A Unified Image and Video Encoder for Multimodal LLMs

Tongtian Yue[1,2*], **Zikang Liu**[1,2*], **Longteng Guo**[1,2*], **Handong Li**[1,2], **Zhibin Wang**[3], **Jun Fu**[1,2], **Zijia Zhao**[4], **Xianing Chen**[3], **Jun Song**[3], **Yu Cheng**[3], **Bo Zheng**[3], **Jing Liu**[1,2†]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Alibaba Group [4] MoonShot AI

Abstract

Multimodal large language models (MLLMs) have rapidly advanced machine intelligence by integrating visual perception with powerful linguistic reasoning. However existing visual backbones predominantly encode videos in a rigid per-frame manner and delegate temporal modeling entirely to the language model. This paradigm fundamentally constrains long-range temporal understanding and imposes severe computational inefficiencies. To resolve these bottlenecks we propose **FuxViT**, a unified Vision Transformer natively supporting a 64K-token context window to seamlessly process diverse inputs ranging from static images to hour-scale videos. By employing a joint spatiotemporal pretraining paradigm, **FuxViT** generates highly coherent and information-rich visual representations. This strategic early fusion frees the downstream language model from reconstructing low-level temporal structures to focus entirely on high-level reasoning. Furthermore we introduce Frame Selective Attention (FSA) as a native sparse routing mechanism. FSA dynamically attends to the most relevant historical frames to achieve exceptional computational efficiency without sacrificing fine-grained spatial or temporal details. Extensive experiments across 16 image and video benchmarks demonstrate that **FuxViT** consistently outperforms existing baselines and establishes itself as a vastly superior alternative to standard static ViTs for multimodal understanding. We will fully open-source all model checkpoints.

1 Introduction

Multimodal large language models (MLLMs) [27, 19, 45, 3] have dramatically expanded the boundaries of machine intelligence by integrating large language models (LLMs) [1, 48, 40, 61] with visual perception. This integration allows abstract linguistic reasoning to be grounded in the physical world, enabling systems that can understand and reason about both images and videos. Achieving robust video comprehension requires capturing both the *spatial* semantics of individual frames [16, 18] and the *temporal* dynamics across frames [67, 55]. Yet, most existing visual backbones are designed for static images and fail to model these dimensions jointly, leaving a gap in MLLM capabilities.

Existing vision backbones, such as CLIP [39] and SigLIP [49], encode each video frame independently, producing per-frame tokens and deferring temporal reasoning entirely to the LLM. While this *per-frame encoding paradigm* is computationally simple, it introduces two key limitations. First, LLMs are pretrained on text and cannot naturally infer temporal correlations from isolated frame embeddings. Second, as video length increases, the LLM must process longer sequences of fragmented tokens, reducing reasoning efficiency and making it harder to capture long-range temporal dependencies. In effect, per-frame encoding forces the LLM to simultaneously reconstruct temporal structure and perform reasoning, which is unnatural and suboptimal.

*Equal Contribution. †Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

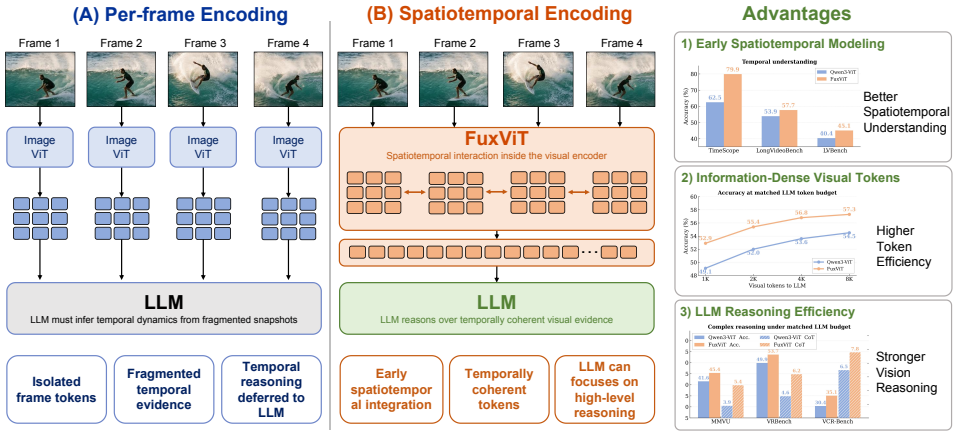

Figure 1: **Video encoding paradigms for MLLMs.** *(A) Per-frame Encoding:* Traditional ViTs encode frames independently, forcing the LLM to infer temporal dynamics from fragmented snapshots. *(B) Spatiotemporal Encoding (**FuxViT**):* Our approach models cross-frame interactions within the visual encoder, directly producing temporally coherent evidence. *Advantages:* This early temporal integration yields information-dense tokens, enhances spatiotemporal grounding, and frees the LLM to focus on complex reasoning rather than temporal reconstruction.

These limitations motivate a shift to a *spatiotemporal encoding paradigm*, where video structure is formed inside the visual encoder before reaching the LLM. As shown in Figure 1, this paradigm offers three intertwined advantages. Temporal dynamics are captured where visual continuity naturally exists, enabling stronger multi-frame reasoning. Cross-frame interactions enrich each token with relevant evidence, providing the LLM with coherent, information-dense inputs. With such consolidated tokens, the LLM can focus on high-level reasoning rather than reconstructing temporal order from fragmented embeddings.

While conceptually appealing, this paradigm introduces new challenges. Videos can be extremely long and high-resolution, making naive attention mechanisms computationally prohibitive. Moreover, training a visual encoder to produce coherent spatiotemporal tokens requires objectives that capture both fine-grained spatial structure and temporal evolution. Effectively balancing these constraints is crucial: the representations must be expressive enough to encode detailed visual dynamics, yet structured and compact enough to be efficiently consumed by the LLM.

Guided by the above limitations, we propose **FuxViT**, a unified Vision Transformer that seamlessly handles diverse visual inputs, from static images to hour-scale videos, for MLLMs. To our knowledge, **FuxViT** is the first ViT to natively scale its context window to 64K tokens, flexibly supporting dynamic spatiotemporal resolutions while preserving original aspect ratios.

At its core **FuxViT** leverages a joint spatiotemporal pretraining paradigm across a massive corpus of images and videos. Alongside standard text supervision, the model explicitly captures visual dynamics through Next-Patch Prediction and Next-Moment Prediction. This synergistic strategy yields highly coherent and information-rich visual tokens. Consequently the downstream LLM is freed from the burden of reconstructing low-level temporal orders and can focus entirely on high-level reasoning. Besides, to efficiently handle long videos, we propose Frame Selective Attention (FSA), a sparse attention mechanism that routes each query token to only the Top-K most relevant historical frames. By focusing on key temporal dependencies, FSA reduces unnecessary computation without sacrificing fine-grained detail or long-range spatiotemporal interactions.

Extensive experiments validate our approach. Under rigorous MLLM probing protocols, **FuxViT** achieves state-of-the-art performance across 16 image and video benchmarks, including tasks that require fine-grained temporal understanding and long-horizon reasoning. Ablation studies further confirm the synergy of dual autoregressive objectives and the computational efficiency of FSA.

In summary, our contributions are as follows:

- We propose **FuxViT**, a unified ViT with a 64K-token context window that can seamlessly process both static images and hour-scale videos with flexible spatiotemporal resolutions.

- We design a joint spatiotemporal pretraining paradigm that encourages the model to build coherent, information-rich visual representations, and introduce Frame Selective Attention (FSA) to efficiently capture long-range dependencies while reducing unnecessary computation without sacrificing fine-grained spatiotemporal detail.
- We demonstrate that **FuxViT** serves as an effective alternative to existing image ViTs in MLLMs, improving performance across 16 image and video benchmarks and supporting zero-shot context extrapolation up to 256K tokens.
- We will release **FuxViT** as an open-source model to promote future advances in vision understanding within multimodal LLMs.

2 Related Work

Advances in Visual Backbones. The evolution of visual front-ends has largely diverged into image- and video-centric trajectories. Image-centric backbones, spanning foundational ViTs [12], self-supervised frameworks like DINO [64], and contrastive models such as CLIP [39] and SigLIP [63, 49], establish robust spatial representations. Recent MLLM-tailored encoders [46, 3] further push these boundaries. However, they inherently process videos as isolated static snapshots, transferring the burden of temporal reasoning entirely to the LLM. Conversely, video-centric backbones capture spatiotemporal dynamics directly through masked modeling (*e.g.*, VideoMAE [47, 50], V-JEPA [5, 2]) or large-scale alignment [54, 53]. Yet, they heavily rely on fixed frame counts and rigid resolutions, often exhibiting suboptimal compatibility with pure image inputs. While recent unified architecture like OV-Encoder [44] attempt to bridge this gap, they still adhere to rigid temporal sampling. To our knowledge, **FuxViT** is the first visual backbone to overcome these constraints by natively supporting adaptive spatiotemporal resolutions. Equipped with a 64K-token context window, **FuxViT** enables a flexible trade-off between spatial granularity and temporal sampling rate within a unified budget. It dynamically adapts to tasks ranging from fine-grained static parsing to long-form motion tracking, without structural modifications.

Efficient Long-Video Processing. Long-video understanding is fundamentally bottlenecked by the quadratic complexity induced by massive visual sequences. Current MLLMs primarily address this challenge through three paradigms. (1) *Video sampling*. Existing methods utilize motion-aware selection [20] or token dropout [17, 30] to discard redundant frames or patches. However, this approach risks missing crucial transient events. (2) *Token compression*. Dedicated modules [6, 25, 10, 14] employ temporal pooling or query-based bottlenecks to aggregate visual semantics into condensed representations, yet such fusion often irreversibly blurs fine-grained spatiotemporal evidence. (3) *Sparse computation*. Recent MLLMs explore modality-aware patterns [26], streaming memory [42], or block selection [34] to dynamically allocate attention. By adaptively preserving visual tokens without rigid pruning, it generally yields a higher performance upper bound. Crucially, existing solutions adopt a decoupled paradigm: they rely on static backbones for spatial extraction, delegating temporal modeling and redundancy reduction to the LLM. In contrast, **FuxViT** intrinsically unifies both dimensions within ViT. By embedding Frame Selective Attention (FSA) directly into visual backbone, it enables spatial semantics and temporal dynamics to mutually reinforce each other, yielding robust and efficient long-range spatiotemporal representations.

3 Method

In this section, we present **FuxViT**, a unified vision backbone for inputs ranging from static images to long-form videos. We begin by formulating a native spatiotemporal tokenization scheme that supports variable spatial resolutions and frame rates within a 64K context budget. We then introduce Frame Selective Attention (FSA), a native block-causal mechanism that enables efficient and content-aware temporal interactions over frames. Finally, we detail the pretraining objectives, incorporating a standard language autoregressive loss alongside dual visual autoregressive objectives, designed to jointly foster unified representations of spatial coherence and temporal evolution.

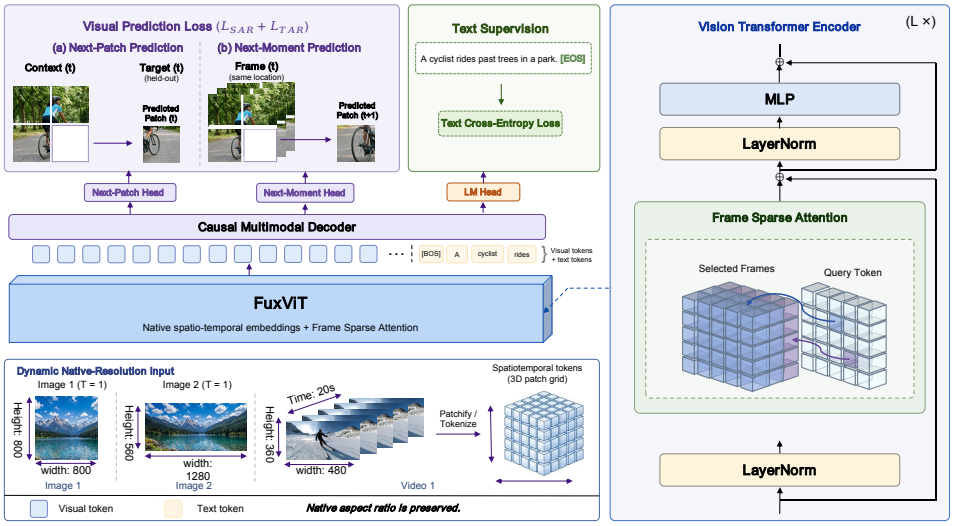

Figure 2: **The overall framework of FuxViT.** It natively ingests images and videos at their original spatiotemporal resolutions, processing them through a Vision Transformer equipped with sparse attention for efficient long-range modeling. The extracted visual tokens are then fed into a multimodal decoder and jointly optimized through standard text supervision alongside a dual visual autoregressive objective to enforce both intra-frame spatial coherence and inter-frame temporal dynamics.

3.1 Native Resolution Inputs for Images and Videos

To preserve the structural integrity of visual signals, **FuxViT** avoids rigid spatial resizing and fixed temporal sampling. Instead, any visual input is unified as a spatiotemporal volume $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$, where H and W maintain the native aspect ratio, and T denotes the frame count dynamically sampled at an arbitrary frame rate. A static image is naturally accommodated as the special case of $T = 1$. Given a patch size P, **FuxViT** allows for arbitrary adjustments to the spatiotemporal resolution within a maximum context budget L_{ctx}:

$$\mathcal{C} = \left\{ (T, H, W) \;\middle|\; T \times \left\lfloor \frac{H}{P} \right\rfloor \times \left\lfloor \frac{W}{P} \right\rfloor \leq L_{\text{ctx}} \right\} \tag{1}$$

This flexibility of configuration space $\mathcal{C}$ enables a seamless trade-off between spatial granularity and temporal coverage, allowing the model to dynamically adapt to tasks requiring either fine-grained visual details or high-frequency motion tracking without architectural modifications.

3.2 Frame Selective Attention

While expanding the context window empowers the visual backbone to capture long-range spatiotemporal dependencies, the quadratic complexity of dense self-attention renders this computationally prohibitive. To mitigate this, we propose Frame Selective Attention (FSA), which treats each frame as a semantically coherent unit to enable sparse yet content-adaptive interactions.

For a given input configuration $(T, H, W) \in \mathcal{C}$, each frame contributes $M = \lfloor \frac{H}{P} \rfloor \times \lfloor \frac{W}{P} \rfloor$ spatial tokens. The flattened token sequence X can thus be decomposed into T consecutive frame-level blocks $[\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_T]$, defined as:

$$\mathcal{B}_t = \{x_n \mid (t-1)M < n \leq tM\} \tag{2}$$

FSA elevates the temporal routing granularity from individual tokens to entire frames. Specifically, for a given layer and attention head, let q_n denote the query of a token $x_n \in \mathcal{B}_t$. FSA first evaluates the relevance of q_n against a candidate frame block $\mathcal{B}_s$ via:

$$a_{(n,s)} = \frac{q_n^\top \bar{k}_s}{\sqrt{d_h}} = \frac{q_n^\top}{\sqrt{d_h}} \left(\frac{1}{|\mathcal{B}_s|} \sum_{j\in\mathcal{B}_s} k_j \right) \tag{3}$$

where d_h is the head dimension and $\bar{k}_s$ is the mean key of all tokens within the frame block $\mathcal{B}_s$.

After quantifying frame-level relevance $a_{(n,s)}$, FSA defines the interaction scope based on visual structural priors. Since dynamics unfold temporally while individual frames capture cohesive spatial snapshots, we adopt a native *block causal* paradigm: intra-frame tokens interact bidirectionally to extract dense spatial semantics, whereas inter-frame interactions remain strictly causal to model temporal progression. However, given the massive temporal redundancy in videos, uniformly attending to all past frames is computationally prohibitive and dilutes representations. To resolve this, we employ a sparse routing strategy, leveraging the relevance scores $a_{(n,s)}$ to isolate highly correlated historical frames via a Top-K selection for x_n:

$$\mathcal{S}_n = \left\{ s \;\middle|\; 1 \le s < t,\ a_{(n,s)} \in \text{Top-K}\left(\{a_{(n,i)} \mid 1 \le i < t\}, K\right) \right\} \tag{4}$$

The attention mask matrix $\mathcal{M}$ for a query token $x_n \in \mathcal{B}_t$ and a key token $x_j \in \mathcal{B}_s$ is defined as:

$$\mathcal{M}_{n,j} = \begin{cases} 0, & \text{if } s \in \{t\} \cup \mathcal{S}_n, \\ -\infty, & \text{otherwise.} \end{cases} \tag{5}$$

The final output of FSA is then computed as:

$$\text{FSA}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + \mathcal{M}\right) V \tag{6}$$

FSA provides a unified temporal modeling mechanism that gracefully degenerates into standard full attention for single images, while acting as a highly efficient causal router for long-form videos.

3.3 Joint Autoregressive Pretraining

We pretrain **FuxViT** by tightly coupling it with a causal text decoder Θ. Given the extracted vision features $\boldsymbol{v} = \{\boldsymbol{v}_i\}_{i=1}^{V}$ by **FuxViT** and the corresponding text descriptions $\boldsymbol{t} = \{\boldsymbol{t}_i\}_{i=1}^{T}$, we project $\boldsymbol{v}$ into the semantic space of the text decoder via an MLP ϕ, and then concatenate it with $\boldsymbol{t}$. The decoder processes this unified multimodal sequence to yield the output hidden states:

$$[\mathcal{H}_v; \mathcal{H}_t] = \Theta([\phi(\boldsymbol{v}); \boldsymbol{t}]), \tag{7}$$

where $[\cdot\,;\,\cdot]$ denotes sequence concatenation. Here, $\mathcal{H}_v$ and $\mathcal{H}_t$ represent the decoupled output hidden states corresponding to the visual and textual positions, respectively.

Language Autoregressive Modeling. The language objective enforces high-level semantic alignment by applying the standard autoregressive loss over the logits from textual hidden states $\mathcal{H}_t$:

$$\mathcal{L}_{\text{LAR}} = -\sum_{i=1}^{T} \log p_\Theta(\boldsymbol{t}_i \mid \boldsymbol{t}_{<i}, \boldsymbol{v}). \tag{8}$$

Beyond the text-side supervision, we introduce two complementary visual autoregressive objectives to foster a unified representation of spatial coherence and temporal evolution. Specifically, for any given spatiotemporal index (t, h, w), we can uniquely isolate its corresponding input visual feature from $\boldsymbol{v}$ and its output hidden state from $\mathcal{H}_v$, denoted as $\boldsymbol{v}_{t,s}$ and $\boldsymbol{h}_{t,s}$ respectively. Here, t represents the temporal frame index, and s denotes the spatial patch index following a raster-scan order.

Next Patch Prediction. This objective aims to capture fine-grained structural dependencies within each individual frame. We employ a spatial prediction head g_s to map the current hidden state $\boldsymbol{h}_{t,s}$, which aggregates all preceding spatial context, into the feature space of the next patch. The input feature of the subsequent spatial patch, *i.e.*, $\boldsymbol{v}_{t,s+1}$, is the reconstruction target. The loss function is formulated to maximize the cosine similarity between the projected state and this target:

$$\mathcal{L}_{\text{SAR}} = -\frac{1}{|\Omega_\text{s}|} \sum_{(t,s)\in\Omega_\text{s}} \left(\frac{g_\text{s}(\boldsymbol{h}_{t,s})}{\|g_\text{s}(\boldsymbol{h}_{t,s})\|_2}\right)^\top \left(\frac{\text{sg}(\boldsymbol{v}_{t,s+1})}{\|\,\text{sg}(\boldsymbol{v}_{t,s+1})\|_2}\right) \tag{9}$$

To avoid degenerate solutions, following previous works [60, 9], we apply a stop-gradient operation $\text{sg}(\cdot)$ to the target embeddings. Here, Ω_s represents the set of valid intra-frame prediction positions.

Next Moment Prediction. To enforce temporal consistency, we employ a temporal head g_t to predict the patch embedding at the identical spatial location s in the subsequent frame:

$$\mathcal{L}_{\text{TAR}} = -\frac{1}{|\Omega_\text{t}|} \sum_{(t,s)\in\Omega_\text{t}} \left(\frac{g_\text{t}(\boldsymbol{h}_{t,s})}{\|g_\text{t}(\boldsymbol{h}_{t,s})\|_2}\right)^\top \left(\frac{\text{sg}(\boldsymbol{v}_{t+1,s})}{\|\,\text{sg}(\boldsymbol{v}_{t+1,s})\|_2}\right) \tag{10}$$

where Ω_t contains all visual positions with a valid subsequent frame. The unified pretraining objective of **FuxViT** is thus:

$$\mathcal{L} = \mathcal{L}_{\text{LAR}} + \lambda_s \mathcal{L}_{\text{SAR}} + \lambda_t \mathcal{L}_{\text{TAR}} \tag{11}$$

3.4 Progressive Training Stages

We initialize the vision backbone of **FuxViT** with SigLIP2-So400M-384 [49] and the text decoder with Qwen3-0.6B [61]. To systematically foster holistic spatiotemporal understanding, we design a four-stage progressive training curriculum. The details of each stage are summarized in Table 1, and more details about the training settings and data are shown in the Appendix.

Table 1: **Overview of training stages**. It includes context lengths, attention modes, trainable parameters, training token volumes, and objective states.

Configuration	Stage 1 *Alignment*	Stage 2 *Image*	Stage 3 *Short Video*	Stage 4 *Long Video*
Context Length	4K	4K	16K	64K
Attention Mode	Full	Full	Block-Causal	FSA
Trainable Module	MLP	All	All	All
Training Tokens	10B	200B	400B	300B

Specifically, **Stage 1** serves as an initial warm-up phase to bridge the inter-modality gap. Optimized exclusively on image-text pairs, this stage establishes a robust cross-modal alignment, projecting visual features directly into the semantic latent space of the text decoder. Building upon this foundation, **Stage 2** discards the original absolute positional embeddings in favor of VRoPE [32], alongside adapting the image processor to accommodate dynamic aspect ratios. Supervised by high-quality recaptioned image data, this phase empowers the model with fine-grained spatial perception across diverse native resolutions. In **Stage 3**, we extend this established spatial mastery into the temporal dimension. By transitioning from standard full attention to a dense *block-causal* paradigm, **FuxViT** unlocks its preliminary video understanding capabilities. The training corpus is expanded to a mixture of static images, short videos, and long videos, sampled at a token ratio of 4:5:1. Finally, **Stage 4** activates the proposed FSA to achieve efficient temporal modeling over hour-scale videos. To explicitly optimize for extended temporal horizons, the data mixture ratio is adjusted to 4:3:3.

4 Experiment

4.1 Experimental Settings

Implementation Details. We trained **FuxViT** on 128 NVIDIA H100 GPUs for 8 days using the AdamW optimizer [33]. The comprehensive pretraining process spans four progressive stages. Across these stages, the batch sizes are set to 256, 1024, 512, and 512, with maximum learning rates of 1×10^{-4}, 2×10^{-5}, 2×10^{-5}, and 1×10^{-5}, respectively. A warm-up ratio of 0.03 and a cosine decay learning rate scheduler are consistently applied throughout all stages. During training, the visual token budget for static image inputs is capped at 4096. For video inputs, the frame rate is randomly sampled from 0.5 to 5.0 FPS with a step size of 0.5, while a default rate of 1 FPS is adopted for downstream evaluation. Additionally, the maximum token count per video frame is restricted to 1024. Whenever the initial token count exceeds either the global 64K context budget or the respective single-frame limits, we proportionally downscale the spatiotemporal resolution to strictly satisfy all constraints. For the auxiliary visual objectives, both prediction heads are implemented as 8-layer MLPs, and their corresponding loss weights, λ_s and λ_t, are both set to 0.5. Finally, we set the routing parameter K to 32 for FSA.

Evaluation Protocol. To evaluate the representation quality of different visual backbones, we adopt the *MLLM Probing Evaluation* protocol. Specifically, we couple various ViTs with the same LLM and train them using identical data recipes for cross-modal alignment and instruction tuning. The resulting MLLMs are then systematically benchmarked using the LMMs-eval [65] framework. To ensure that our findings are robust and backbone-agnostic, we employ two distinct LLMs: Qwen2-7B-Instruct [62] and Qwen3-8B [61]. For image-centric evaluation, the models are trained on LLaVA-558K [27] for the alignment stage and LLaVA-OneVision [23] for the instruction-tuning stage. The evaluation spans a comprehensive suite of image benchmarks, including MMBench [29], OCRBench [31], TextVQA [43], DocVQA [38], ChartQA [36], InfoVQA [37], AI2D [21], RealWorldQA [57], and MMStar [8]. For video-centric evaluation, we utilize Valley-702K [35] and

Table 2: **Performance comparison on 9 image-centric benchmarks.** "Res" and "Param" denote the supported spatial resolution and the parameter count of the visual backbone, respectively. Overall, **FuxViT** achieves the highest average accuracy across both LLM configurations.

Model Setup			Image Benchmarks (Acc. %)									
Method	**Res**	**Param**	**MMBench**	**OCRBench**	**TextVQA**	**DocVQA**	**ChartQA**	**InfoVQA**	**AI2D**	**RWQA**	**MMStar**	**Avg.**
LLM Backbone: Qwen2-7B-Instruct												
CLIP	336	304M	68.9	28.7	58.7	50.4	47.4	39.1	61.0	55.0	45.9	50.6
AIMv2	336	310M	70.6	30.2	60.4	49.7	50.7	38.8	63.1	55.0	46.4	51.7
AIMv2	448	310M	71.8	34.6	64.2	62.3	53.0	42.6	63.7	56.2	48.1	55.2
OpenVision2	336	304M	70.4	30.7	61.5	52.7	49.1	40.4	62.9	56.9	48.2	52.5
OpenVision2	384	400M	70.0	29.6	61.8	56.3	50.0	39.7	69.4	57.4	47.5	53.5
SigLIP2	384	428M	75.2	57.8	69.2	63.6	64.2	40.0	75.0	57.9	51.8	61.6
SigLIP2	Any	428M	72.9	57.1	69.2	67.9	67.9	43.0	72.1	56.5	49.7	61.8
Qwen3-ViT	Any	576M	71.5	71.6	76.2	87.8	73.2	66.9	69.2	60.8	47.5	69.4
OV-Ecnoder	Any	303M	74.5	70.0	76.6	89.2	77.0	66.6	72.9	60.4	53.4	71.2
FuxViT	Any	412M	**76.5**	**72.8**	**79.3**	**90.4**	**78.4**	**71.4**	**77.0**	**62.6**	**55.9**	**73.8**
LLM Backbone: Qwen3-8B												
CLIP	336	304M	73.7	26.8	56.5	48.7	46.1	38.6	63.5	57.3	48.2	51.0
AIMv2	336	310M	74.3	31.4	60.7	51.2	53.2	40.4	66.0	57.0	48.3	53.6
AIMv2	448	310M	75.0	35.3	64.0	63.0	55.2	43.6	65.4	56.6	50.1	56.5
OpenVision2	336	304M	76.1	32.0	60.4	51.1	50.8	38.8	68.0	58.0	49.9	53.9
OpenVision2	384	400M	74.1	33.7	62.5	57.1	51.6	41.3	63.0	57.7	47.0	54.2
SigLIP2	384	428M	79.9	60.3	69.7	64.0	66.2	41.0	64.1	60.4	55.3	62.3
SigLIP2	Any	428M	80.2	64.7	72.9	73.1	72.5	48.8	66.9	59.9	53.6	65.8
Qwen3-ViT	Any	576M	77.3	75.0	78.1	89.9	78.2	71.0	76.1	60.5	52.8	73.2
OV-Ecnoder	Any	303M	79.4	70.7	77.8	90.3	78.8	69.5	**77.0**	62.9	58.5	73.9
FuxViT	Any	412M	**80.6**	**76.1**	**80.7**	**92.2**	**81.2**	**74.1**	76.4	**66.3**	**59.2**	**76.3**

Table 3: **Performance comparison on seven video benchmarks.** "Res", "Frame", and "Param" denote the supported spatial resolution, the supported frames for jointly encoding, and the parameter count of the visual backbone, respectively. The asterisk (*) indicates that image-centric ViTs inherently process only 1 frame at a time. Their reported results are obtained by independently encoding 32 sampled frames and concatenating the features.

Model Setup				Video Benchmarks (Acc. %)							
Method	**Res**	**Frame**	**Param**	**VideoMME**	**MVBench**	**LVBench**	**MLVU**	**NExT-QA**	**LongVideoBench**	**TimeScope**	**Avg.**
LLM Backbone: Qwen2-7B-Instruct											
CLIP	336	1*	304M	58.1	57.6	40.4	61.2	79.0	52.3	57.1	58.0
SigLIP2	Any	1*	428M	58.7	57.9	40.2	61.8	80.5	52.8	62.1	59.1
Qwen3-ViT	Any	1*	576M	58.5	58.0	40.6	60.8	78.6	51.8	62.0	58.6
VideoMAEv2	224	16	304M	47.3	52.4	33.6	51.1	71.0	46.8	38.6	48.7
V-JEPA-2	256	64	326M	49.8	56.2	35.7	54.9	76.9	50.3	49.0	53.3
InternVideo-Next	224	16	311M	54.1	57.4	36.6	58.1	80.1	50.9	50.6	55.4
OV-Encoder	224	64	303M	59.2	56.7	41.9	61.0	79.7	53.5	64.2	59.5
FuxViT	Any	Any	412M	**61.0**	**59.2**	**44.7**	**62.5**	**80.5**	**56.6**	**74.7**	**62.7**
LLM Backbone: Qwen3-8B											
CLIP	336	1*	304M	60.4	58.0	40.9	63.4	80.0	53.1	60.7	59.5
SigLIP2	Any	1*	428M	60.8	59.0	40.8	62.8	79.5	54.1	63.6	60.1
Qwen3-ViT	Any	1*	576M	59.6	58.8	40.4	63.1	78.8	53.9	62.5	59.6
VideoMAEv2	224	16	304M	49.1	50.2	35.1	52.2	68.1	48.6	40.4	49.1
V-JEPA-2	256	64	326M	52.2	53.9	36.1	57.8	76.3	50.5	51.7	54.1
InternVideo-Next	224	16	311M	54.4	57.5	36.7	60.5	79.9	52.6	53.6	56.5
OV-Encoder	224	64	303M	60.9	58.2	42.4	63.6	80.2	54.5	66.0	60.8
FuxViT	Any	Any	412M	**63.6**	**60.6**	**45.1**	**65.5**	**81.0**	**57.7**	**79.9**	**64.8**

LLaVA-Video-178K [66] for the respective two training stages. The resulting models are tested on prominent video benchmarks, including Video-MME [13], MVBench [24], MLVU [68], NExT-QA [58], LongVideoBench [56], LVBench [52] and TimeScope [28].

4.2 Main Results

Image Understanding Performance. Table 2 reports the results across 9 image-centric benchmarks, where **FuxViT** consistently outperforms existing visual backbones. When coupled with Qwen2-7B-Instruct, it achieves the highest average accuracy, surpassing strong baselines like OV-

Encoder and the parameter-heavy Qwen3-ViT by 2.6% and 4.4%, respectively. This performance edge scales seamlessly to the stronger Qwen3-8B, where **FuxViT** maintains a decisive 2.4% lead over the second-best model. Notably, **FuxViT** exhibits remarkable superiority in fine-grained tasks such as DocVQA, OCRBench, and ChartQA. These substantial gains validate that native spatiotemporal resolution and spatial autoregressive pretraining effectively capture intricate structural details.

Video Understanding Performance. Table 3 reports the results across 7 diverse video benchmarks, demonstrating **FuxViT**'s consistent superiority over both image-centric and video-centric backbones. When paired with Qwen2-7B-Instruct, **FuxViT** achieves the highest average accuracy, outperforming the strongest unified baseline, OV-Encoder, by 3.2%. This advantage scales seamlessly to the stronger Qwen3-8B backbone, where **FuxViT** secures a decisive 4.0% lead over the second-best model. Notably, **FuxViT** exhibits exceptional proficiency in tasks demanding complex temporal reasoning and long-context dependency. For instance, on TimeScope with the Qwen3-8B backbone, our model achieves an impressive 79.9%, outstripping OV-Encoder by a massive 13.9%. Similarly, on LongVideoBench, **FuxViT** seamlessly processes extended temporal contexts, decisively outperforming all rigid-sampling baselines. These profound gains validate that our native FSA and visual autoregressive pretraining successfully empower the backbone to dynamically route and model critical continuous dynamics across extended horizons.

4.3 Ablation Study

To thoroughly investigate the key architectural designs and strategies of **FuxViT**, we conduct extensive ablation studies. Unless otherwise specified, all ablated variants are trained on a 300B-token subset that strictly maintains the original data mixture proportions. For downstream multimodal evaluation, these visual backbones are consistently coupled with the Qwen2-7B-Instruct.

Effect of Visual Objectives. We first isolate the contributions of dual visual autoregressive pretraining. Specifically, we activate or deactivate the spatial ($\mathcal{L}_{\text{SAR}}$) and temporal ($\mathcal{L}_{\text{TAR}}$) prediction, while maintaining the standard language autoregressive loss as the baseline. As shown in Table 4, relying solely on the language loss yields suboptimal performance. Adding the spatial-only objective ($\mathcal{L}_{\text{SAR}}$) boosts fine-grained structural comprehension for image tasks. Conversely, employing the temporal-only objective ($\mathcal{L}_{\text{TAR}}$) enhances temporal understanding in video tasks while exerting minimal impact on spatial perception. Combining both objectives achieves the highest overall score. It validates our core hypothesis: spatial layouts and temporal dynamics share fundamental visual priors, which can be unified into a robust spatiotemporal representation through proper objectives.

Table 4: **Ablation of visual autoregressive objectives.** We compare four pretraining configurations: the standard language-only baseline, spatial-only ($\mathcal{L}_{\text{SAR}}$), temporal-only ($\mathcal{L}_{\text{TAR}}$), and their joint formulation. The default setting is highlighted in blue.

Objectives		Image Tasks		Video Tasks		Avg.
$\mathcal{L}_{\text{SAR}}$	$\mathcal{L}_{\text{TAR}}$	AI2D	DVQA	TScope	LVBench	
✗	✗	69.8	86.6	65.0	39.3	65.2
✓	✗	**71.7**	**89.0**	65.6	40.4	66.7
✗	✓	69.4	86.8	66.5	41.2	66.0
✓	✓	71.4	88.2	**67.3**	**41.5**	**67.1**

Effect of FSA Strategy. FSA is proposed to alleviate the quadratic complexity of long-video processing while preserving critical long-range dependencies. To validate its design, we ablate three core components in Table 5, measuring both computational efficiency and performance. *(1) Attention Choice:* Compared to full attention and block-causal attention, FSA achieves highly competitive performance while requiring only ∼25% of the computational overhead (70.3T *vs.* 253.5T). This underscores its capability to effectively compress temporal redundancy without sacrificing accurate long-range interactions. *(2) Top-K Choice:* Increasing

Table 5: **Ablation of FSA.** We compare attention type, Top-K budget, and block granularity. TFLOPs measure ViT-side computation.

Settings	TFLOPs	LVBench	MLVU	TScope	Avg.
Attention Choice					
Full	253.5	42.0	**61.9**	67.0	**57.0**
Block	149.5	**42.4**	61.1	66.2	56.6
FSA	70.3	41.5	60.6	**67.3**	56.5
Top-K Choice					
16	58.5	40.4	59.8	66.0	55.4
32	70.3	41.5	**60.6**	67.3	56.5
64	91.4	**41.9**	60.0	**68.6**	**56.8**
Block Size Choice					
64	56.0	38.7	58.0	64.6	53.8
256	83.7	39.8	58.5	65.9	54.7
1024	148.1	40.9	59.3	66.4	55.5
Frame	70.3	**41.5**	**60.6**	**67.3**	**56.5**

the routing parameter K enlarges the temporal context but inherently incurs higher computational costs. Scaling K from 16 to 32 yields substantial gains while further increasing it to 64 brings negligible improvements despite surging FLOPs. Thus we adopt $K = 32$ for the optimal efficiency-performance balance. *(3) Block Size Choice:* Finally we evaluate the interaction block granularity. Utilizing fixed token-count splits (*e.g.*, 256 or 1024) consistently degrades performance even at higher costs because arbitrary truncation severely disrupts spatial coherence. In contrast our native frame-wise strategy achieves the best overall performance. This validates that an individual frame naturally serves as the optimal semantic unit for robust temporal modeling.

4.4 Visualization and Analysis

Computational Efficiency Analysis. We further evaluate the computational efficiency of **FuxViT** against Qwen3-ViT as our single-frame baseline. Figure 3 plots the average performance against total FLOPs (ViT & LLM) by sweeping the frame count from 8 to 128. The resulting substantial performance gains demonstrate that **FuxViT** achieves superior representation quality at any given computational budget. Notably our model at only 16 frames matches the accuracy of the baseline at 64 frames while requiring far fewer total FLOPs. This dramatic efficiency gap confirms that modeling shared spatiotemporal visual priors within the visual backbone is significantly more effective than increasing the sampling density of isolated frames. It underscore the fundamental necessity of early spatiotemporal fusion for reaching the optimal frontier of long-video reasoning efficiency.

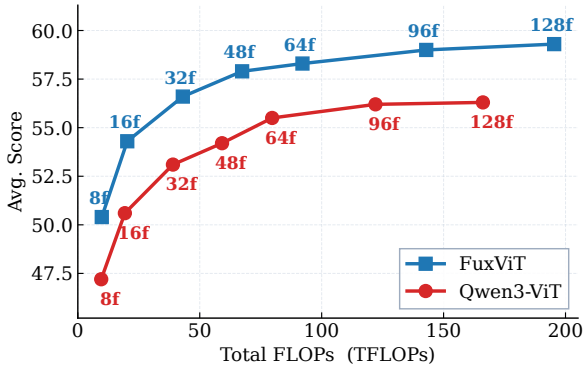

Figure 3: **Performance versus total FLOPs. FuxViT** consistently achieves superior representation quality at any given computational budget compared to the single-frame Qwen3-ViT.

Context Extrapolation. We evaluate the zero-shot extrapolation capability of **FuxViT** beyond the 64K-token training budget. As illustrated in Figure 4, the model demonstrates a steady upward trend within the training limits. More importantly it exhibits continuous performance gains when scaling the inference context up to 128K and 256K tokens without requiring any fine-tuning. The average score successfully climbs from 60.6 at the 64K boundary to a peak of 62.8 at 256K. This remarkable generalization firmly validates the robustness of **FuxViT** under unprecedented sequence lengths. It ultimately highlights its immense potential for ultra-long video understanding.

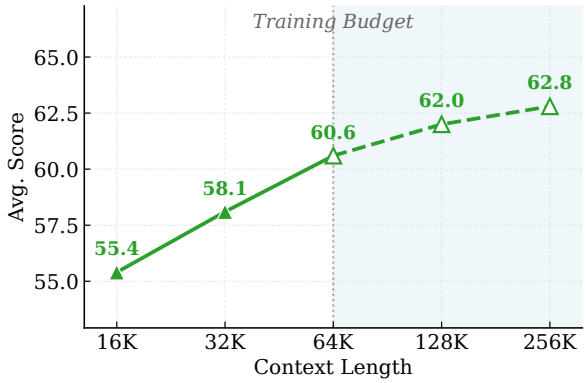

Figure 4: **Zero-shot context extrapolation. FuxViT** maintains continuous performance improvements when scaling the inference context up to 256K tokens well beyond its 64K training budget.

Frame Selection Visualization. We examine how FSA can access relevant visual evidence across temporally separated events. Figure 5 aligns historical frames with their Top-K selection frequencies for the current frame, aggregated across attention heads and spatial positions at a fixed layer. Specifically, a remote control reappears in the current frame after a phone demonstration. Earlier frames highlighted in yellow containing the same object form a sustained region of elevated selection frequency while most intervening frames receive lower scores. This pattern illustrates the behavior

of FSA: maintaining access to visually relevant historical content across long temporal gaps through content-based frame selection.

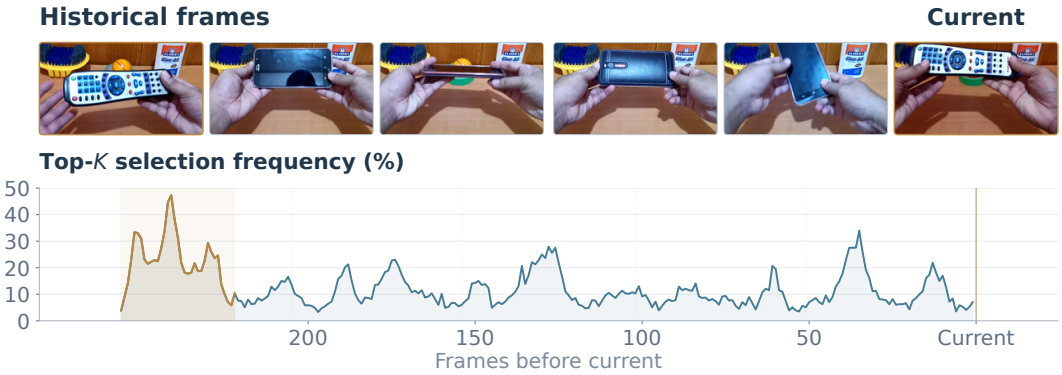

Figure 5: **Temporal selectivity of FSA.** Historical frames are aligned with their Top-K selection frequencies for the current frame. Earlier appearances of the remote control exhibit higher selection frequencies than most intervening frames, highlighting visually relevant content across a long temporal gap.

5 Conclusion

We present **FuxViT**, a unified Vision Transformer that generates high-information-density visual tokens for both images and long-form videos at arbitrary spatiotemporal resolutions. By adopting an early spatiotemporal encoding paradigm, the model consolidates spatial and temporal structures before reaching the language model, enabling more coherent and efficient reasoning. Extensive experiments on 16 image and video benchmarks demonstrate that **FuxViT** consistently outperforms existing backbones and can serve as an effective alternative to standard image ViTs in MLLMs. This work highlights the benefits of unified spatiotemporal visual representations, providing a scalable framework for fine-grained and long-range vision understanding in MLLMs.

6 Limitations and Future Work

FuxViT currently uses a fixed historical-frame routing budget K across all layers. This static allocation may be suboptimal for videos with different temporal complexity and context requirements. Future work will investigate adaptive routing budgets that allocate computation according to visual content. In addition, our evaluation focuses on an approximately 400M-parameter backbone. Its scaling behavior toward compact models for resource-constrained deployment and billion-parameter models remains to be systematically characterized. Exploring these model scales will help clarify the trade-offs among representation quality, training cost, and inference efficiency.

Acknowledgments

This research is supported by the Artificial Intelligence–National Science and Technology Major Project (Grant No. 2023ZD0121200); the National Natural Science Foundation of China (Grant Nos. 62531026, 62437001, and 62436001); the Natural Science Foundation of Jiangsu Province (Grant No. BK20243051); the Development Program of Jiangsu Province (Grant No. BE2023016-3); and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB1350103).

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. 1

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 3

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1, 3

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 16

[5] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023. 3

[6] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3

[7] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 16

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 5

[10] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 16

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[13] Chaoyou Fu, Yuhan Dai, Yongdong 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025. 7

[14] Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. *arXiv preprint arXiv:2501.01986*, 2024. 3

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 16

[16] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024. 1

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. Turbo training with token dropout. *arXiv preprint arXiv:2210.04889*, 2022. 3

[18] Runhui Huang, Xinpeng Ding, Chunwei Wang, Jianhua Han, Yulong Liu, Hengshuang Zhao, Hang Xu, Lu Hou, Wei Zhang, and Xiaodan Liang. Hires-llava: Restoring fragmentation input in high-resolution large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29814–29824, 2025. 1

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[20] Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024. 3

[21] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 6

[22] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 16

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[24] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 7

[25] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 3

[26] Yucheng Li, Huiqiang Jiang, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Amir H Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, et al. Mminference: Accelerating pre-filling for long-context vlms via modality-aware permutation sparse attention. *arXiv preprint arXiv:2504.16083*, 2025. 3

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 6, 16

[28] Xiangrui Liu, Minghao Qin, Yan Shu, Zhengyang Liang, Yang Tian, Chen Jason Zhang, Bo Zhao, and Zheng Liu. Timescope: Towards task-oriented temporal grounding in long videos. *arXiv preprint arXiv:2509.26360*, 2025. 7

[29] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 6

[30] Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, and Kevin Smith. Patchdropout: Economizing vision transformers using patch dropout. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3953–3962, 2023. 3

[31] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 6

[32] Zikang Liu, Longteng Guo, Yepeng Tang, Tongtian Yue, Junxian Cai, Kai Ma, Qingbin Liu, Xi Chen, and Jing Liu. Vrope: Rotary position embedding for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14471–14483, 2025. 6

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[34] Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, et al. Moba: Mixture of block attention for long-context llms. *arXiv preprint arXiv:2502.13189*, 2025. 3

[35] Ruipu Luo, Ziwang Zhao, Min Yang, Zheming Yang, Minghui Qiu, Zhongyu Wei, Yanhao Wang, and Cen Chen. Valley: Video assistant with large language model enhanced ability. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 6

[36] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 6

[37] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 6

[38] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 6

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 3

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 16

[42] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169, 2025. 3

[43] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6

[44] Feilong Tang, Xiang An, Yunyao Yan, Yin Xie, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Chunyuan Li, Shikun Feng, et al. Onevision-encoder: Codec-aligned sparsity as a foundational principle for multimodal intelligence. *arXiv preprint arXiv:2602.08683*, 2026. 3

[45] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[46] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026. 3

[47] 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. 3

[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[49] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 1, 3, 6

[50] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 3

[51] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 16

[52] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025. 7

[53] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pages 396–416. Springer, 2024. 3

[54] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 3

[55] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 1

[56] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 7

[57] X.ai. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, 2024. 6

[58] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 7

[59] Yin Xie, Kaicheng Yang, Xiang An, Kun Wu, Yongle Zhao, Weimo Deng, Zimin Ran, Yumeng Wang, Ziyong Feng, Roy Miles, et al. Region-based cluster discrimination for visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1793–1803, 2025. 16

[60] Sihan Xu, Ziqiao Ma, Wenhao Chai, Xuweiyi Chen, Weiyang Jin, Joyce Chai, Saining Xie, and Stella X Yu. Next-embedding prediction makes strong vision learners. *arXiv preprint arXiv:2512.16922*, 2025. 5

[61] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 6

[62] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 6

[63] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3

[64] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[65] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025. 6, 18

[66] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 7

[67] Zijia Zhao, Yuqi Huo, Tongtian Yue, Longteng Guo, Haoyu Lu, Bingning Wang, Weipeng Chen, and Jing Liu. Efficient motion-aware video mllm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24159–24168, 2025. 1

[68] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701, 2025. 7

A Training Details

Table 6: Detailed hyperparameter settings across the four progressive training stages of FuxViT. We consistently adopt the AdamW optimizer with a cosine learning rate scheduler.

Hyperparameter	**Stage 1: Alignment**	**Stage 2: Image**	**Stage 3: Short Video**	**Stage 4: Long Video**
Optimizer	AdamW	AdamW	AdamW	AdamW
Base Learning Rate	1×10^{-4}	2×10^{-5}	2×10^{-5}	1×10^{-5}
Weight Decay	0.05	0.05	0.05	0.05
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.999$	$\beta_1, \beta_2 = 0.9, 0.999$	$\beta_1, \beta_2 = 0.9, 0.999$	$\beta_1, \beta_2 = 0.9, 0.999$
LR Schedule	Cosine Decay	Cosine Decay	Cosine Decay	Cosine Decay
Warmup Ratio	0.03	0.03	0.03	0.03
Global Batch Size	256	1024	512	512
Numerical Precision	bfloat16	bfloat16	bfloat16	bfloat16
DeepSpeed Stage	Zero2	Zero3	Zero3	Zero3
Max Spatial Resolution	384×384	Dynamic (up to 4K)	Dynamic (up to 4K)	Dynamic
Max Frames	1 (Image)	1 (Image)	16	128
Context Budget (L_{ctx})	4K	4K	16K	64K
FSA Top-K (K)	-	-	-	32
Loss Weights (λ_s, λ_t)	-	0.5, 0.0	0.5, 0.5	0.5, 0.5

In this section, we provide the comprehensive hyperparameter configurations used across the four progressive training stages of FuxViT, as summarized in Table 6. We consistently employ the AdamW optimizer with a weight decay of 0.05 and a cosine learning rate decay schedule, preceded by a 3% warmup ratio. To ensure stable and memory-efficient scaling across tens of billions of tokens, we utilize DeepSpeed ZeRO optimization, transitioning from ZeRO-2 in the initial alignment stage to ZeRO-3 in subsequent stages, all executed under `bfloat16` numerical precision.

As the model progresses through the training curriculum, the global context budget (L_{ctx}) is incrementally expanded from 4K tokens in Stages 1 and 2, to 16K in Stage 3, and finally reaching an expansive 64K in Stage 4. This expansion is intrinsically tied to the maximum supported frames, evolving from single-image processing to jointly encoding up to 128 frames for long-video sequences. Furthermore, the auxiliary visual autoregressive loss weights (λ_s and λ_t) are introduced post-alignment (set to 0.5) to enforce both intra-frame spatial coherence and inter-frame temporal consistency. For our proposed Frame Selective Attention (FSA) mechanism, introduced in Stage 4 to manage the extended temporal horizon, we set the routing parameter K to 32.

B Training Data

Table 7: Overview of the training datasets across the four progressive stages of FuxViT. The mixture is dynamically adjusted, incorporating image replays in video stages to preserve fine-grained spatial comprehension.

Training Stage	**Dataset Name**	**Data Type**	**Samples**
Stage 1: Alignment	LLaVA-558K [27]	Image-Text Alignment	558K
Stage 2: Image	LAION-400M [41]	WebImages	250M
	COYO-700M [7]	WebImages	400M
	OBELICS [22]	Documents	15M
	Zero250M [59]	CuratedImages	15M
	ImageNet-21K [11]	Images	14M
	DataComp [15]	Image-Text Mixture	1B
Stage 3: Short Video	Stage 2 Dataset Replay	Image Mixture	-
	WebVid-5.7M [4] & Koala-20M [51]	Short Videos ($\leq$ 10 min)	$\sim$24.7M
Stage 4: Long Video	Stage 2 Dataset Replay	Image Mixture	-
	WebVid & Koala Subset	Long Videos ($>$ 10 min)	1M

Table 7 details the composition of the diverse datasets utilized throughout our pretraining curriculum. Our data scaling strategy is meticulously designed to progressively build spatiotemporal comprehension while actively preventing the catastrophic forgetting of fine-grained spatial features.

In **Stage 1 (Alignment)**, we exclusively use 558K image-text pairs from LLaVA-558K to establish a foundational cross-modal semantic mapping. **Stage 2 (Image)** scales up spatial learning by introducing a massive mixture of image data, encompassing web-scraped images (LAION-400M, COYO-700M), interleaved documents (OBELICS), and curated high-quality datasets (Zero250M, ImageNet-21K, DataComp). In **Stage 3 (Short Video)**, we unlock preliminary temporal dynamics by incorporating short video clips (under 10 minutes) from WebVid and Koala. Crucially, we concurrently replay a substantial portion ($\sim$ 24.7M samples) of the static image mixture from Stage 2 to maintain spatial granularity. Finally, **Stage 4 (Long Video)** extends the temporal horizon by leveraging longer videos (over 10 minutes) alongside continuous image replays. This mixed-modality sampling guarantees FuxViT's proficiency in tracking extended motion dynamics without sacrificing its capability for intricate static spatial parsing.

C Frame Selective Attention (FSA) Details

Algorithm 1 Frame Selective Attention (FSA)

Require: Token sequence X, frame count T, spatial tokens per frame M, routing parameter K, head dimension d_h.
Ensure: Updated sequence Y.

Initialize: Partition sequence X into T frame-level blocks $[\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_T]$, where $\mathcal{B}_t = \{x_n \mid (t-1)M < n \leq tM\}$.
Compute Query Q, Key K, and Value V for all tokens in X.
for each candidate frame block $\mathcal{B}_s$ where $1 \leq s \leq T$ **do**
  Compute the mean key vector representing the frame's semantics:
  $\overline{k}_s = \frac{1}{|\mathcal{B}_s|} \sum_{j \in \mathcal{B}_s} k_j$
end for
for each frame block $\mathcal{B}_t$ where $1 \leq t \leq T$ **do**
  for each query token $x_n \in \mathcal{B}_t$ **do**
    for each past frame $s \in \{1, \ldots, t-1\}$ **do**
      Compute frame-level relevance: $a_{(n,s)} = \frac{q_n^\top \overline{k}_s}{\sqrt{d_h}}$
    end for
    Select Top-K relevant historical frames:
    $\mathcal{S}_n \leftarrow \text{Top-K}(\{a_{(n,1)}, \ldots, a_{(n,t-1)}\}, K)$
    for each key token $x_j \in \mathcal{B}_s$ where $1 \leq s \leq T$ **do**
      if $s = t$ **or** $s \in \mathcal{S}_n$ **then**
        $\mathcal{M}_{n,j} \leftarrow 0$ {Intra-frame (bidirectional) or selected inter-frame (causal)}
      else
        $\mathcal{M}_{n,j} \leftarrow -\infty$ {Unselected past frames or future frames}
      end if
    end for
  end for
end for
Apply sparse attention: $Y = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + \mathcal{M}\right) V$
return Y

To formalize the sparse temporal routing mechanism discussed in the main text, Algorithm 1 outlines the step-by-step operations of Frame Selective Attention (FSA). The process inherently respects the visual structural prior by partitioning the flattened token sequence into discrete frame-level blocks $(\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_T)$.

To compute historical relevance efficiently, we extract a consolidated semantic representation for each candidate frame by averaging its constituent key vectors ($\overline{k}_s$). For each query token x_n residing in the current frame t, we compute its relevance scores against all preceding frames. We then apply a Top-K selection operation to isolate the most critical historical contexts ($\mathcal{S}_n$).

The attention mask ($\mathcal{M}$) is subsequently constructed using a native block-causal paradigm: tokens interact bidirectionally with other tokens within their own frame (dense intra-frame modeling), while inter-frame interactions are strictly causal and confined only to the selected Top-K historical frames.

All unselected past frames and future frames are masked out ($\mathcal{M}_{n,j} \leftarrow -\infty$). Finally, the standard scaled dot-product attention is applied over this sparsely routed topology, gracefully slashing the quadratic computational overhead while preserving high-fidelity, long-range dependencies.

D Evaluation Prompts and Benchmarks

In this section, we provide the detailed response formatting prompts used for our evaluation across various image and video benchmarks. To ensure a fair, rigorous, and standardized comparison with existing multimodal large language models, the majority of our evaluation prompts are directly adopted from the LMMs-Eval framework [65]. Table 8 summarizes the specific text prompts utilized to guide the model's expected response format for each corresponding benchmark during our experimental evaluation.

Table 8: **Summary of the evaluation benchmarks.** Prompts are mostly borrowed from LMMs-Eval [65].

Benchmark	Response formatting prompts
TextVQA	Answer the question using a single word or phrase.
OCRBench	Answer the question using a single word or phrase.
ChartQA	Answer the question using a single word or phrase.
DocVQA	Answer the question using a single word or phrase.
InfoVQA	Answer the question using a single word or phrase.
MMBench	Answer with the option's letter from the given choices directly.
AI2D	Answer with the option's letter from the given choices directly.
RealWorldQA	Answer with the option's letter from the given choices directly.
MMStar	Answer with the option's letter from the given choices directly.
MLVU	–
MVBench	Only give the best option.
Video-MME	Answer with the option's letter from the given choices directly.
NEXT-QA	Answer with the option's letter from the given choices directly.
LongVideoBench	Answer with the option's letter from the given choices directly.
LVBench	Answer with the option's letter from the given choices directly.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. They effectively capture the core objectives, methodology, and key findings of the research, providing a clear overview of the paper's content.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the fixed routing budget and the limitations of our model-scaling evaluation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper has provided detailed descriptions on how to reproduce the main experimental results in Experiment Section and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code and Model will be publicly available soon.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specify all the training and test details in Experiment Section and Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: No error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provide sufficient information on the computer resources in Experiment Section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impacts in our paper. Positively, our efficient ViT design lowers the computational barrier for MLLM deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code or dataset we used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: No new asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: In this work, we propose a novel ViT for Multimodal Large Language Models (MLLMs).

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.