The Appendix of Dynamic Cooperative Network Empowers LLMs with Video Understanding

Abstract of Appendix

This appendix provides the implementation of redundancy estimation (Appendix A), additional discussions (Appendix B), more implementation details (Appendix C), more visualization results (Appendix D), and case study (Appendix E).

A. Redundancy Estimation

In this section, we provide the details about how to estimate the ratio of temporal repetitive frames and answer-irrelevant frames, which is denoted as r_d and r_a, respectively. Specifically, given a T-frame video, we use CLIP-ViT [19] to extract the representation for each video frame and its text part.

For temporal repetitive frames, we calculate the cosine similarities of features between consecutive frames, denoted as $s_t = \cos(\mathbf{f}_t, \mathbf{f}_{t+1})$. We then collect all scores into a score vector $\mathbf{s} = \{s_t\}_{t=1}^{T-1}$ and apply min-max normalization. This process can be summarized as,

$$r_d = \frac{\sum_{t=1}^{T-1} \mathbf{I}(s_t > 0.6)}{T-1}, \tag{1}$$

where $\mathbf{I}(\cdot)$ is the indicator function, defined as $\mathbf{I}(\mathbf{x}) = 1$ if $\mathbf{x}$ is true, and $\mathbf{I}(\mathbf{x}) = 0$ if $\mathbf{x}$ is false.

For answer-irrelevant frames, we compute their similarity using $s_t = \cos(\mathbf{f}_t, \mathbf{q}||\mathbf{a})$, where $\mathbf{q}||\mathbf{a}$ represents the token-wise concatenation of question and answer feature. After applying min-max normalization, we mark a frame as redundant when its frame-to-text similarity falls below a certain threshold. This is summarized as,

$$r_a = \frac{\sum_{t=1}^{T} \mathbf{I}(s_t < 0.4)}{T}. \tag{2}$$

Notably, for each benchmark dataset, we randomly sample 20 videos to calculate the average value of redundancy ratio r_d and r_a as a rough redundancy estimation.

B. Additional Discussions

B.1. Component-wise Training State on Model Performance

We conduct the extensive experiment to explore the effect of different components with different training state. As can be seen in Table 1, only locking LLM and DPE+CCE module in the first stage exhibits the best, which achieves a obvious performance gain of 0.16 on VCG-Bench. This can be explained that DPE+CCE† primarily undertakes the effective feature encoding, whereas the projector $\mathcal{F}_{\text{fine}}, \mathcal{F}_{\text{coarse}}$ may be only responsible for bridging the semantic gap between video content and LLM, respectively. Therefore, the learned knowledge preserved in DPE+CCE† in the first stage may not be well adapted to learning of the second stage. In the second stage, unlocking DPE+CCE† achieves the substantial performance gain. This may be due to that the knowledge learned in the second stage focuses on video reasoning (for example, which part need to be focused?), which keeps consistent with the design motivation of DPE+CCE†.

Vision-Language Alignment			Instruction Tuning			**MSVD-QA**		**VCG-Bench**
DPE+CCE†	$\mathcal{F}_{\text{fine}}, \mathcal{F}_{\text{coarse}}$	LLM	DPE+CCE†	$\mathcal{F}_{\text{fine}}, \mathcal{F}_{\text{coarse}}$	LLM	Acc	Score	Score
🔓	🔓	🔒	🔓	🔓	🔓	65.45	3.56	2.65
🔒	🔒	🔒	🔒	🔓	🔓	61.07	3.20	2.31
🔓	🔓	🔒	🔒	🔓	🔓	62.21	3.34	2.38
🔒	🔓	🔒	🔓	🔓	🔓	67.90	3.72	2.81

Table 1. Performance Comparisons with training state for different components, which is only pretrained and fine-tuned with video dataset. 🔒 indicates parameters are frozen while 🔓 denotes the trainable state. DPE+CCE† denotes the DPE module and CCE module without $\mathcal{F}_{\text{fine}}, \mathcal{F}_{\text{coarse}}$.

B.2. Parameter, Runtime and Memory Complexity

Training Time. Table 2 reports the training hours on 8 A100 GPU w/ and w/o the added modules (CCE and DPE). Notably, the model without DPE+CCE refers to that we represents each video frame with two only tokens similar to LLaMA-VID, whereas the model with DPE+CCE additionally generates the finer tokens for important video frames. The increased training time probably comes from the computation time of the extra tokens in LLM backbone, rather than the actual computation time in DPE+CCE module.

Model	Stage1 (PT)	Stage2 (SFT)	Total
w/o DPE+CCE	5.85	19.63	25.48
w DPE+CCE	7.75	25.35	33.10

Table 2. Comparison on training hour of methods without DPE+CCE and with DPE+CCE.

Computation Complexity. Table 3 reports the inference cost of each added components on LVBench with 1000 input frames on one A100 GPU. The calculated event prototypes correspond to T-DPC, the filtered event prototypes correspond to Dyn. Select., multi-grained spatial object prototypes correspond to S-DPC, and Dyn. Enc. corresponds to *Cones* and *Rods* as depicted in CCE. The S-DPC and T-DPC modules do not have trainable parameters.

Modules		Inference GFLOPs	Param. (M)	Inference Latency (ms)
CCE	S-DPC	0.00	0.00	50.84
	Dyn. Enc.	112.16	30.31	15.96
DPE	T-DPC	0.00	0.00	608.28
	Dyn. Select.	12.91	11.87	3.02

Table 3. Ablative analysis on computation efficiency of added modules.

Parameter Budget. The additional parameter introduced by our designed modules compared with LLaMA-VID are listed in follows:

(a) DPE module: (1) Dynamic Selection (Three MLPs): $\left[d, \frac{d}{2}\right] \rightarrow \left[\frac{d}{2}, \frac{d}{4}\right] \rightarrow \left[\frac{d}{4}, 1\right]$.

(b) CCE module: (1) CA module (Two MLPs): $[d, d]$, $[d, d]$; (2) $\mathcal{F}_{\text{coarse}}$ and $\mathcal{F}_{\text{fine}}$ (Two MLPs): $[d, d]$, $[d, d]$
Inference Latency with other baselines. As shown in Table 4, we showcase the comparison of image resolution, averaged inference latency, and input strategies when training. Notably, we achieve the comparable computational efficiency with LLaMA-VID.

Methods	Res.	Inference Latency (s) ↓			Training Setting
		MSVD	ANet-QA	VideoMME	
LLaMA-VID [13] [ECCV 24]	224^2	1.3	3.8	6.3	1 fps
Flash-Vstream [24]	224^2	1.7	6.9	8.2	1 fps
DynFocus ($L = 25, K/L = 0.8$)	224^2	1.4	6.4	7.8	1 fps

Table 4. Comparison on image resolution, average inference latency, and input strategies when training.

B.3. Comparison of Method Design with other Methods.

In this section, we compare the design details with two closely related studies: LLaMA-VID and Chat-Univ. **(a) Comparison with LLaMA-VID:** LLaMA-VID compresses the each frame into only two tokens: a visual content token and a text-guided context token. Our compression design in *Rods* is somewhat similar to LLaMA-VID. However, the main difference lies in the resolution of input visual signals processed by the text-guided compression module (i.e., *Context Attention*). Specifically, LLaMA-VID directly use visual feature at their original resolution. In contrast, our method uses the generated semantic prototypes as the input of *Rods*. These prototypes are generated by merging the patch feature with different weight $\rho_i \cdot \delta_i$, where i denotes the patch index in single frame. (b) **Comparison with Chat-Univ.** Chat-Univ adopts DPC-KNN clustering algorithm to form clusters both spatially and temporally. Our method differs from Chat-Univ in the following aspects during the clustering process: **(1) Temporally:** We cluster the frames by calculating the similarity using downsampled features to model more fine-grained temporal relationship, rather than using the feature after global average pooling as in Chat-Univ. This effectively avoids the information loss when performing clustering. **(2) Spatially:** We use $\exp(\rho_i \cdot \delta_i)$ as weight coefficient when generating the prototype from patch features. **(3) Token Budget:** The maximum number of tokens per frame in our method is approximately 60% less than that in Chat-Univ, i.e., 40 tokens versus 112 tokens. **Essentially**, our model highlights adopting the dynamic encoding, which not only reduces the visual nuisance but also effectively reconciles the spatial details with temporal clues using affordable tokens.

Model Variants	MSVD-QA		LV-Bench
	Acc	Score	Acc
K-means [17]	66.5	3.6	23.7
Weighted K-means [6]	66.8	3.6	25.1
DPC-KNN	67.9	3.7	25.8

Table 5. Effects of different clustering algorithm.

Model Variants	MSVD-QA		VCG-Bench
	Acc	Score	Score
Cross-attention (*Soft*)	64.74	3.61	2.56
Concat.	66.20	3.67	2.66
Concat. + Multi-grained	67.90	3.72	2.81

Table 6. Effects of different components in CCE module. Concat. is the concatenation operation.

B.4. Comparison with other Clustering Methods.

There are multiple clustering algorithm [6, 17] available to form the spatial and temporal prototype. To assess the effect of different clustering on model performance, we report the results on two traditional clustering algorithms, K-means and weighted K-means in Table 5. To save the time overhead, we train our model using only the video-based dataset.

B.5. The Effect of Compact Encoding in CCE.

As shown in Table 6, we introduce several variants to assess the impact of fusion strategies between filtered event prototypes $\mathbf{h}_t$ and spatial multi-grained prototypes $\mathbf{G}_t$ on model performance. Although direct concatenation uses slightly more tokens compared to cross-attention, it offers performance advantages with greater parameter efficiency, making it our paramount choice.

B.6. The Effect of Different Training Datasets

In this section, we delve into the effect of data scaling on our model. We begin with adopting the only video-based dataset for training. Specifically, we use WebVid-Cap for vision-language alignment in the first stage and VideoChatGPT-100K for instruction tuning in the second stage. Com-

Table 7. **Ablation of structure and training data.** † represents the results running their official open-sourced code, which adopts the same experimental setting with our *DynFocus*. For fairness, we adopt GPT-3.5-Turbo-16k version for evaluation for all the model in this table.

Methods	Vision-Language Alignment	Instruction Tuning	MSVD-QA		VCG-Bench	VideoMME
	Training Datasets	Training Datasets	Acc	Score	Score	Acc
LLaMA-VID† [13] [ECCV 24]	WebVid-Cap	VideoChatGPT-100K	62.20	3.5	2.67	-
Flash-Vstream† [24]	WebVid-Cap	VideoChatGPT-100K	65.29	3.6	2.76	-
DynFocus ($K = 25, K/L = 0.8$)	WebVid-Cap	VideoChatGPT-100K	67.90	3.7	2.91	35.1
LLaMA-VID† [13] [ECCV 24]	WebVid-Cap, LLaVA-CC3M	VideoChatGPT-100K, LLaVA-625K	68.70	3.6	2.67	-
LLaMA-VID (Reported) [13] [ECCV 24]	WebVid-Cap, LLaVA-CC3M	VideoChatGPT-100K, LLaVA-625K	69.70	3.7	2.89	-
Flash-Vstream† [24]	WebVid-Cap, LLaVA-CC3M	VideoChatGPT-100K, LLaVA-625K	69.86	3.8	2.97	-
DynFocus ($K = 25, K/L = 0.8$)	WebVid-Cap, LLaVA-CC3M	VideoChatGPT-100K, LLaVA-625K	71.20	3.9	3.05	41.2
DynFocus ($K = 25, K/L = 0.8$)	WebVid-Cap, LLaVA-CC3M	+ Science-QA	71.70	3.9	3.05	41.8
DynFocus ($K = 25, K/L = 0.8$)	WebVid-Cap, LLaVA-CC3M	+ Science-QA, CLEVRER	71.60	3.9	3.07	42.6
DynFocus ($K = 25, K/L = 0.8$)	WebVid-Cap, LLaVA-CC3M	+ Science-QA, CLEVRER, NeXT-QA, WebVid-QA	72.30	3.9	3.17	44.1

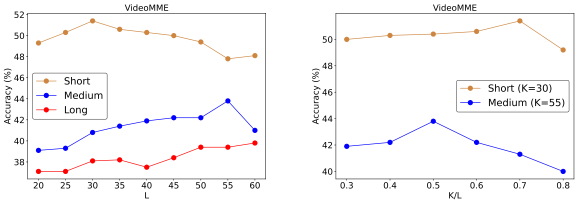
Table 8. Performance comparison of existing VideoLLM on VideoHallucer Benchmark for hallucination diagnosis. To evaluate the accuracy, we present the performance of all these models on basic questions, hallucinated questions, and the overall score. † represents the results by adding rectified prompt *"Please Carefully Think."*, and †† denotes the model with DPO tuning.

Models	LLM Size	Object-Relation (%)			Temporal (%)			Semantic Detail (%)			Factual (%)			Non-Factual (%)			Overall
		Basic	Halluc.	Final	Basic	Halluc.	Final	Basic	Halluc.	Final	Basic	Halluc.	Final	Basic	Halluc.	Final	
VideoChatGPT [10]	7B	95.5	7.0	6.0	100.0	0.0	0.0	96.5	4.0	2.0	86.5	13.5	7.0	85.5	27.5	17.0	6.4
LLaMA-VID [ECCV 24] [14]	7B	78.5	59.0	43.5	86.0	25.0	21.0	89.0	24.0	17.0	98.0	2.5	2.5	16.0	14.0	3.5	21.0
LLaMA-VID [ECCV 24] [14]	13B	87.5	55.5	44.5	78.5	35.0	27.0	90.5	30.0	25.5	85.0	17.5	12.5	84.5	46.5	36.5	23.5
Video-LLaMA2 [14]	7B	88.5	21.5	18.0	91.5	8.5	7.5	99.0	1.5	1.0	88.0	8.5	6.5	87.5	23.5	17.0	10.0
VideoChat2 [CVPR 24][11]	7B	26.0	41.5	10.5	23.5	25.0	7.5	33.0	26.0	9.0	32.0	16.5	7.0	34.0	20.0	5.0	7.8
VideoLLaVA [EMNLP 24] [14]	7B	95.0	38.0	34.5	97.5	13.5	13.5	97.0	14.0	12.0	93.0	4.5	3.0	93.0	31.5	26.0	17.8
VideoLaVIT	-	94.5	39.0	35.5	88.5	27.0	25.5	96.5	13.0	10.5	97.5	6.0	4.0	97.5	21.5	19.0	18.9
MiniGPT4-Video [2]	7B	80.5	34.5	27.5	68.5	27.0	18.0	68.5	27.0	23.5	86.0	16.5	12.0	83.5	37.5	30.5	22.3
PLLaVA [22]	-	76.0	76.5	60.0	46.5	58.0	23.5	83.0	71.5	57.0	85.0	18.0	9.5	85.0	53.5	40.5	38.1
LLaVA-NeXT†† [25]	7B	72.0	73.0	51.5	53.0	61.0	28.0	63.5	69.0	38.0	62.5	41.0	14.0	61.5	60.5	28.5	32.0
DynFocus	7B	86.5	56.0	48.0	86.0	21.5	18.5	92.0	34.0	29.0	96.5	9.0	7.5	-	-	-	-
DynFocus†	7B	88.0	62.0	52.5	87.0	37.5	33.5	91.5	42.0	38.5	98.5	15.0	13.0	96.5	40.0	38.5	35.1

pared with two strong baselines, our model scores 67.9% on MSVD-QA, even outperforming several models that uses additional image-based dataset for training. As we introduce more image-based dataset, our method consistently shows improving performance, maintaining its leading position. Notably, the addition of CLEVRER appears to degrade the model performance. This possibly because that the visual scene involved in CLEVRER differs significantly from those in the targeted evaluation benchmarks, despite it potentially enhances the spatial reasoning and counting abilities of our model.

B.7. Different L and $\frac{K}{L}$ towards Long-term Video

We assess the performance variation with different L and $\frac{K}{L}$ when handling longer and more complex videos, as shown in the following figure,



(a) Number of Initial Prototypes (b) Ratio of Filtering Prototypes

We have two observations for longer videos: (a) the optimal L shifts progressively to the right, from 30 to 55, and further to 60; (b) a smaller $\frac{K}{L}$ yields better performance. This is primarily due to long videos introducing more redundant visual events, while a smaller portion of events should be adaptively selected for question answering. The default parameters towards L and $\frac{K}{L}$ are set to 25 and 0.8 in the main paper when performing evaluation without specification, to achieve a trade-off between accuracy and efficiency.

B.8. Robustness on Video Hallucination

Several researches have pointed that existing MLLMs suffers from the issues of hallucination, which means that they tend to generate irrelevant or nonsensical content that deviates from the original visual context. To comprehensively demonstrate the robustness of our method, we compare the extent of video hallucination of our method with existing video MLLMs. The evaluated benchmark VideoHallucer categorizes hallucinations into two main types: intrinsic and extrinsic, offering further subcategories for detailed analysis, including object-relation, temporal, semantic detail, extrinsic factual, and extrinsic non-factual hallucinations. The overall results are delineated in Table 8. We have several following observations: (1) Although all models demonstrate strong capabilities in answering basic questions, they experience a significant decline in accuracy when dealing with halluci-

Table 9. Video-Language instructional data statistics for training.

Modality	Dataset	Task
Video-Text	VideoChatGPT [16]	Instruction
	WebVidQA [18]	VQA
	CLEVRER [23]	VQA
	NeXT-QA [3]	VQA
Image-Text	COCO [15]	Captioning
	Visual Genome [9]	Captioning
	GQA [7]	VQA
	OCRVQA [21]	VQA
	TextVQA [1]	VQA
	ScienceQA [12]	VQA
Vision-Language	Total	Mixture

Table 10. Video-Language pre-training data statistics for training. We directly adopt the filtered version following LLaVA-VID [13].

Modality	Dataset Source	Task
Video-Text	WebVid-Cap [4]	Captioning
Image-Text	LLaVA-filtered CC3M [20]	Captioning
Vision-Language	Total	Captioning

nated questions. This huge gap implies a widespread conclusion that existing models are vulnerable to the "Yes/NO' 'bias. In other words, most models tend to generate the "Yes" answers. (2) Our *DynFocus* ranks second among all the baselines. VideoChat2 and PLLaVA share the same video-based instructional data but obtain the diametrical results, and the difference stems from source of image-based knowledge. Specifically, the image-based knowledge preserved in PLLaVA originates from a pre-existing image-based MLLM, whereas the knowledge in VideoChat2 is learned from scratch based on collected image QA pairs. On contrary, our model achieves a clear-cut performance gain of **28.3%** compared with VideoChat2, and comparable results to PLLaVA. It is noteworthy that our method employs a dynamic encoding strategy, where each frame is encoded with 40 tokens or 2 tokens depending on its contribution to question answering, which is much less than VideoChat2 and PLLaVA.

C. More Implementation Details

C.1. Training Details

For most of input videos, we sample the frame at 1 *fps* following LLaVA-VID [13] and Flash-Vstream [24], except excessive long video. All input images or frames are resized to 224 × 224 and encoded as 16 × 16 visual features via pre-trained EVA-G [5], and the hidden dimension d is 1408. We set $I = 22$, $J = 2$, $P = 16$, $K = 20$, and $L = 25$ when training to achieve a trade-off between performance and memory efficiency. During vision-language alignment, we pre-train our model with a batch size of 256, employing AdamW [8] optimizer with a cosine schedule. The learning rate is set to 2e-3, and the warmup rate is 0.03. For instruction tuning, the batch size is 32, and the learning rate is 2e-5. We empirically observe that training more than 1 epoch would hamper performance, we thus set the optimal training epoch to 1. Our model is trained using 8 × NVIDIA A100 80G GPUs. All training and inference experiments were conducted under BF16 precision to save time and resources. The training settings are summarized in Table 11.

Table 11. Training settings of our *DynFocus.*

Settings	Stage-1	Stage-2
Batch size	256	32
Learning rate	1e-3	2e-5
Learning schedule	Cosine decay	
Warmup ratio	0.03	
Weight decay	0	
Epoch	1	
Optimizer	AdamW	
DeepSpeed stage	1	0
Visual encoder	Freeze	
Projector $\mathcal{F}_{\text{coarse}}, \mathcal{F}_{\text{fine}}$	Open	
LLM	Freeze	Open

C.2. Statistics of Training datasets

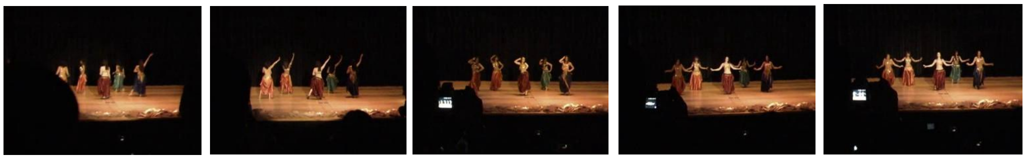
The used training dataset for training are listed in Table 9 and Table 10, respectively.

C.3. Details of Long-Term Video Benchmark

LV-Bench. It encompasses a diverse set of tasks aimed at long video comprehension and information extraction, which tests six core capabilities. Temporal Grounding (TG) focuses on understanding sequences and dynamics within the video. Summarization (Sum) requires an entire understanding of video from start to finish. Reasoning (Rea) involves four advanced reasoning skills: casual relationship identification, understanding for emotional development of character, understanding for underlying intentions of characters, future prediction. Entity Recognition (ER) requires the key entities tracking (such as people, places, and objects) throughout the video. Event Understanding (EU) needs to summarize the semantic concept for question answering. Key Information Retrieval (KIR) emphasizes retrieval of crucial detailed clues within videos.

MLVU. The evaluation task of MLVU can be categorized into three types: (1) *holistic LVU* (TR: Topic Reasoning, AR: Anomaly Recognition, VS: Video Summary), which requires to make use of global perspectives from the entire video; (2) *single-detail LVU* (NQA: Needle QA, ER: Ego Reasoning, PQA: Plot QA, SSC: Sub-Scene Captioning), which needs to pinpoint one critical details in a haystack; (3) *multi-detail LVU* (AO: Action Order, AC: Action Count), which calls for the joint utilization of multiple detailed plots within the long video to collaborately infer the answer.

What happened in the video?



In the video, a group of women walked onto a stage and began dancing. They moved their hips and arms in a rhythmic manner, and the audience cheered them on. The women continued dancing until the end of the performance

How many dancers are there in this video?

There are five dancers in this video.

Write a story about this video.

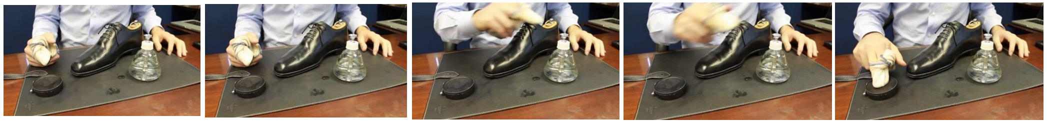
Once upon a time, in a small village nestled in the mountains, there lived a group of women who loved to dance. They would gather every week to practice their moves and perform for the villagers. One day, a group of travelers passing through the village noticed their talent and offered to invite them to perform at a grand festival in the city. The women were thrilled and practiced day and night to prepare for the big event.

On the day of the festival, the women arrived at the grand stage and the audience was amazed by their graceful and synchronized movements. The women's costumes were adorned with vibrant colors and intricate designs that added to the beauty of their performance. The audience cheered and applauded as the women danced and twirled around the stage.

The women's performance was so captivating that the audience demanded an encore. The women happily complied and returned to the stage, this time with even more energy and enthusiasm. The audience was left in awe of their talent and the women's performance was hailed as one of the best at the festival.

From that day on, the women continued to practice and perform, bringing joy and happiness to the people of the village and beyond

Can you guess this man's profession from this outfit?

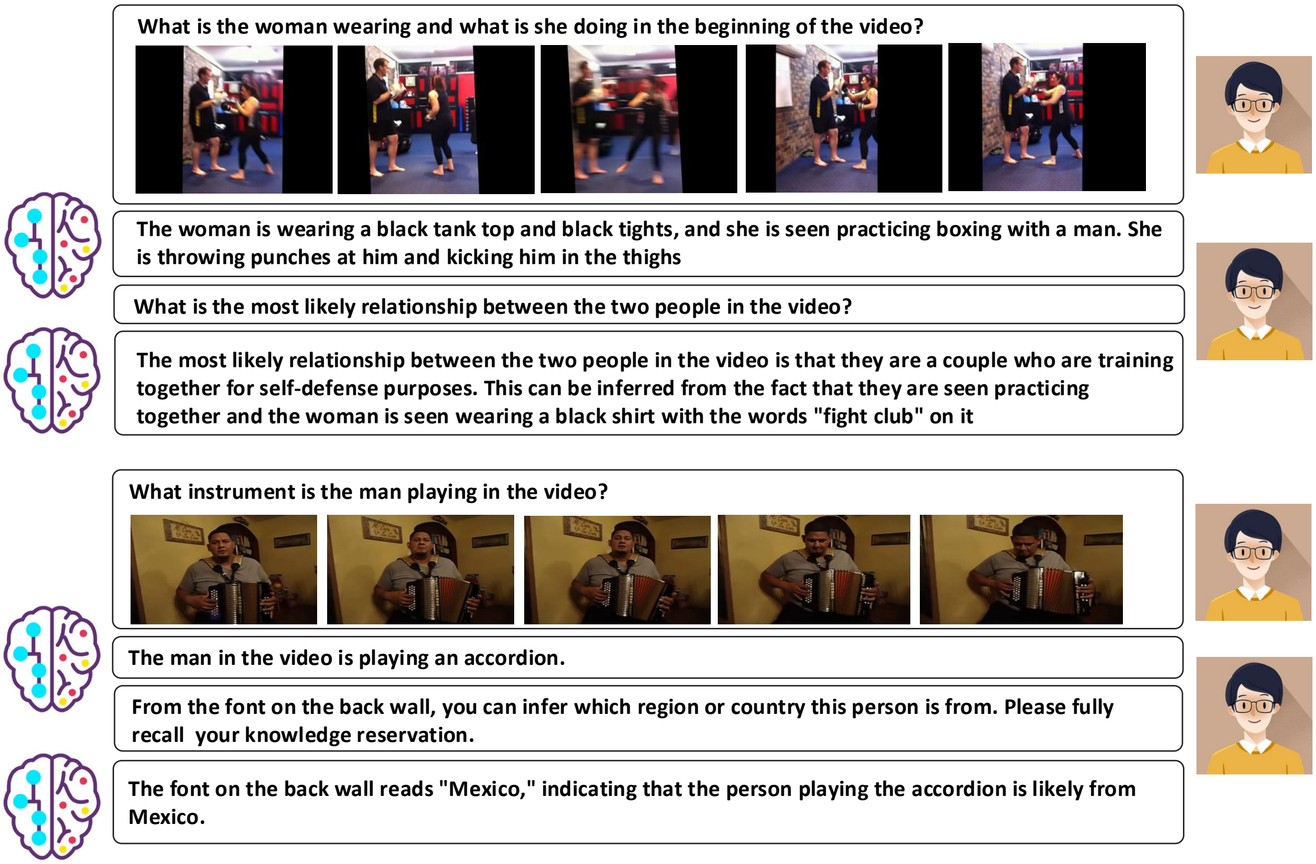


Based on the video, it is likely that the man is a professional shoemaker. He is shown meticulously cleaning and polishing a shoe using various tools, which suggests that he is a skilled craftsman in the field of shoe care.

What movie can you associate with this scene?

This scene can be associated with the movie "The Godfather." In the movie, a man is shown polishing a shoe with a cloth, which is a common task in the world of organized crime. The act of polishing a shoe can be seen as a metaphor for the importance of attention to detail and the value of presenting oneself in a polished and professional manner, which are important values in the world of organized crime.

Figure 1. Case Studies.



What is the woman wearing and what is she doing in the beginning of the video?

The woman is wearing a black tank top and black tights, and she is seen practicing boxing with a man. She is throwing punches at him and kicking him in the thighs

What is the most likely relationship between the two people in the video?

The most likely relationship between the two people in the video is that they are a couple who are training together for self-defense purposes. This can be inferred from the fact that they are seen practicing together and the woman is seen wearing a black shirt with the words "fight club" on it

What instrument is the man playing in the video?

The man in the video is playing an accordion.

From the font on the back wall, you can infer which region or country this person is from. Please fully recall your knowledge reservation.

The font on the back wall reads "Mexico," indicating that the person playing the accordion is likely from Mexico.

Figure 2. Case Studies.

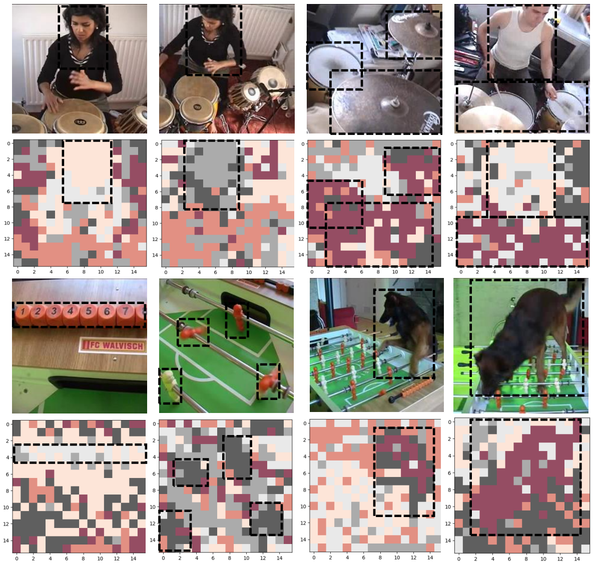


Figure 3. Illustration of learned spatial prototypes in S-DPC. We highlight the region with dotted line for better understanding.

D. More Visualization Results

In Figure 3, we illustrate the learned semantic prototypes, where the patches with similar semantic are first clustered. The formation of spatial prototypes effectively reduces the token number while enhancing the semantic representation of each video frame.

E. Case Study

Figure 1 and Figure 2 illustrates the conversation example towards video understanding. Our method could harness the information of contextual clues to provide appropriate and coherent responses based on user prompts. The illustrative examples showcase the remarkable ability of *DynFocus* on capturing the temporal dynamics and delicate visual details, addressing the counting problem as well as imagination across multiple conversational turns.