CrossLMM: Decoupling Long Video Sequences from LMMs via Dual Cross-Attention Mechanisms

Shilin Yan[1]♠, **Jiaming Han**[2], **Joey Tsai**[3], **Hongwei Xue**, **Rongyao Fang**[2]
Lingyi Hong[1], **Ziyu Guo**[4], **Ray Zhang**♣

[1]Fudan [2]CUHK MMLab [3]Tsinghua [4]CUHK MiuLar Lab

tattoo.ysl@gmail.com

♠Project Leader ♣Corresponding Author

Figure 1: **Comparisons of Different Visual Token Compression Methods.** (a) keeps all visual tokens. (b) and (c) merge visual tokens before and within LLMs, respectively. (d) Our method decouples visual tokens from LLMs with a dual cross-attention mechanism. We first merge visual tokens before LLMs to reduce computational cost in LLMs. Then we propose a Visual-to-Visual Cross-Attention (V2V CA) to preserve fine-grained details of original visual tokens into merged tokens, and a Text-to-Visual Cross-Attention (T2V CA) to enhance text tokens with visual information.

Abstract

The advent of Large Multimodal Models (LMMs) has significantly enhanced Large Language Models (LLMs) to process and interpret diverse data modalities (e.g., image and video). However, as input complexity increases, particularly with long video sequences, the number of required tokens has grown significantly, leading to quadratically computational costs. This has made the efficient compression of video tokens in LMMs, while maintaining performance integrity, a pressing research challenge. In this paper, we introduce CrossLMM, decoupling long video sequences from LMMs via a dual cross-attention mechanism, which substantially reduces visual token quantity with minimal performance degradation. Specifically, we first implement a significant token reduction from pretrained visual encoders through a pooling methodology. Then, within LLM layers, we employ a visual-to-visual cross-attention mechanism, wherein the pooled visual tokens function as queries against the original visual token set. This module en-

Preprint. Under review.

ables more efficient token utilization while retaining fine-grained informational fidelity. In addition, we introduce a text-to-visual cross-attention mechanism, for which the text tokens are enhanced through interaction with the original visual tokens, enriching the visual comprehension of the text tokens. Comprehensive empirical evaluation demonstrates that our approach achieves comparable or superior performance across diverse video-based LMM benchmarks, despite utilizing substantially fewer computational resources. Code is available at: https://github.com/shilinyan99/CrossLMM.

1 Introduction

Large Multimodal Models (LMMs) [18, 26, 31, 38, 41, 42, 44, 74] enhance Large Language Models (LLMs) [1, 15, 54, 66] with visual perception capabilities, demonstrating remarkable proficiency in image-language [12, 16, 17, 23–25, 77] and video-language [4, 4, 13, 20, 43, 50, 60, 82] tasks. Contemporary research on LMMs predominantly employs an intermediate module, i.e., the projector, to map visual token representations into LLM embedding spaces, which subsequently serve as prefix content for textual tokens into the LLMs as demonstrated in Figure 1 (a). Nevertheless, the increasing complexity of input modalities, particularly long video sequences, generates substantial quantities of visual tokens. This proliferation of tokens necessitates significant computational resources, thereby constraining the applicability of such models in resource-limited environments and long-context scenarios. Consequently, the development of efficient visual compression methodologies that minimize performance degradation has emerged as a critical research imperative, garnering substantial attention from both academic and industrial communities.

In recent literature, a diverse array of methodologies for visual token compression has emerged, which can be categorized into two principal paradigms as follows. *1) Token Merging before LLM* [22, 47, 81], as illustrated in Figure 1 (b), predominantly entails the compression of visual tokens through semantic similarity metrics before feeding into LLMs. Nevertheless, this methodology exhibits significant limitations: it fails to adequately identify visual tokens corresponding to relevant textual elements, compromising spatial relationships and attenuating the model's capacity for comprehensive image interpretation. *2) Token Pruning within LLM* [45, 52, 69, 80], as depicted in Figure 1 (c), implements a systematic reduction of visual tokens within the input layers of LLMs. This methodology typically employs the quantification of attention weights between visual and textual tokens, subsequently eliminating those tokens that manifest lower weight values. While this approach effectively illuminates multi-modal interactions, it might instead overlook crucial semantic information within the visual modality. Furthermore, both methods experience significant performance degradation as the number of visual tokens decreases. Consequently, an intriguing question arises: *Is it possible to maintain comparable performance while significantly reducing the number of video tokens?*

To achieve this objective, we introduce an efficient architecture designated as **CrossLMM**. The fundamental component comprises a dual cross-attention module. In the visual domain, we initially compress image tokens derived from the visual encoder along the spatial dimension to reduce the token quantity inputted into LLMs. To facilitate comprehensive image information capture, we implement a visual-to-visual (V2V) cross-attention mechanism within the LLM layer. This process enables sufficient interaction between the compressed visual tokens (as queries) and the original long-sequence visual representations (as keys and values). Correspondingly, in the textual domain, we employ a text-to-visual (T2V) cross-attention mechanism to enhance text tokens (as queries) with multimodal information from original visual tokens (as keys and values). This further complements the text generation process with fine-grained visual semantics. Through the implementation of this dual cross-attention mechanism, we endeavor to maintain the fidelity of the original visual tokens, effectively mitigating performance deterioration, while simultaneously achieving substantial token compression.

We evaluate CrossLMM on an extensive spectrum of video-based multimodal benchmarks. Our model, utilizing very few tokens (1 or 9 or 16), demonstrates promising performance across various video assessments, including VideoMME [13] and MLVU [82] and so on. These findings substantiate that the implementation of our dual cross-attention module, coupled with a restricted number of visual tokens, enables LMMs to effectively address diverse visual tasks.

Our contributions can be summarized as follows:

1. We introduce CrossLMM, which effectively compacts long video sequences into efficient representations via a dual cross-attention mechanism.
2. We propose the V2V and T2V cross-attention layers, providing semantics from original visual tokens for compact visual tokens and text tokens, respectively.
3. CrossLMM archives comparable or state-of-the-art performance across various video understanding benchmarks with only few visual tokens, demonstrating superior efficiency.

2 Related Works

2.1 Large Multimodal Models

Large language models (LLMs) [1, 15, 54, 66] trained on extensive datasets have demonstrated remarkable capabilities in text understanding and generation tasks, establishing the foundation for developing multi-modal LLMs to solve many traditional multi-modal comprehensive tasks [10, 11, 21, 32, 37, 61, 63–65]. Among large multimodal models (LMMs) [2, 3, 26, 31, 38, 39, 41, 42, 44], LLaVA [34] has emerged as the predominant architecture, valued for its data efficiency and streamlined design. This approach incorporates a projector that effectively bridges visual and language modalities, while employing instruction-tuning for both the projector and the LLM using comprehensive instruction-following datasets. Recent scholarly work [6, 7, 14, 26, 27, 33, 75] has sought to enhance model performance by implementing high-resolution visual inputs. Concurrently, video-based LLMs [8, 30, 35, 46, 48, 51, 62, 78, 79] have advanced through the extension of visual instruction tuning datasets to accommodate video modality. However, the substantial number of tokens generated by high-resolution and video inputs presents significant computational challenges. Consequently, there is a pressing need to develop efficient methods for visual token compression to address these limitations.

2.2 Visual Token Compression

With high-resolution image and video inputs, the token count has increased substantially, often exceeding textual token quantities by one to two orders of magnitude. Consequently, the efficient compression of visual tokens has emerged as a critical research challenge. Prior approaches [22, 45, 47, 52, 69, 80, 81] addressing this issue can be categorized into two principal directions. The first category encompasses methods that compress tokens generated by visual encoders or implement efficient projections of visual modalities. For instance, LLaMA-VID [31] employs a QFormer [28] architecture to compress visual tokens into a minimal representation of two tokens. Similarly, Deco [68] implements adaptive pooling at the image patch level. However, these approaches lack integration with textual information for guided compression. Alternative methods gradually reduce visual tokens within Large Language Models (LLMs), yet such approaches frequently compromise the semantic integrity of visual information. Therefore, we propose a dual cross-attention module that simultaneously updates text-relevant visual tokens and visually-relevant textual tokens, thereby effectively compressing visual representations while minimizing information loss.

3 Method

In this section, we illustrate the details of our CrossLMM for multi-modal video understanding. We first describe the overall pipeline in Section 3.1. Then, in Section 3.2 and Section 3.3, we respectively elaborate on the proposed designs of visual-to-visual and text-to-visual cross-attention mechanisms.

3.1 Overall Architecture

Our proposed approach, CrossLMM, inspired by current mainstream architecture LLaVA [34], constructs around three core components, including the frame-wise visual encoder, the visual-language projector, and the Large Language Model (LLM), as demonstrated in Figure 2.

Visual Feature Encoding. For visual feature extraction, we employ an image-based encoder rather than video-based architectures, such as VideoMAE [53] or Video-Swin [36]. This approach extracts features from individual frames sequentially. Specifically, we utilize the pre-trained SigLIP [71] vision encoder for encoding visual information. Our methodology is justified by two key considerations: 1)

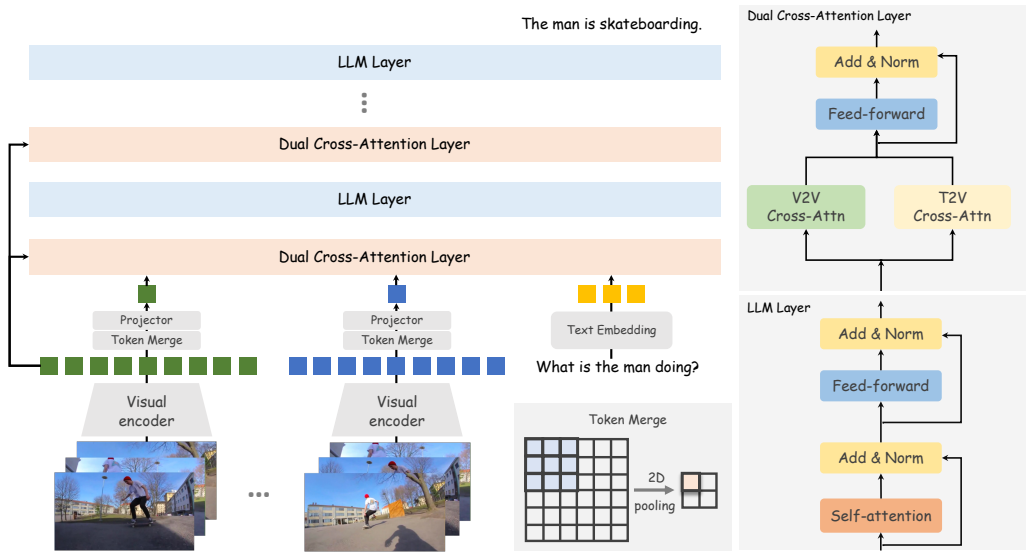


Figure 2: **Architecture of CrossLMM,** which consists of a visual encoder, a visual projector and an LLM. For a pretrained LLM, we insert the proposed Dual Cross-Attention Layer (DCAL) to it every n layers. The DCAL is a variant of general cross-attention layer with two parallel blocks: Visual-to-Visual (V2V) Cross-Attention and Text-to-Visual (T2V) Cross-Attention. Both V2V Cross-Attn and T2V Cross-Attn aggregate fine-grained information from the original visual tokens to produce visual-enhanced video tokens and text tokens.

image encoders normally contain better generalization capabilities with larger-scale training, and 2) the temporal information can be retained by concatenating multiple frame representations. Given an input video-text pair, we sample T frames $v \in \mathbb{R}^{T\times 3\times H\times W}$ from the video clip and apply the visual encoder to extract image features. This process can be formulated as:

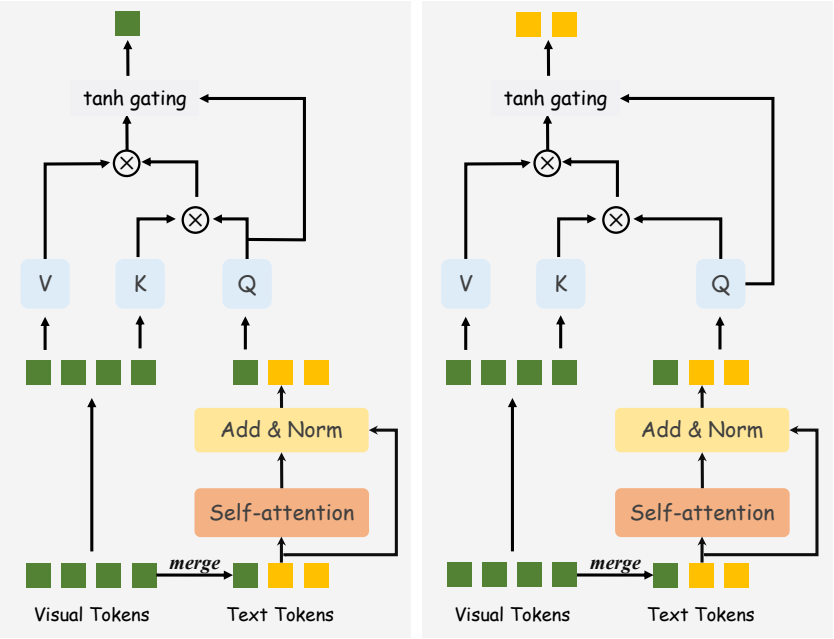
$$\mathcal{V} = \{x_i = \mathcal{F}(v_i) \mid \forall i = 1, \ldots, T\} \tag{1}$$

Here, $\mathcal{V} \in \mathbb{R}^{T\times N\times D}$ corresponds to the feature representations of the T input frames. The visual encoder $\mathcal{F} : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^{N\times D}$ processes each RGB frame v_i (with spatial resolution $H \times W$) to generate $x_i \in \mathbb{R}^{N\times D}$, where N denotes the cardinality of visual tokens and D specifies the latent embedding dimension per token.

Initial Token Merge. After that, the bilinear pooling operator $\mathcal{B} : \mathbb{R}^{N\times D} \to \mathbb{R}^{N/9\times D}$ is applied preceding the projection operation, performing 3×3 local patch aggregation. This spatially downsamples the token grid from original 27×27 ($N = 729$) to 3×3 ($N = 9$) resolution through non-overlapping fusion of spatially adjacent patches, while preserving the feature dimension D through parameter-free structural compression. This architecture introduces geometrically meaningful inductive bias through structural token aggregation while maintaining the integrity of inter-patch spatial relations, which also enhances computational efficiency.

Visual-language Projector. The cross-modal projection module $\Phi : \mathbb{R}^D \to \mathbb{R}^{D'}$ is architected as a parameterized two-layer MLP (MLP : $\mathbb{R}^D \to \mathbb{R}^{d_h} \to \mathbb{R}^{D'}$), operating on visual tokens $x_i \in \mathbb{R}^D$ through successive nonlinear transformations (GeLU [19] activation with layer normalization). This learnable transformation achieves bidirectional semantic alignment by establishing: (1) an injective mapping from vision feature space $\mathcal{V}$ to the LLM's textual embedding manifold $\mathcal{T}$, and (2) a differentiable inverse approximation $\Phi^{-1}_{\text{approx}}$ preserving topological consistency. As the pivotal interface enabling visio-linguistic fusion, Φ induces latent space isomorphism through ℓ_2-sphere projection regularization, thus constituting the mathematical foundation for constructing joint embedding space.

Large Language Model. The architecture builds upon a standard decoder-only LLM foundation, augmented with a hierarchical cross-modal integration scheme. At every K decoder layers, we introduce dual-stream attention mechanisms:



(a) V2V Cross-Attention (b) T2V Cross-Attention

Figure 3: **Implementation Details of Dual Cross-Attention.** For a detailed illustration, please refer to Sec. 3.2 and Sec. 3.3.

- **Visual-to-Visual**: Positioned after textual self-attention, this module computes

$$\text{Attn}(Q_{v_p}, K_{v_o}, V_{v_o}) = \text{Softmax}\left(\frac{Q_{v_p} K_{v_o}^{\top}}{\sqrt{d_k}}\right) V_{v_o} \tag{2}$$

  where Q_{v_p} derives from efficient visual representations and $\{K_{v_o}, V_{v_o}\}$ are projected original visual tokens. This enables text-conditioned vision-guided fine-grained refinement of efficient visual representations from original visual tokens.

- **Text-to-Visual**: Operating in parallel, it establishes

$$\text{Attn}(Q_{\text{text}}, K_{v_o}, V_{v_o}) \tag{3}$$

  creating visual-conditioned text feature updating.

This staggered integration strategy implements multi-round cross-modal fusion through alternating attention directions, where $\mathcal{V}$-Cross-Attn extracts text-related visual specifics to ground language generation, while $\mathcal{T}$-Cross-Attn maintains semantic coherence through visual-related linguistic feedback forming a coupled attention dynamical system. Further details of this $\mathcal{V}$-Cross-Attn and $\mathcal{T}$-Cross-Attn are provided in Section 3.2 and Section 3.3, respectively.

3.2 Visual-to-Visual Cross-attention

To boost both the multi-modal and multi-frame feature fusion, we introduce visual-to-visual (V2V) cross attention module for text-condictioned visual aggregation from coarse-grained to fine-grained. For a multimodal vision-language model, the projected and pooled visual feature $\mathcal{V} \in \mathbb{R}^{L_v \times D'}$ would concat with text feature $\mathcal{T} \in \mathbb{R}^{L_t \times D'}$ into the LLM, where D' denotes the embedding dimension of LLM, and L_v, L_t represents the length of visual and text tokens, respectively. As shown in Figure 3(a), the V2V module take the visual feature $\mathcal{V}$ (pooled and projected), and text features $\mathcal{T}$ as input $\mathcal{I} = [\mathcal{V}; \mathcal{T}]$, where $[;]$ denotes concatenation. we adopt gated cross-modal fusion mechanisms for input tokens to progressively capture multi-modal information. There consists of two stages, self-attention processing and cross-attention fusion. That can be formulated as,

$$\mathcal{I}' = \text{LayerNorm}(\text{SelfAttn}(\mathcal{I}) + \mathcal{I}), \tag{4}$$

where $\mathcal{I}'$ represents the integrated representation of visual and textual features. Subsequently, we extract the coarse text-conditioned visual feature $\mathcal{I}'[:L_v]$. To obtain a more fine-grained visual representation, we employ $\mathcal{I}'[:L_v]$ as query tokens, as expressed by:

$$Q = W_q \mathcal{I}'[:L_v] \in \mathbb{R}^{L_v \times D'}, \tag{5}$$

while utilizing the original tokens as the Key and Value tokens:

$$K, V = (W_k / W_v)\mathcal{V} \in \mathbb{R}^{T \times N \times D'}, \tag{6}$$

where W_q, W_k, and W_v denote linear projection matrices. The fine-grained visual feature is then acquired through cross-attention and gating mechanisms, which can be formulated as:

$$\mathcal{I}'[:L_v] = \mathcal{I}'[:L_v] + \gamma * \text{Attn}(Q, K, V), \tag{7}$$

where $\gamma \in [-1, 1]$ is a learnable gating parameter. Through this process, we derive a hierarchical text-conditioned visual representation that systematically progresses from coarse-grained to fine-grained feature abstraction.

3.3 Text-to-Visual Cross-attention

To enhance the visual comprehension of the text tokens, we implement a text-to-visual (T2V) cross-attention mechanism that facilitates interaction between text tokens and original visual tokens. While structurally parallel to the V2V module previously described, the T2V module operates in the complementary direction, transforming text token representations through visual context. The mathematical formulation follows a similar pattern:

$$Q = W_q \mathcal{I}'[L_v : L_v + L_t] \in \mathbb{R}^{L_t \times D'}, \tag{8}$$

where text tokens serve as queries that attend to the visual information. The original visual tokens function as Key and Value tokens:

$$K, V = (W_k / W_v)\mathcal{V} \in \mathbb{R}^{T \times N \times D'}, \tag{9}$$

with W_q, W_k, and W_v representing the respective linear projection matrices. The enhanced text features are computed through cross-attention and modulated by a learnable gating mechanism:

$$\mathcal{I}'[L_v : L_v + L_t] = \mathcal{I}'[L_v : L_v + L_t] + \gamma * \text{Attn}(Q, K, V), \tag{10}$$

where $\gamma \in [-1, 1]$ controls the influence of the visually-contextualized information. This mechanism complements the V2V module by providing the reciprocal flow of information, resulting in visual-conditioned text representations that capture multi-level semantic relationships between the two modalities.

4 Experiments

4.1 Experiment settings

Implementation Details. CrossLMM employs SigLIP2 [55] as the vision encoder. Qwen2.5-1.5B-Instruct [66] serves as the large language model (LLM) for CrossLMM-2B, while Qwen2.5-7B-Instruct [66] is utilized for CrossLMM-7B. For the visual-to-visual module, we integrate connections at every K layers within the LLM, following the same approach as the text-to-visual module. During the pretraining phase, we implement a learning rate of 1×10^{-4}, while in the instruction tuning stage, the learning rate is adjusted to 1×10^{-5}, with the exception of the Vision Transformer (ViT) component, which uses 2×10^{-6}. Our model undergoes training only one epoch during both pretraining and instrcution tuning stage. More details can be found in supplementary materails.

Evaluation Benchmarks. To validate CrossLMM's general capabilities, we evaluate our model on five video understanding benchmarks spanning short video benchmarks, long video benchmarks, and comprehensive benchmarks. These include two short video benchmarks: MVBench [29] and Perception Test [43], and two long video benchmarks: LongVideoBench [60] and MLVU [82], and a comprehensive benchmark, VideoMME [13], covering videos ranging from minute-level to hour-level.

Table 1: **Comprehensive evaluation of video understanding models.** Performance comparison across multiple video understanding benchmarks for different categories of models (proprietary, small-size LMMs, general open-source LMMs, and specialized long video LMMs). The CrossLMM model (highlighted) achieves competitive or superior performance while using significantly fewer tokens per frame (1 or 9 or 16). † represents the vision encoder is ViT-G.

Model	Size	#tokens per frame	MVBench *Avg*	PerceptionTest *Val*	LongVideoBench *Val*	MLVU *M-Avg*	VideoMME *w/o sub.*	VideoMME *w sub.*
Avg. Duration			16s	23s	473s	651s	1010s	1010s
Proprietary Models								
GPT4-V [41]	-	-	43.7	-	59.1	49.2	59.9	63.3
GPT4-o [42]	-	-	64.6	-	66.7	64.6	71.9	77.2
Gemini-1.5-Pro [44]	-	-	60.5	-	64.0	-	75.0	81.3
Small Size LMMs								
Qwen2-VL [56]	2B	-	63.2	-	-	-	55.6	60.4
InternVL2.5 [6]	2B	256	**68.8**	-	46.0	**61.4**	51.9	54.1
CrossLMM	2B	1	58.6	58.1	47.3	57.6	55.5	58.9
CrossLMM	2B	9	63.5	**63.5**	**51.8**	61.0	**59.1**	**61.3**
Open-Source LMMs								
VideoLLaMA2 [9]	7B	72	54.6	51.4	-	48.5	47.9	50.3
VideoLLaMA2 [9]	72B	72	62.0 57.5	-	-	62.4	64.7	
VideoChat2-HD [29]	7B	72	62.3	-	-	47.9	45.3	55.7
InternVideo2-HD [59]	7B	72	67.2	63.4	-	-	49.4	-
IXComposer-2.5 [72]	7B	400	69.1	34.4	-	37.3	55.8	58.8
InternVL2 [7]	8B	256	65.8	-	54.6	64.0	54.0	56.9
InternVL2 [7]	76B	256	69.6	-	61.1	69.9	61.2	62.8
InternVL2.5 [6]	8B	256	72.0	-	60.0	68.9	64.2	66.9
Qwen2-VL [56]	7B	-	67.0	62.3	-	-	63.3	69.0
Qwen2-VL [56]	72B	-	73.6	68.0	-	-	71.2	77.8
LLaVA-NeXT-Video [78]	7B	144	53.1	48.8	49.1	-	-	46.5
LLaVA-OneVision [26]	7B	196	56.7	57.1	56.3	64.7	58.2	61.5
LLaVA-OneVision [26]	72B	196	59.4	66.9	61.3	68.0	66.2	69.5
LLaVA-Video [79]	7B	676	58.6	67.9	58.2	70.8	63.3	69.7
Open-Source Long Video LMMs								
LLaMA-VID [31]	7B	2	41.9	44.6	-	33.2	25.9	-
Kangaroo [35]	8B	256	61.0	-	54.8	61.0	56.0	57.6
LongVILA [62]	7B	196	67.1	58.1	57.1	-	60.1	65.6
LongVA [73]	7B	144	-	-	-	56.3	52.6	54.3
LongLLaVA [57]	9B	144	49.1	-	-	-	43.7	-
LongVU [46]	7B	64	66.9	-	-	65.4	-	60.6
CrossLMM	7B	1	62.7	64.8	54.5	64.0	61.3	63.7
CrossLMM	7B	9	68.2	68.4	56.0	67.2	62.6	64.7
CrossLMM †	7B	16	**68.8**	**71.4**	**60.4**	**70.7**	**65.4**	**67.6**

4.2 Comparison with State-of-the-art Methods

In this section, we present a comparative analysis of CrossLMM against state-of-the-art LMMs, encompassing commercial implementations [41, 42, 44], open-source LMMs [6, 7, 9, 26, 29, 56, 59, 72, 78, 79], and specialized open-source long video LMMs [31, 35, 46, 57, 62, 73].

Our experimental results, as presented in Table 1, demonstrate several significant findings in the landscape of video understanding models. CrossLMM exhibits remarkable efficiency-performance trade-offs compared to existing models. Despite utilizing only 9 tokens per frame—significantly fewer than competitors that employ between 64 and 676 tokens—CrossLMM achieves competitive performance across multiple benchmarks. Moreover, CrossLMM with 7B parameters and 16 tokens per frame surpasses several mainstream open-source LMMs, while maintaining lower memory and computation overhead. This efficiency is consistently observed under both 2B and 7B model scales, which highlight the potential of CrossLLM that designed efficiently that balance token efficiency and performance for practical video understanding applications.

4.3 Efficiency Analysis

Figure 4 presents a comprehensive efficiency comparison between LLaVA-OV [26] and CrossLMM across different frame numbers (32, 64, 128, and 256) with 8 H800. The analysis focuses on three critical metrics: CUDA memory consumption, computational complexity, and prefill time.

Memory Efficiency. As shown in Figure 4(a), CrossLMM demonstrates remarkable memory efficiency compared to LLaVA-OV. At 32 frames, CrossLMM requires only 1,753MB of CUDA memory, which is 63.9% less than the 4,858MB needed by LLaVA-OV. This memory advantage

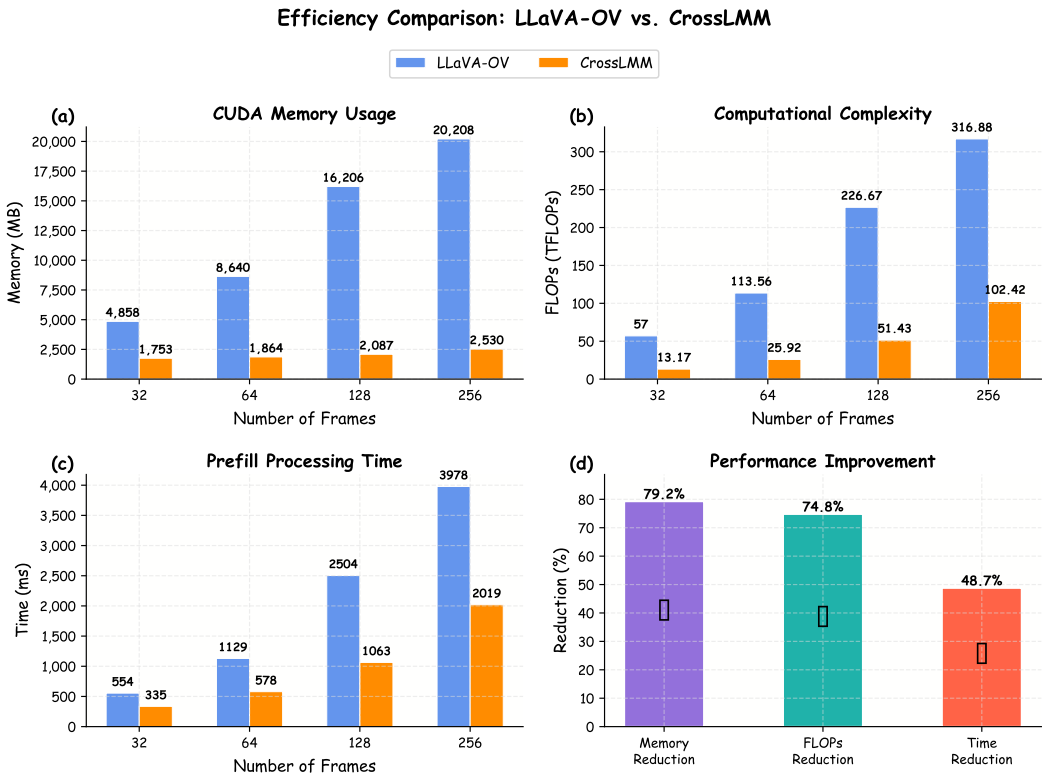


Figure 4: **Efficiency comparison between LLaVA-OV and CrossLMM across different frame counts (32, 64, 128, and 256).** (a) CUDA memory consumption measured in MB, showing CrossLMM's significantly lower memory footprint that scales more efficiently with increasing frames. (b) Computational complexity measured in TFLOPs, demonstrating CrossLMM's reduced computational requirements. (c) Prefill processing time measured in milliseconds, illustrating CrossLMM's faster processing capability. (d) Average performance improvement of CrossLMM over LLaVA-OV across all frame counts, showing substantial reductions in all metrics.

becomes increasingly pronounced as the frame count increases. At 256 frames, CrossLMM consumes just 2,531MB, representing an 87.5% reduction from LLaVA-OV's 20,208MB. Notably, the memory consumption of CrossLMM scales much more gradually with increasing frame counts, exhibiting an almost sub-linear growth pattern compared to LLaVA-OV's near-linear growth.

Computational Efficiency. Figure 4(b) illustrates the computational requirements measured in TFLOPs. CrossLMM consistently achieves lower computational complexity across all frame counts. At 32 frames, CrossLMM requires 13.17 TFLOPs, which is 76.9% less than LLaVA-OV's 57 TFLOPs. The computational advantage persists at higher frame counts, with CrossLMM requiring 102.42 TFLOPs at 256 frames compared to LLaVA-OV's 316.88 TFLOPs, representing a 67.7% reduction. This substantial decrease in computational demands contributes to CrossLMM's overall efficiency and potentially enables deployment on resources-limited devices.

Processing Time Efficiency. The prefill processing time, depicted in Figure 4(c), shows that CrossLMM consistently outperforms LLaVA-OV [26] in terms of speed. At 32 frames, CrossLMM completes processing in 335ms, which is 39.6% faster than LLaVA-OV's 554.91ms. This advantage is maintained across higher frame counts, with CrossLMM processing 256 frames in 1,975ms compared to LLaVA-OV's 3,978.83ms, representing a 50.4% reduction in processing time. The time efficiency of CrossLMM is particularly significant for real-time applications where latency is a critical factor.

Overall Efficiency Improvement. Figure 4(d) summarizes the average performance improvements achieved by CrossLMM across all tested frame counts. The most substantial gain is observed in memory usage, with an average reduction of 79.2%, followed by computational requirements at

Table 2: **Ablation studies on key components of our method.** (a) Visual-to-Visual (V2V) and Text-to-Visual (T2V) modules (✓: applied, ✗: not applied). (b) T2V and V2V modules insertion frequency K. (c) T2V activation stage (PT: pretraining, SFT: instruction-tuning).

(a) V2V and T2V modules

Modules		VideoMME	MVBench
V2V	*T2V*	*Overall*	*Avg.*
✗	✗	47.2	47.5
✗	✓	47.3	47.7
✓	✗	48.4	48.9
✓	✓	**48.7**	**49.9**

(b) Insertion frequency K

K	VideoMME	MVBench
	Overall	*Avg.*
None	47.2	47.5
1	48.5	47.8
2	48.5	49.7
4	**48.7**	**49.9**
8	47.7	48.5

(c) T2V activation stage

Modules		VideoMME	MVBench
PT	*SFT*	*Overall*	*Avg.*
✗	✗	48.4	48.9
✓	✓	48.0	47.8
✗	✓	**48.7**	**49.9**

74.8%, and processing time at 48.7%. These consistent improvements indicate that CrossLMM's architectural design effectively addresses the efficiency limitations of previous approaches.

The efficiency gains can be attributed to CrossLMM's novel approach to multi-modal processing, which employs a strategic token pruning to reduce redundancy in the video frame representations. On top of this, by leveraging V2V and T2V modules capture the multi-modal representations efficiently, CrossLMM minimizes computational overhead while maintaining model performance. These significant efficiency improvements enable CrossLMM to process longer video sequences with substantially lower resource requirements, making it more suitable for practical applications and deployment in resource-constrained environments.

4.4 Ablation Study

In this section, we present comprehensive ablation experiments to evaluate the contributions of individual components. Given computational resource constraints, we employ the CrossLMM-2B model as the foundation with 1 tokens for our ablation studies. During the pretraining phase, we utilize a balanced corpus comprising 3.75 million image-text pairs and 1.25 million video-text pairs, maintaining a consistent sampling ratio of 1:1 between modalities. For the instruction-tuning phase, we leverage the established LLaVA-Video 178K dataset to optimize model performance.

Visual-to-Visual and Text-to-Visual: The V2V module is specifically designed to capture text-conditioned fine-grained visual features from the original visual tokens. On top of this, the T2V module is designed to enhance the visual comprehension of the text information. As shown in Table 2a, in the absence of the V2V or T2V modules, performance decreases across all benchmarks, directly demonstrating the critical roles of the V2V and T2V modules in extracting fine-grained visual information and text features.

Insertion frequency: We conducted systematic ablation experiments by varying the insertion frequency K, which is shwon in Table 2b. When $K = 1$ (indicating maximum insertion frequency), we observed performance significantly below our proposed configuration, primarily attributable to the excessive number of parameters introduced, which compromises the inherent capabilities of the original LLM. Similarly, when $K = 8$, performance remains suboptimal compared to our method, suggesting that at this insertion frequency, the fine-grained visual information becomes overly sparse.

T2V activation stage: Our experimental design incorporates three configurations of the Text-to-Visual (T2V) module: (1) omission of the T2V module during both pretraining and instruction-tuning phases; (2) implementation of the T2V module throughout both pretraining and instruction-tuning phases; and (3) inclusion of the T2V module during pretraining but exclusion during instruction-tuning. The ablation study results, presented in Table 2c, yield two significant findings. First, the absence of the T2V module in both stages results in diminished performance across all benchmarks compared to our optimal configuration, demonstrating the module's efficacy in enhancing the visual representation of text tokens. Second, while incorporating the T2V module solely in the pretraining stage produces improvements, the efficacy is significantly attenuated due to the considerable textual noise present in pretraining data, such as alt-text descriptions.

5 Conclusion

We present CrossLMM, which offers a rigorous solution to the computational inefficiency that has long constrained LMMs when processing extended video sequences. Our CrossLMM framework demonstrates that through judicious token reduction strategies and architectural innovations, significant computational efficiency can be achieved while preserving model fidelity. The dual cross-attention mechanism, comprising V2V cross attention for maintaining fine-grained representational integrity and T2V cross attention for enhanced multimodal comprehension, constitutes a methodological advancement in token management for the practical real-world multimodal systems.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Ruichuan An, Sihan Yang, Ming Lu, Renrui Zhang, Kai Zeng, Yulin Luo, Jiajun Cao, Hao Liang, Ying Chen, Qi She, et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024.

[3] Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, Bocheng Zou, Chaoqun Yang, and Wentao Zhang. Unictokens: Boosting personalized understanding and generation via unified concept tokens, 2025.

[4] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024.

[5] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024.

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *CoRR*, abs/2404.16821, 2024.

[8] Jen-Hao Cheng, Vivian Wang, Huayu Wang, Huapeng Zhou, Yi-Hao Peng, Hou-I Liu, Hsiang-Wei Huang, Kuang-Ming Chen, Cheng-Yen Yang, Wenhao Chai, et al. Tempura: Temporal event masked prediction and understanding for reasoning in action. *arXiv preprint arXiv:2505.01583*, 2025.

[9] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024.

[10] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

[11] Rongyao Fang, Shilin Yan, Zhaoyang Huang, Jingqiu Zhou, Hao Tian, Jifeng Dai, and Hongsheng Li. Instructseq: Unifying vision tasks with instruction-conditioned multi-modal sequence generation. *arXiv preprint arXiv:2311.18835*, 2023.

[12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

[13] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[14] Peng Gao*, Jiaming Han*, Renrui Zhang*, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Ziyu Guo, Ray Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multi-modal scientific problems. *arXiv preprint arXiv:2503.10627*, 2025.

[17] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

[18] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

[19] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[20] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

[21] Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19079–19091, 2024.

[22] Minbin Huang, Runhui Huang, Han Shi, Yimeng Chen, Chuanyang Zheng, Xiangguo Sun, Xin Jiang, Zhenguo Li, and Hong Cheng. Efficient multi-modal large language models via visual token grouping. *arXiv preprint arXiv:2411.17773*, 2024.

[23] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

[24] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

[25] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326, 2024.

[27] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206. IEEE, 2024.

[30] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

[31] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, volume 15104 of *Lecture Notes in Computer Science*, pages 323–340. Springer, 2024.

[32] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable MLLMs to comprehend what you want. In *The Thirteenth International Conference on Learning Representations*, 2025.

[33] Ziyi Lin*, Chris Liu*, Renrui Zhang*, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *ECCV 2024*, 2023.

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[35] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024.

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[37] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pages 235–252. Springer, 2024.

[38] Feipeng Ma, Hongwei Xue, Yizhou Zhou, Guangting Wang, Fengyun Rao, Shilin Yan, Yueyi Zhang, Siying Wu, Mike Zheng Shou, and Xiaoyan Sun. Visual perception by large language model's weights. *Advances in Neural Information Processing Systems*, 37:28615–28635, 2025.

[39] Feipeng Ma, Yizhou Zhou, Zheyu Zhang, Shilin Yan, Hebei Li, Zilong He, Siying Wu, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. Ee-mllm: A data-efficient and compute-efficient multimodal large language model. *arXiv preprint arXiv:2408.11795*, 2024.

[40] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024.

[41] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[42] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, May 2024.

[43] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NIPS*, volume 36, 2024.

[44] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024.

[45] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

[46] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

[47] Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal llms. *arXiv preprint arXiv:2409.10994*, 2024.

[48] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[49] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[50] Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *arXiv preprint arXiv:2504.14693*, 2025.

[51] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024.

[52] Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Batista-Navarro. Lvpruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. *arXiv preprint arXiv:2501.13652*, 2025.

[53] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[55] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[57] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *CoRR*, abs/2409.02889, 2024.

[58] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[59] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. In *ECCV*, 2024.

[60] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

[61] Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. Dynaprompt: Dynamic test-time prompt tuning. *arXiv preprint arXiv:2501.16404*, 2025.

[62] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024.

[63] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

[64] Shilin Yan, Xiaohao Xu, Renrui Zhang, Lingyi Hong, Wenchao Chen, Wenqiang Zhang, and Wei Zhang. Panovos: Bridging non-panoramic and panoramic views with transformer for video segmentation. In *European Conference on Computer Vision*, pages 346–365. Springer, 2024.

[65] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6449–6457, 2024.

[66] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[67] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *arXiv preprint arXiv:2406.06040*, 2024.

[68] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024.

[69] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. *arXiv preprint arXiv:2409.10197*, 2024.

[70] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024.

[71] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[72] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *CoRR*, abs/2407.03320, 2024.

[73] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *CoRR*, abs/2406.16852, 2024.

[74] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ECCV 2024*, 2024.

[75] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*, 2024.

[76] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.

[77] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

[78] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024.

[79] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[80] Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N Metaxas, and Licheng Yu. Accelerating multimodel large language models by searching optimal vision token reduction. *arXiv preprint arXiv:2412.00556*, 2024.

[81] Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. Aim: Adaptive inference of multi-modal llms via token merging and pruning. *arXiv preprint arXiv:2412.03248*, 2024.

[82] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[83] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Implementation Details

A.1 Training Strategies and Data

Training Strategies. In contrast to current mainstream LMMs [6, 7, 56], which typically employ intricate three- or four-stage training pipelines, we argue that such methodologies introduce unnecessary complexity and accessibility barriers. To address this, we adopt a streamlined two-stage training framework inspired by earlier LMM practices [34, 83].

- **Stage I: Vision-Language Pretraining.** During this stage, CrossLMM aligns visual and linguistic representations using visual captioning datasets. Notably, due to the inherent noise in textual data, the text-to-visual module is intentionally excluded at this phase. Training instead focuses exclusively on optimizing the projection and visual-to-visual modules, while both the vision encoder and the LLM remain frozen to preserve their pretrained knowledge.

- **Stage II: Instruction-tuning.** In this stage, CrossLMM undergoes comprehensive training to execute diverse video-related tasks (e.g., question answering, multiple-choice assessment, etc.) utilizing computationally efficient vision tokens based on instruction-tuning data. The model's parameters are jointly optimized through end-to-end training methodology.

Training Data. In contrast to numerous contemporary LMMs that rely extensively on proprietary in-house datasets, CrossLMM exclusively utilizes publicly available open-source data for its development and training.

- **Stage I: Pretraining Data.** The pre-training corpus for CrossLMM consists of two primary components: (1) 15M image-text pairs from CapsFusion [70], and (2) 5M video-text pairs sourced from InternVid [58]. A sampling ratio of 1:1 is implemented between image-text and video-text pairs throughout the training process.

- **Stage II: Instruction Data.** The instruction-tuning data consists of about 3 million samples including image instruction data, short video instruction data and long video instruction data.

  - **Image Instruction data.** In our methodology, we employed a corpus of single-image instruction data derived from multiple established datasets, specifically LLaVA-NeXT [78], ALLaVA [5], and ShareGPT4-o [7, 59]. To enhance the model's capacity for processing complex visual inputs, we further augmented our training regime with multi-image sequential data obtained from LLaVA-Interleave [27]. This comprehensive data integration approach facilitated a more robust visual-linguistic representational framework.
  - **Short Video Instruction data.** For the instruction fine-tuning phase of our research, we predominantly employed short video sequences from VideoChat2 [29] and InternVideo2 [59] datasets. To enhance the model's comprehension capabilities, we further supplemented the training corpus with annotations derived via GPT4-o from several established datasets, including ShareGPT4o [7, 59], VideoChatGPT-Plus [40], LLaVA-Video-178K [79], and LLava-Hound [76]. This methodical integration of diverse video-based instructional data facilitated the development of a more comprehensive temporal-visual understanding framework.
  - **Long Video Instruction data.** In our experimental framework, we primarily utilized long-form video instruction datasets, specifically those sourced from MovieChat [49] and Vript [67], supplemented by LongVid corpus [30]. This methodological approach to data selection enabled comprehensive training on extended temporal sequences, thereby facilitating the model's capacity to process, interpret, and generate responses to complex narrative structures and sustained visual information across prolonged video segments.

A.2 Training hyperparameters.

As delineated in Table 3, we present a comprehensive documentation of the training protocols and associated hyperparametric configurations implemented across the sequential developmental stages of our CrossLMM model.

Table 3: **Training details of each training stage for the CrossLMM-2B model.**

		Stage 1	Stage 2
Vision	**Resolution×Num. frames**	384	384 ×8
	#Tokens	9 × 32	9 × (64∼512)
Data	**Dataset**	Image & Short Video	(Multi)-Image & Short/Long Video
	#Samples	20M	3M
Model	**Trainable**	Projector & T2V Cross Attention	Full Model
	#Parameters	43M	2B
Training	**Batch Size**	512	128
	LR of *vision encoder*	1×10^{-3}	2×10^{-6}
	LR *of connector & LLM*	1×10^{-3}	1×10^{-5}
	Epoch	1	1

Table 4: **Quantitative analysis of pretraining data volume impact on multi-benchmark performance.**

Data Volume	MVBench	PerceptionTest	LongVideoBench	VideoMME	VideoMME
	Avg	*Val*	*Val*	*w/o sub.*	*w sub.*
10M samples	51.4	60.4	48.1	53.1	56.4
15M samples	51.3	60.5	48.5	52.7	56.5
20M samples	**52.1**	**60.8**	**49.2**	**54.2**	**57.0**

B More Experiments

B.1 Ablation Study.

Pretraining data volume analysis. To establish a robust experimental foundation, we utilize the CrossLMM-2B model for conducting comprehensive ablation studies. For the instruction fine-tuning phase, we incorporate the LLaVA-Video-178K dataset. Our investigation focuses on the impact of varying pretraining data volume, with results presented in Table 4. This analysis aims to determine optimal data exposure during the critical pretraining stage.

Vision encoder trainability investigation. Table 5 presents our systematic investigation into vision encoder parameter adaptation strategies. We examine the differential effects between maintaining a frozen encoder versus allowing parameter updates during training. This comparison elucidates the significance of vision encoder plasticity on cross-modal representation learning and downstream performance across multiple evaluation benchmarks.

B.2 Computational Efficiency Analysis

The comprehensive efficiency analysis presented in Table 6 demonstrates how the CrossLMM-2B model scales with increasing frame counts under controlled experimental conditions. Memory utilization exhibits a sub-linear growth pattern, increasing from 715.61 MB at 32 frames to 1490.25 MB at 256 frames—approximately a 2.08× increase despite the 8× expansion in input dimensionality. This favorable scaling characteristic indicates efficient memory management within the model architecture.

Computational requirements, quantified in TFLOPs (Trillion Floating Point Operations), demonstrate a near-linear relationship with frame count, escalating from 10.93 TFLOPs to 86.82 TFLOPs across the evaluated range. This represents an 7.94× increase, closely approximating the theoretical linear scaling factor. Similarly, the prefill processing latency exhibits proportional growth, with measured times increasing from 264.06 ms to 1698.12 ms as the frame count expands.

These empirical measurements suggest that CrossLMM-2B maintains computational efficiency across varying temporal resolutions, with memory utilization scaling particularly well. Such characteristics

Table 5: **Comparative analysis of vision encoder training strategies.**

Parameter State	MVBench	PerceptionTest	LongVideoBench	VideoMME	VideoMME
	Avg	*Val*	*Val*	*w/o sub.*	*w sub.*
Frozen	60.8	61.3	51.0	56.7	59.2
Trainable	**63.5**	**63.5**	**51.8**	**59.1**	**61.3**

Table 6: **Computational efficiency metrics for CrossLMM-2B across varying temporal resolutions**

Model	Frame Count	CUDA Memory (MB)	FLOPs (TFLOPs)	Prefill Time (ms)
CrossLMM-2B	32	715.61	10.93	264.06
CrossLMM-2B	64	825.21	21.77	505.50
CrossLMM-2B	128	1046.68	43.46	909.77
CrossLMM-2B	256	1490.25	86.82	1698.12

are essential for applications requiring flexible processing of variable-length video inputs within constrained computational environments.

C Limitation and Future Work

Although CrossLMM has demonstrated promising results in video understanding, an important future direction lies in extending its framework to 2D image and 3D point cloud understanding. This expansion is particularly crucial for handling high-resolution images and large-scale 3D scenes, where capturing fine-grained spatial details while maintaining processing efficiency within LLMs remains challenging.

D Broader impacts

CrossLMM improves the efficiency of large multimodal models processing video, which may allow wider adoption in real-world applications such as smart surveillance, healthcare, and education. However, the enhanced accessibility of such technologies also raises potential ethical concerns, including risks to privacy and misuse for malicious purposes. We encourage responsible development and deployment of these systems, with attention to fairness, privacy protection, and prevention of harmful applications.