLPv2 with 224×224 video frames, such tiny objects are often hard to be distinguished. However, higher-resolution frames will be more helpful in addressing such intricate scenarios, which we plan to explore in future works.

H. Qualitative Downstream Performance

EgoMCQ: In Figure H.1, we show example predictions made by EgoVLP [16] and EgoVLPv2 on multiple choice questions from EgoMCQ validation set. EgoVLPv2 beats EgoVLP substantially on the challenging intra-video setting, where all 5 choices are visually similar. The VTM head pre-trained with hard-negative sampling helps EgoVLPv2 to distinguish between similar videos and boosts the performance over EgoVLP.

QFVS: Figure H.2 shows some examples of query-focused summaries generated by EgoVLPv2 on the QFVS dataset. Given a long egocentric video and a natural language query, our model can summarize all relevant scenes successfully. Notably, the input videos on this dataset are very long (3-5 hours), and the length of the generated summary is 2% input video, which makes this task challenging.

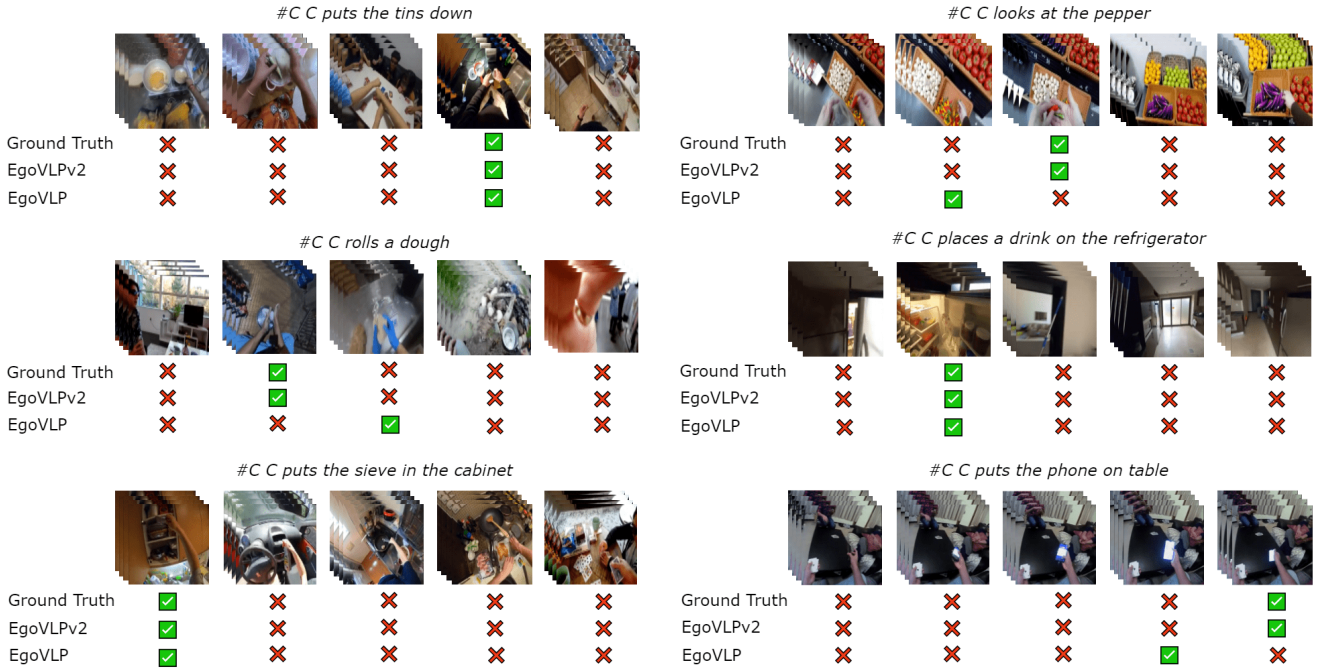
EgoNLQ: Figure H.3 shows examples of predictions made by EgoVLP [16] and EgoVLPv2 on text-guided video localization from the EgoNLQ dataset. Given an untrimmed video and a natural language query, this task aims to predict a single temporal window to answer the query. The predictions of EgoVLPv2 are significantly more aligned with the ground truth than EgoVLP, which supports the impressive quantitative performance gain by EgoVLPv2 over EgoVLP across all metrics.

References

[1] 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. 2, 4, 5

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 3

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?



(a) Inter-video MCQ.

(b) Intra-video MCQ.

Figure H.1: **Examples of predictions made by EgoVLP [16] and EgoVLPv2 on multiple choice questions from EgoMCQ validation set.** *Left:* The "inter-video" setting, each question contains 5 clips from different videos. *Right:* The "intra-video" setting, each question contains 5 contiguous clips from the same video, making it more challenging.

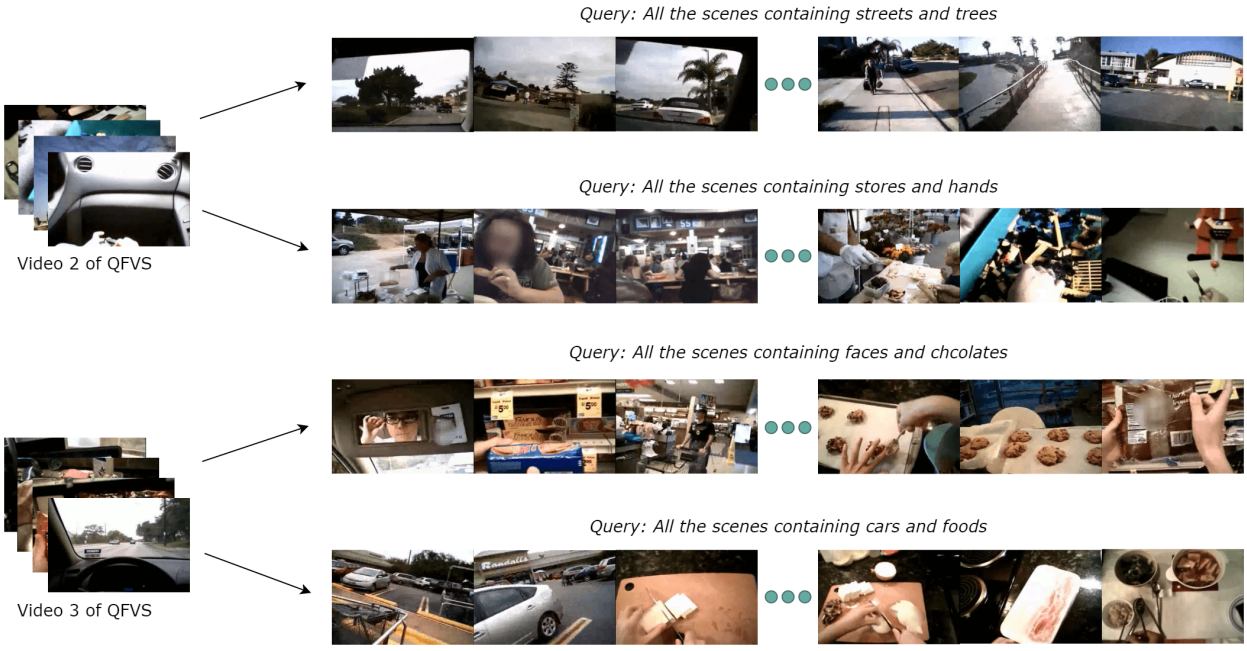


Figure H.2: **Examples of query-focused video summary generated by EgoVLPv2 on the QFVS daatset.** Given a long egocentric video and a natural language query, the generated summary includes all relevant scenes. For example, the query "All the scenes containing streets and trees" summarizes the scenes containing streets and trees in the long input video.

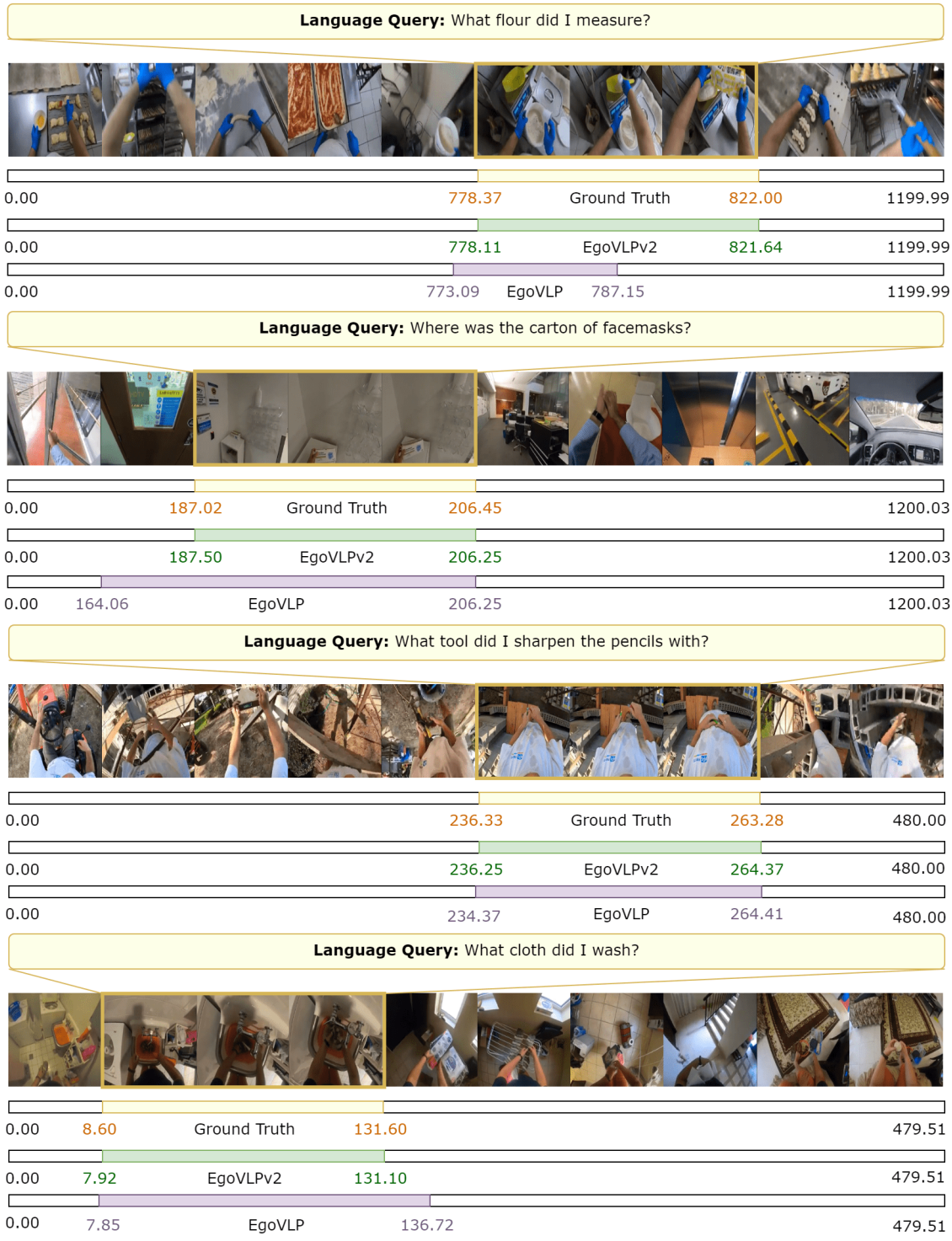


Figure H.3: **Examples of predictions made by EgoVLP [16] and EgoVLPv2 on text-guided video localization from the EgoNLQ dataset.** Given an untrimmed video and a language query, the prediction is a single temporal window containing the answer to the query. The predictions of EgoVLPv2 are significantly more aligned with the ground truth than EgoVLP.

In *International Conference on Machine Learning*, pages 813–824. PMLR, 2021. 2, 3

[4] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 459–460, 2013. 1

[5] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3584–3592, 2015. 2

[6] 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. 2

[7] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*, pages 39–54. Springer, 2019. 3

[8] 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. 5

[9] 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. 2

[10] 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

[11] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 767–786. Springer, 2020. 2

[12] 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. 2

[13] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2698–2705, 2013. 2

[14] Gunhee Kim, Leonid Sigal, and Eric P Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4225–4232, 2014. 2

[15] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012. 1

[16] 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

[17] 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. 2, 3

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 3, 4

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

[20] 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. 3

[21] 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

[22] 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. 1

[23] 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. 1, 3

[24] 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. 2

[25] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 2

[26] 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 all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1

[27] 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. 4

[28] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2244, 2015. 2

[29] Serena Yeung, Alireza Fathi, and Li Fei-Fei. Videoset: Video summary evaluation through text. *arXiv preprint arXiv:1406.5824*, 2014. 1

[30] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[31] 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. 2, 3, 5

[32] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1059–1067, 2016. 2

[33] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 766–782. Springer, 2016. 3

[34] Bin Zhao and Eric P Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2513–2520, 2014. 2

[35] 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. 3

[36] 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. 2, 3, 4, 5