Video-XL:
Extra-Long Vision Language Model for Hour-Scale Video Understanding

Supplementary Material

Overview of Supplementary Material

A. Limitations and Future Works

Although Video-XL has a strong capacity to handle extremely long videos, it also has several limitations: (i) Large training memory cost: During training, we unfreeze all parameters of CLIP, the vision-language projector, and the LLM, requiring substantial GPU memory. Additionally, processing video frame extraction and handling a large number of visual tokens demand extra computational resources. We will continue to optimize the training process to improve efficiency or integrate smaller-scale LLMs. (ii) Performance decline with increasing visual tokens: As shown in the Needle-in-the-haystack evaluation, Video-XL occasionally makes errors when the context exceeds 1,000 frames. In the future, we will continue to improve the visual compression module and reduce the information decay for longer video understanding.

B. Relation to Concurrent Works

In this section, we compare and discuss the relation between our Video-XL with the concurrent works including LongVA [53] and VoCo-LLaMA [51].

Comparison to LongVA. Though Video-XL shares a similar model architecture with LongVA, they have several important distinctions. First, although the unified visual encoding mechanism resembles that of LongVA, Video-XL simultaneously encodes single images, multiple images, and videos during training, whereas LongVA only uses image data. Furthermore, we explore the knowledge transfer extensively, conducting experiments to demonstrate that single images and multiple images contribute to long video understanding from different perspectives. Regarding technical contributions to improving long-context processing capability, LongVA fine-tunes the LLM to extend its context length, while Video-XL designs Visual Summarization Token and learns to compress visual tokens. Consequently, Video-XL can handle more visual tokens than LongVA using the same computing device.

Comparison to VoCo-LLaMA. Like VoCo-LLaMA, Video-XL also leverages the inherent capability of LLMs to compress visual tokens. Beyond VoCo-LLaMA for image token compression, Video-XL introduces several significant technical contributions that distinguish it from that one. Firstly, VoCo-LLaMA appends all special tokens at the end of a chunk, while our Video-XL splits long visual token sequences into fine-grained intervals and interleaved special tokens. Thus, VoCo-LLaMA has difficulty in handling long videos. Secondly, the dynamic compression is utilized in Video-XL to ensure the control of compression granularity within a video. Thirdly, we optimize the training process by utilize composite data curation and curriculum learning techniques. Since the official weights of VoCo-LLaMA have not been released, we cannot comprehensively compare the models on long video benchmarks. Therefore, we report our performance (in Table 1) on image understanding benchmarks for reference, though our model mainly designed is for video understanding. It shows that Video-XL achieves significantly better results than VoCo-LLaMa even though on image benchmarks.

Model	Compression Ratio	MMB	GQA	SEED
VoCo-LLaMA	-	64.0	61.1	57.9
VoCo-LLaMA	8	60.5	60.4	56.3
VoCo-LLaMA	16	59.4	60.2	56.2
Video-XL	-	71.6	60.0	61.6
Video-XL	8	71.4	59.3	61.2
Video-XL	16	70.9	59.1	61.0

Table 1. Comparison with VoCo-LLaMA under different compression ratios on image understanding benchmarks.

C. Details in LLM Compressor

In our work, we split long visual sequences into shorter intervals and introduce the special tokens, namely the VSTs, which condense LLM's raw activations into more compact ones. Consequently, the same context window can intake more information from the previous context, which will benefit the prediction of new tokens. In each decoding layer of the LLM, let D denote the LLM's hidden size and L is the size of the chunk, the input hidden states of VSTs ($H_{vst} \in \mathbb{R}^{k \times D}$) are transformed to query the raw KV activations within the chunk: $\{K, V \mid K \in \mathbb{R}^{L \times D}, V \in \mathbb{R}^{L \times D}\}$, where the condensed activations can be produced.

Formally,

$$Q_{vst} \leftarrow H_{vst}W'_Q, \quad K_{vst} \leftarrow H_{vst}W'_K, \quad V_{vst} \leftarrow H_{vst}W'_V \tag{4}$$

$$A \leftarrow \mathrm{softmax}\big(\mathrm{mask}(Q_{vst}\{K \oplus K_{vst}\}^T)\big) \tag{5}$$

$$V_{vst} \leftarrow A\{V \oplus V_{vst}\}^T, \quad O_{vst} \leftarrow V_{vst}{W'_O}^T. \tag{6}$$

The newly generated KV activations for the VSTs, i.e., $K_{vst}, V_{vst} \in \mathbb{R}^{k \times D}$, which leads to a condensing ratio of $\alpha = L/k$ ($k \ll L$). Moreover, the VSTs are parameter-efficient because they primarily rely on the LLM's original parameters, introducing only a few additional projection matrices. For instance, they add no more than 1B parameters to the Qwen2 7B base model.

D. Further Discussion of Video-XL

In this section, we discuss more details about Video-XL, including the training method, computational efficiency and generalization.

Ablation studies on dynamic compression strategy. Our dynamic compression strategy enables Video-XL to control the granularity of compression. We conduct empirical experiments to demonstrate the effectiveness of this strategy. Specifically, we compare dynamic compression with fixed methods that use different fixed interval sizes to encode video segments. As shown in Table 2, the performance of Video-XL is sensitive to the hyperparameter of fixed size L. Generally, larger L results in poorer performance, as coarse-grained compression can damage detailed visual information. While smaller L improves performance, it is highly benchmark-dependent and incurs longer training times. In contrast, dynamic compression allows Video-XL to achieve consistent performance across all benchmarks.

L (tokens)	Video-MME	MLVU	MMB
144×2	41.3	52.3	71.5
144×4	40.7	52.1	70.6
144×8	39.4	51.0	69.6
144×16	38.9	50.5	69.5
144×32	38.6	50.1	69.2
Dynamic	41.6	52.3	71.3

Table 2. Ablations on dynamic compression strategy.

Ablation studies on training methods. To train Video-XL more effectively, we explored various methods based on visual instruction tuning, primarily differing in the fine-tuning phase. In the first, two-stage method, we initially trained for one epoch without setting compression parameters. Next, we froze the parameters of CLIP, the projector, and all pre-trained LLM parameters, optimizing only the newly introduced parameters in the compression module for an additional epoch using the same data. In the second, single-stage method, we activated all parameters and trained for two epochs with the same training data. The results of both methods are reported in the table. Compared to the second method, the first method achieved better training efficiency, though the second method produced superior results. We speculate that this is mainly because the projector pre-aligns CLIP and the LLM, making it less adaptable to compressed knowledge.

Model	Video-MME	MLVU	MMB	Training time
Upper-bound	41.8	52.6	71.6	-
One-Stage	35.3	46.8	66.7	1.5 days
Two-stage	41.4	52.0	70.9	2 days

Table 3. Ablation studies on training methods of Video-XL.

Discussion on the computational efficiency. Video-XL reduces the KV cache by α times where α is the *average compression ratio* and hence the memory cost. This is because it only needs to store the compressed activations of the preceding chunks instead of the raw activations. In terms of computation, the situation is a bit more complex. Specifically, Video-XL significantly reduces the computation in self-attention, because each token only needs to interact with local tokens within the chunk and preceding VSTs, which are approximately α times shorter than the raw context. However, it also triggers more computation to encode the inserted VSTs in other modules (e.g., MLP). Formally, given an LLM with a fixed number of layers, attention heads, and hidden size, let s denote the input context length, s^{pst} denote the cached context length, the forward FLOPs is:

$$\mathrm{FLOPs} = F^{Att}(s, s^{pst}) + F^{Oth}(s), \tag{7}$$

where F^{Att} is the computation during self attention, and F^{Oth} is the computation of other modules. For full-attention models, $s = n, s^{pst} = 0$. For the LLM compressor in Video-XL, the FLOPs is:

$$\mathrm{FLOPs}^{bcn} = \sum_{i=1}^{\lceil \frac{n}{w} \rceil} F^{Att}\left(\frac{(\alpha+1)w}{\alpha}, \frac{(i-1)w}{\alpha}\right) + F^{Oth}(n + \lceil \tfrac{n}{\alpha} \rceil) \tag{8}$$

Specifically, denote the input sequence length as s, the cached sequence length as s^{pst}, query head number as h^q, key/value head number as h^k, the hidden size D, head dimension as d, intermediate size I, and vocabulary size V, FLOPS can be calculated as follows:

$$
\begin{aligned}
F^{Att} &= F^{qkv} + F^{qk} + F^{softmax} + F^{av} + F^{out} \\
F^{qkv} &= 2 \times s \times D \times d \times h^q + 2 \times 2 \times s \times D \times d \times h^k \\
F^{qk} &= 2 \times h^q \times s \times (s + s^{pst}) \times d \\
F^{soft} &= h^q \times (s + s^{pst}) \times (s + s^{pst}) \\
F^{av} &= 2 \times h^q \times s \times (s + s^{pst}) \times d \\
F^{out} &= 2 \times s \times d \times h^q \times D \qquad (9) \\
F^{Oth} &= F^{up} + F^{gate} + F^{down} + F^{lm} \\
F^{up} &= 2 \times s \times D \times 2 \times I \\
F^{gate} &= s \times \times I \\
F^{down} &= 2 \times s \times D \times I \\
F^{lm} &= 2 \times s \times D \times V \qquad (10)
\end{aligned}
$$

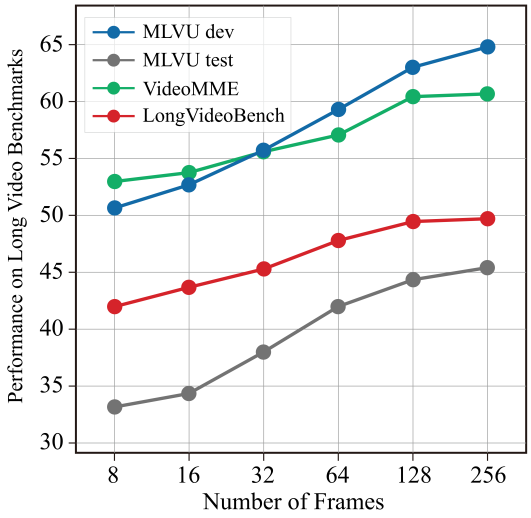


Figure 1. Video-XL can achieve better performance on long video benchmarks with increased context length.

Discussion on the Generalization. We find that Video-XL demonstrates strong generalization capabilities. On one hand, the language compressor can be flexibly applied across various language models. In addition to Qwen2, we conducted experiments with Vicuna and LLaMA2, with results shown in the Table 5. In the future, we plan to introduce smaller language models to enable Video-XL to process longer videos on a single GPU. On the other hand, Video-XL requires only relatively short videos (under two minutes) for training but can handle videos nearly an hour long during inference. We believe this is largely due to the design of both the training data and the model itself. In the training data, the model learns to understand long videos by leveraging images, multi-image sequences, and short video data. Furthermore, in the model's architecture, Video-XL applies relative positional encoding to the visual tokens within each chunk, enabling it to generalize its understanding to infinitely long videos during inference. As shown in Figure 3, the increase of the context length can boost the performance of Video-XL on several long video benchmarks.

LLM	Video-MME	MLVU	MMB
Vicuna-7B	36.5	48.1	63.2
LLaMA2-7B	38.3	49.6	66.8
Qwen2-7B	41.4	52.0	70.9

Table 4. Video-XL has strong generalization to different LLMs.

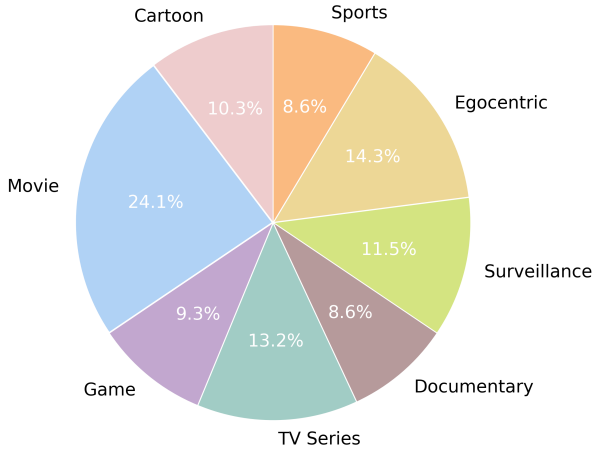


Figure 2. The distribution of video source for VICO.

E. Analysis of VICO Dataset

To enhance long video understanding and unlock the potential of visual compression, we developed an automated long-video data production pipeline and a high-quality dataset called Visual Clue Order (VICO). In this section, we highlight its key features across several aspects.

Diversified Video Categories. VICO offers a comprehensive collection of videos across various genres. Initially, we source videos from CinePile, which includes movies, TV series, and cartoons. Additionally, we collect real-world videos such as egocentric videos, documentaries, games, sports, tutorials, and surveillance footage. The proportion of each video type is illustrated in Figure 2.

Versertile video length. VICO comprises videos of diverse lengths, ranging from 1 minute to over 9 minutes, as shown in Figure 3. Additionally, each video is annotated with QA pairs for event/action ordering and detailed captions for individual video clips. This allows MLLMs to leverage the dataset to enhance their long video comprehension capabilities.

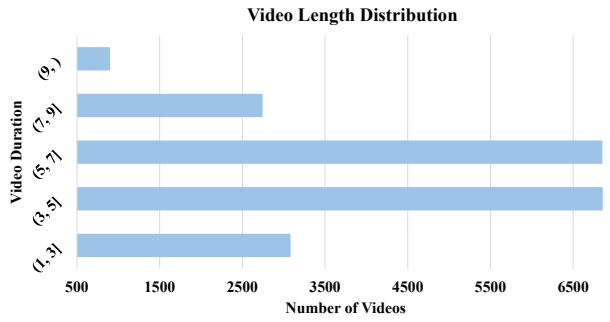


Figure 3. The distribution of video length for VICO.

Robust Generalization. With the assistance of VICO, Video-XL can boost its performance on general long video tasks like MLVU and Video-MME. Here, we demonstrate more experimental results, where LongVA and VideoXL are trained with different scaled VICO data. As shown in Table, both LongVA and VideoXL can benefit from the scaling up of VICO, proving that VICO strength the precise and comprehensive retrieval ability of captured information. We provide the prompt for QA generation in Figure 4 and more visualization results of VICO in Figure 5.

Methods	Size	MLVU	Video-MME
LongVA	5k	51.5	59.4
LongVA	10k	51.9	60.2
LongVA	20k	52.6	61.3
Video-XL	5k	53.9	60.1
Video-XL	10k	54.3	60.9
Video-XL	20k	54.9	61.8

Table 5. The performance of scaling up VICO.

F. Experimental Settings & Additional Results

We elaborate on the training and inference details of Video-XL. Since our method only modifies the workflow of LLM, the hyperparameters reported are specific to the fine-tuning stage, as shown in Table 6. For the inference details, we emphasize the particular context length for different benchmarks, as shown in Table F.

Although Video-XL is designed for long video understanding, it also demonstrates strong proficiency in image understanding. We conduct extensive experiments on several image QA benchmarks, where Video-XL exhibits significant advantages over previous methods, as shown in Table 8.

Hyperparameter	Value
Overall batch size	8
Learning rate	1e-5
LR Scheduler	Cosine decay
DeepSpeed ZeRO Stage	ZeRO-2-offload
Optimizer	Adam
Warmup ratio	0
Epoch	1
Weight decay	0
Precision	bf16

Table 6. Hyperparameters of Video-XL.

Dataset	Context Length
MLVU	256 frms
Video-MME	128 frms
VNBench	1 fps
LongVideoBench	256 frms
VideoVista	128 frms
VideoChatGPT Bench	16 frms
MVBench	16 frms

Table 7. Experimental settings of Video-XL.

Methods	MME	MMB	GQA	POPE	SeedBench
LLaMA-VID	1521.4	65.1	64.3	86.0	59.9
VoCo-LLaMA	1323.3	58.8	57.0	81.4	53.7
Video-XL	1530.2	75.3	65.1	83.2	59.4

Table 8. The performance of Video-XL on mainstream image understanding benchmarks.

G. Qualitative Results

We present qualitative illustrations in Figures 6–8, where video samples are selected from the long video benchmark MLVU, with durations ranging from 10 to 30 minutes. To showcase the long video understanding capabilities of Video-XL, we focus on two representative tasks. The first is PlotQA, which requires the MLLM to reason about questions related to the plot of a narrative video. The second is video summarization, where the MLLM must summarize the key events in a long video. Video-XL demonstrates its ability to accurately locate relevant video segments based on a given query and provide precise answers. Additionally, it effectively captures the overall content of long videos, including summarizing main plots, describing the actions of key characters, and more.

Prompt for VICO QA generation

I will provide you with a series of video clip descriptions. Please summarize 4 unique clues from these clips in sequence. Each clue should consist of sevral words and must be a unique event or action that only appears in one specific clip. Do not include objects, people, or places unless they are integral to describing an event or action. Ensure that the clues are distinct across all clips and are listed in the order they occur.
Please return me the index of the clip and the clue.
The output should be in JSON format as follows:
{index1:"clue1", index2:"clue2", index3:"clue3",index4:"clue4"}

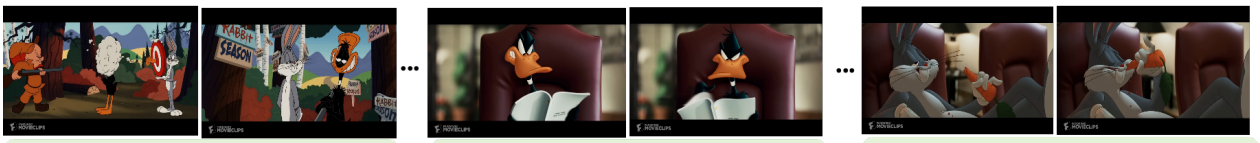
Figure 4. The GPT prompt for VICO QA generation.

The video shows a man is talking to a woman with blue shirt... | The woman is then shown sitting in the bathtub, looking angry and upset.... | The woman quickly gets up and runs out of the room ....

Q: Please select the correct order of visual elements as they occur in the video.
(A) man pushes woman into bathtub -> woman runs out of room -> woman talking to man in window
(B) woman talking to man in window -> man pushes woman into bathtub -> woman runs out of room
(C) woman talking to man in window -> woman runs out of room -> man pushes woman into bathtub

The video shows a talk between two man, a man lightes a cigar... | The man suddenly stands up and takes a sip of water... | The standing man opens the door and leaves the room ....

Q: Select the option that matches the sequence of objects appearing in the video.
(A) man takes a sip of water -> standing man leaves the room -> seated man lights a cigar
(B) man takes a sip of water -> seated man lights a cigar -> standing man leaves the room
(C) seated man lights a cigar -> man takes a sip of water -> standing man leaves the room



An orange hunter shots to the Daffy Duck, while Bunny ... | The Daffy Duck reads a book, sits in the red chair... | The Bugs Bunny lies in the chair, eating a carrot ....

Q: Please identify the correct chronological sequence of the items in the video.
(A) Daffy Duck gets shot by a hunter -> Daffy Duck reads a book -> Bugs Bunny eats carrots
(B) Daffy Duck reads a book -> Daffy Duck gets shot by a hunter -> Bugs Bunny eats carrots
(C) Bugs Bunny eats carrots -> Daffy Duck gets shot by a hunter -> Daffy Duck reads a book

Figure 5. Some visualization cases of VICO data, which includes clip cation (green) and QA pairs for event/action order (blue).

Please summarize this video, including its main content.

USer

Video-XL

The video begins with SpongeBob SquarePants and Patrick Star in a whimsical setting, transitioning through various scenes that include a pirate ship, a stage performance, and a series of comedic and dramatic events. Initially, SpongeBob is seen in a pirate-themed environment, followed by a scene where he and Patrick are on a stage, engaging in a performance or presentation. The narrative progresses to show them in different settings, including a room with a green curtain backdrop, where they interact with objects like a large blue bucket and a microphone.\n\nAs the video continues, the focus shifts to a more elaborate stage setup.The characters engage in various activities, such as cleaning the stage, interacting with props, and reacting to each other's actions. The scene evolves from a simple cleaning task to a more complex scenario involving a large blue curtain, which is manipulated by the characters, leading to a chaotic situation with the curtain .falling and revealing a red substance on the floor.\n\nThroughout the video, the characters' interactions and the unfolding events suggest a storyline centered around cleaning, performance, and unexpected outcomes. The video concludes with SpongeBob and Squidward engaged in a cleaning activity amidst a chaotic environment, highlighting a comedic and dynamic conclusion to their adventures on stage.

Can you summarize the main contents of this video?

USer

Video-XL

The video captures a series of events at an outdoor wedding ceremony, beginning with a group of musicians performing on stage, followed by a procession of individuals in formal attire, including a bride and groom, walking down the aisle towards the altar. The ceremony unfolds with the couple standing at the altar, surrounded by floral arrangements and guests seated in rows, indicating the commencement of the wedding proceedings.\n\nAs the ceremony progresses, the bride and groom are seen engaging in various rituals, such as exchanging rings, signing documents, The presence of a choir singing adds to the ceremonial atmosphere, while the guests' attention is consistently directed towards the couple, highlighting the significance of their actions.\n\nThroughout the video, the camera captures these moments from different angles, emphasizing the emotional and celebratory nature of the event. The bride's expressions and gestures, along with the groom's participation in the rituals, underscore the personal and intimate aspects of the wedding. The setting, adorned with white flowers and draped fabric, complements the festive mood, while the choir's performance contributes to the auditory experience of the ceremony.\n\nThe video concludes with the couple sharing a moment of affection, symbolizing the bond they have formed, before transitioning to the cutting of the wedding cake, a traditional act marking the beginning of their shared life.

Can you provide a summary of this video that covers the key points?

USer

Video-XL

The video depicts a sequence of events unfolding in various settings, beginning with a man and a woman inside a vehicle, suggesting a journey or an escape. The narrative progresses to show them navigating through different terrains, including a sunflower field and a dense forest, indicating a significant journey or adventure. The focus then shifts to a confrontation between two individuals in a forested area, hinting at a tense situation.\n\nSubsequently, the scene transitions to a more urban setting where a man is seen interacting with a laptop, possibly working or searching for information. This is followed by a shift to a residential or office environment where a woman is engaged in a conversation over a video call, suggesting a professional or personal interaction. The narrative continues with a man in a suit, possibly in a formal setting, engaging in a conversation that hints at a serious or intense discussion. Throughout the video, the presence of text overlays provides context and narrative progression, guiding the viewer through the story's development from a personal journey to a potentially life-threatening situation."

Figure 6. Some visualization cases of Video-XL on visual summarization task.

Please summarize this video, including its main content.

USer

Video-XL

The video depicts a sequence of events unfolding in an urban setting, beginning with two individuals in a dilapidated cityscape at night, followed by a scene where one character is seen running towards a car amidst a backdrop of illuminated buildings and a bridge. The narrative then transitions to a more intimate setting, focusing on a character in a green suit and another in a red outfit, suggesting a confrontation or significant interaction between them.\n\nAs the story progresses, the characters are shown in various states of distress and emotional turmoil, with scenes depicting intense expressions, physical confrontations, and moments of vulnerability. The environment shifts from the initial urban decay to a more industrial or possibly abandoned area, indicating a change in location or a progression in the storyline.\n\nThroughout the video, the camera work plays a crucial role in framing the characters' emotions and actions, often zooming in for close-up shots that emphasize facial expressions and reactions. This technique helps to convey the intensity of the interactions and the characters' internal struggles. The consistent use of lighting and color grading adds to the atmospheric quality of the scenes, enhancing the emotional impact and visual storytelling.

Can you summarize the main contents of this video?

USer

Video-XL

The video begins with a scene of a penguin standing on a rocky cliff overlooking a snowy landscape, with a bird in flight and a small animal near the water's edge. The narrative then transitions to a snowy field where a polar bear is seen walking towards the camera, followed by a close-up of the bear's face and a shift to a different polar bear in motion across the snow. The focus moves to a group of birds flying over the ocean, transitioning to a polar bear swimming in icy waters, and then to a polar bear on ice.\n\nAs the video progresses, it showcases a polar bear on ice, surrounded by a flock of birds, suggesting a shared habitat. The scene shifts to a polar bear on land, with a flock of birds nearby, before moving to a snowy landscape with a herd of deer. The narrative continues with a focus on a polar bear in a snowy environment,The video concludes with a polar bear swimming in icy waters, surrounded by birds, highlighting the interaction between these species in their natural habitats. Throughout, the video captures the essence of wildlife in various environments, emphasizing the coexistence of different species in their respective ecosystems.

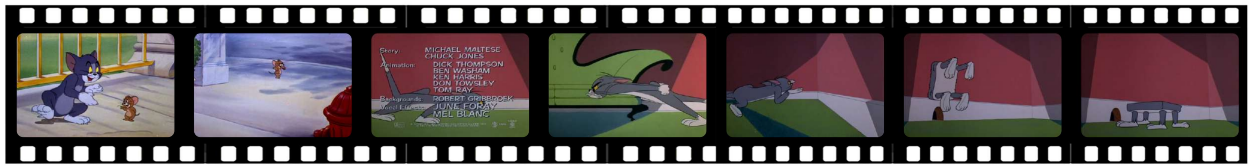
Can you provide a summary of this video that covers the key points?

USer

Video-XL

The video depicts a series of events unfolding in an urban setting, beginning with a woman standing on a bridge overlooking a cityscape at dusk or dawn, dressed in a dark outfit and appearing contemplative. The scene transitions to a nighttime street view where a car is parked by the curb, and a man approaches it from behind, suggesting a potential interaction between them.\n\nAs the video progresses, the focus shifts to a woman inside a vehicle, engaged in a conversation over her phone, indicating a narrative involving communication and possibly urgency. The setting then changes to an outdoor night scene with a woman walking alone, followed by a shift to an indoor environment where she interacts with a computer terminal, suggesting a transition to a more technologically oriented storyline.\n\nThe narrative continues with the woman navigating through a dimly lit corridor, moving towards a door marked \"ATTENTION,\" which hints at a restricted or secure area. Her actions suggest she is searching for something or someone within this space. The video culminates in a dramatic scene where the woman is engulfed in flames, marking a significant escalation in the intensity of the events depicted.\n\nThroughout the video, the presence of Chinese text overlays provides context or commentary on the scenes, while the consistent \"CCTV\" logo indicates that these scenes are part of a television broadcast. The camera work, including panning and zooming, effectively captures the evolving dynamics and settings, guiding the viewer through the unfolding story.

Figure 7. Some visualization cases of Video-XL on visual summarization task.



Why did the cartoon cat's body become flat like a pancake?

USer

Video-XL

It hits the wall.

What color is the clothes of the woman who leaves after the conversation in the video?

USer

Video-XL

Purple.

What color is the cars?

USer

Video-XL

Black.

What shape of musical instrument is the building in the video?

USer

Video-XL

Guitar.

What is the woman doing at the end of the video?

USer

Video-XL

Taking photos.

Figure 8. Some visualization cases of Video-XL on plotQA task.