Spatial-TTT: Streaming Visual-based Spatial Intelligence with Test-Time Training

Fangfu Liu*,1, Diankun Wu*,1, Jiawei Chi*,1, Yimo Cai[1], Yi-Hsin Hung[1], Xumin Yu[2], Hao Li[3], Han Hu[2], Yongming Rao†,2, Yueqi Duan†,1

[1] Tsinghua University, Beijing, China
[2] Tencent Hunyuan, Beijing, China
[3] NTU, Singapore

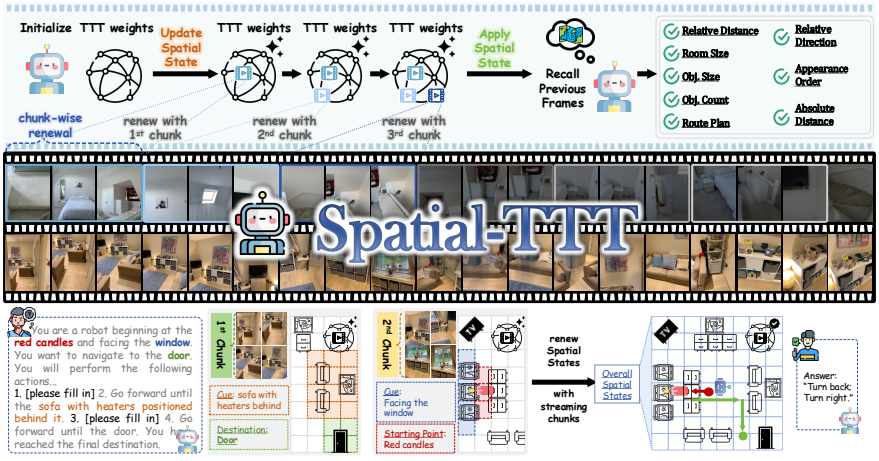

Fig. 1: Spatial-TTT. Given a Visual-based Spatial task, our proposed Spatial-TTT updates spatial state with streaming chunks then answers the question.

Abstract. Humans perceive and understand real-world spaces through a stream of visual observations. Therefore, the ability to streamingly maintain and update spatial evidence from potentially unbounded video streams is essential for spatial intelligence. The core challenge is not simply longer context windows but how spatial information is selected, organized, and retained over time. In this paper, we propose *Spatial-TTT* towards streaming visual-based spatial intelligence with test-time training (TTT), which adapts a subset of parameters (fast weights) to capture and organize spatial evidence over long-horizon scene videos. Specifically, we design a hybrid architecture and adopt large-chunk updates parallel with sliding-window attention for efficient spatial video processing. To further promote spatial awareness, we introduce a spatial-predictive mechanism applied to TTT layers with 3D spatiotemporal convolution, which encourages the model to capture geometric correspondence and temporal continuity across frames. Beyond architecture design, we construct a dataset with

* Equal contribution. † Corresponding author.

dense 3D spatial descriptions, which guides the model to update its fast weights to memorize and organize global 3D spatial signals in a structured manner. Extensive experiments demonstrate that Spatial-TTT improves long-horizon spatial understanding and achieves state-of-the-art performance on video spatial benchmarks. Project page: `https://liuff19.github.io/Spatial-TTT`.

Keywords: Spatial Intelligence · Test-Time Training · MLLM

1 Introduction

Learning to perceive, understand, and reason about 3D structure and geometric relationships of the physical world is a cornerstone capability for spatial intelligence [11, 19, 64], which finds broad applications in diverse fields including embodied robots [17, 26], autonomous driving [25, 57], and augmented reality devices [22]. In these real-world scenarios, spatial information is rarely captured from a single static viewpoint but rather emerges from a continuous stream of visual observations, where the camera moves, objects become occluded, and viewpoints change over time [35, 64]. This necessitates streaming spatial understanding—the capability to selectively maintain, progressively update, and reason over spatial memory from long-horizon video inputs.

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated impressive results in 2D visual understanding [4, 23, 27, 33, 70]. However, their performance degrades significantly on tasks requiring spatial understanding, primarily due to the inherent lack of 3D geometric priors [62, 66], as these models are predominantly trained on 2D semantic-level image-text pairs without the supervision of spatial structure. While recent efforts have explored spatial-aware MLLMs through augmenting the input representations with geometric cues [18, 24, 58] with more 3D spatial VQA data [11, 63], they remain confined to single images or short video clips (*i.e.,* 16 or 32 images) and cannot scale to the long-horizon video streams encountered in practical scenarios [64], where spatial cues are scattered across thousands of frames and must be progressively aggregated as the observer navigates through the environment. Naively extending the input sequence leads to prohibitive computational costs due to quadratic attention complexity [54, 69], while aggressive temporal subsampling inevitably discards fine-grained spatial details critical for accurate 3D reasoning [52, 64].

To address these challenges, we introduce *Spatial-TTT*, a novel framework for streaming visual-based spatial intelligence built on the Test-Time Training (TTT) paradigm [49]. Instead of using fixed parameters for inference, Spatial-TTT maintains adaptive fast weights that are updated online, acting as a compact non-linear memory to accumulate 3D evidence from unbounded video streams. To retain cross-modal alignment and semantic reasoning ability of pretrained MLLM, we employ a hybrid architecture that interleaves TTT layers with self-attention anchor layers, balancing efficient long-context compression with full-context reasoning. To make TTT practical for long spatial videos, we further adopt a large chunk update strategy [69] for higher parallelism and hardware efficiency,

and use sliding-window attention (SWA) in parallel to preserve intra-chunk spatiotemporal continuity.

While these designs enable TTT scalable to long-horizon videos and are compatible with pretrained MLLMs, the Q/K/V used for updates are produced by point-wise linear projections, which ignore neighborhood structure among visual tokens and make the memory update dynamics less spatially coherent. To this end, we propose a *spatial-predictive mechanism* that injects spatiotemporal inductive bias directly into the TTT branch. Instead of using point-wise linear projections, we apply lightweight depth-wise 3D spatiotemporal convolutions to aggregate local neighborhood context for visual tokens. This encourages the fast weights to learn predictive mappings between spatiotemporal contexts rather than isolated tokens, thereby better capturing geometric correspondence and temporal continuity and improving the stability and effectiveness of online updates.

Beyond architecture, effective TTT requires supervision that teaches the model how to update fast weights to preserve globally useful 3D evidence over long streams. However, existing spatial datasets are sparse and local, providing weak gradient signals for model to learn fast-weight update dynamics. To address this, we construct a dense scene-description dataset where the model is required to generate comprehensive 3D scene descriptions covering global context, objects and counts, and spatial relations. This dense description provides rich supervision for training fast-weight update dynamics to preserve structured, scene-level spatial information along video stream. Extensive experiments demonstrate that Spatial-TTT significantly improves long-horizon spatial understanding and achieves state-of-the-art performance on video spatial benchmarks. Our main contributions are summarized as follows:

- We propose Spatial-TTT for streaming visual-based spatial intelligence with test-time training, which performs online fast-weight updates as compact nonlinear memory to accumulate spatial evidence from long-horizon spatial video streams.
- We design a hybrid test-time training architecture together with large-chunk updates and parallel sliding-window attention, achieving efficient long spatial-context compression and reasoning.
- We introduce a spatial-predictive mechanism that captures geometric correspondence and spatial-temporal continuity, and further construct a dense scene-description dataset to provide rich supervision for learning effective fast-weight update dynamics.
- We conduct extensive experiments on video spatial benchmarks, which demonstrate that our method achieves state-of-the-art performance on a wide range of visual-based spatial understanding tasks.

2 Related Work

2.1 Visual-based Spatial Intelligence

Visual-based Spatial Intelligence of MLLMs focuses on MLLMs' ability to perceive, reason about, and recall spatial relationships as well as layouts from visual inputs.

While most existing MLLMs demonstrate strong performance on 2D perception and reasoning tasks [3, 40], they still struggle with tasks that require precise 3D spatial alignment, such as robotic manipulation [36] and 3D question answering [11, 24, 61]. To minimize this gap, previous researchers have constructed several benchmarks [34, 62, 64, 66], for example, VSI-Bench for evaluating comprehensive video-based visual-spatial intelligence [62], STI-Bench for examining spatial-temporal understanding [37], and VSI-Super for challenging spatial recall and continual counting [64]. Meanwhile, several methods were proposed to enhance visual-based spatial intelligence of MLLMs, like incorporating metric depth and multi-view inputs in MM-Spatial [15], adopting feed-forward visual geometry models in Spatial-MLLM [58] and VLM-3R [18], SFT and RL methods explored by SpaceR [43] and MindCube [66], and 3D feature alignment at output proposed by 3DThinker [12]. Recently, VST [63] constructed 4.1M SFT dataset for spatial perception and 135K RL dataset for spatial reasoning. SpatialLadder [32] built a 26K dataset, and Cambrian-S [64] proposed a four-stage training framework with VSI-590K dataset. However, existing methods mostly focused on pre-training or post-training stages while test-time strategy for natively adapting diverse and streaming data is still not fully explored.

2.2 Test-Time Training

Test-time training (TTT) refers to methods that improve model performance using only unlabeled test data at inference time [2, 8, 20, 50, 55]. Unlike test-time scaling (TTS), which relies on sampling multiple trajectories and selecting the most promising ones with frozen model parameters at test time [16, 41, 46], test-time training continually updates model parameters during inference to adapt to diverse inputs and tasks [38, 44]. Prior work has shown that adapting fast weights over large chunks at inference time enables efficient memorization, with applications spanning novel view synthesis, language modeling, and autoregressive video diffusion [69]. These results highlight the potential of TTT for long-context visual tasks, such as spatial intelligence in streaming video settings. More recent study on Nested Learning has demonstrated that TTT supports continal weight adaptation with multi-timescale and massive context windows, which benefits its application on streaming video understanding [7]. Work on TTT for few-shot learning [1] also showed that TTT yields substantial reasoning improvements beyond in-context learning. To effectively stabilize these fast-weight updates, researchers have introduced the Muon optimizer and momentum algorithms [6, 8], while also exploring a broader design space that includes various loss functions and neural representations of memory [30, 56]. While TTT has achieved notable success in language modeling, its application to enhancing the visual capabilities of MLLMs has received comparatively limited attention [45, 48].

3 Spatial-TTT

In this section, we first introduce the overall framework (Sec. 3.2) of Spatial-TTT shown in Fig. 2, which consists of a hybrid TTT architecture that interleaves

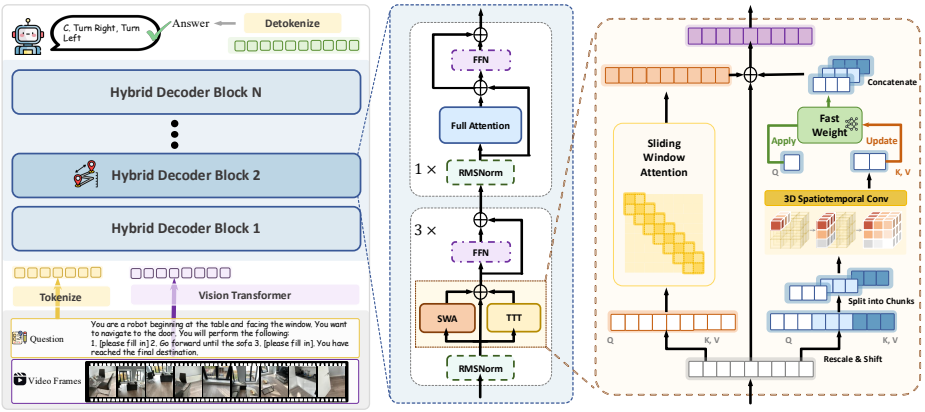

Fig. 2: Overview of Spatial-TTT. The model employs a hybrid architecture that interleaves TTT layers with self-attention anchor layers at a 3:1 ratio to preserve pretrained knowledge while enabling efficient long spatial-context compression. Within each TTT layer, sliding window attention (SWA) and TTT branch operate in parallel with shared Q/K/V projections, where the TTT branch applies spatial predictive mechanism with depthwise spatiotemporal convolution to capture geometric structure and temporal continuity.

TTT layers with standard self-attention layers to preserve pretrained visual-semantic knowledge, and a spatial-predictive mechanism that enhances TTT with lightweight depthwise convolutions to learn spatiotemporal structure. We then describe how to bridge sparse spatial supervision with dense scene descriptions that provide rich, scene-level gradient signals for learning effective fast-weight update dynamics (Sec. 3.3). Finally, we present a spatial-aware progressive training strategy (Sec. 3.4) towards effective visual-based spatial reasoning capability. Before introducing our method in detail, we first review the theory of Test-Time Training (TTT).

3.1 Preliminary

Test-Time Training. Traditional neural networks follow a static paradigm where model parameters θ remain frozen after training, limiting their ability to dynamically adapt to evolving inputs during inference. Test-Time Training (TTT) offers an alternative paradigm that enables on-the-fly parameter adaptation by updating a designated subset of parameters (*i.e.,* fast weights) while processing each test sequence. The TTT mechanism maintains fast weights $W \in \mathbb{R}^{d_{out} \times d_{in}}$ that parameterize a small neural network $f_W : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$. Unlike conventional model parameters that are fixed post-training, these fast weights function as an adaptive memory that continuously encodes contextual information from the input stream. Given an input sequence $\mathbf{x} = [x_1, x_2, \ldots, x_T]$ where $x_t \in \mathbb{R}^d$, each token is projected to derive a key k_t, a query q_t, and a value v_t. TTT alternates between two operations at each timestep:

1. **Update Operation:** The fast weights are modified to associate the current key-value pair (k_t, v_t) by performing a gradient descent step on a self-supervised loss function $\mathcal{L}$:

$$W_t \leftarrow W_{t-1} - \eta \nabla_W \mathcal{L}\big(f_{W_{t-1}}(k_t), v_t\big). \tag{1}$$

   where η denotes the learning rate. This step encodes information from the pair (k_t, v_t) into the neural memory W_t.
2. **Apply Operation:** The updated network produces the output by processing the query:

$$o_t = f_{W_t}(q_t). \tag{2}$$

   The output o_t is enriched with contextual information from all preceding tokens, as their key-value associations have been progressively encoded into W_t.

3.2 Overall Framework

Hybrid TTT Architecture. The inherent linear complexity of Test-Time Training makes it well-suited for processing unbounded spatial streams from long-horizon videos. However, directly replacing all core attention layers with TTT layers is prone to disrupting the pretrained cross-modal alignment and visual semantics, while recovering these would require heavily retraining [49, 50]. To address this, we employ a *hybrid TTT architecture* that interleaves TTT layers with self-attention layers at a 3:1 ratio. In a transformer with L decoder layers, 75% use TTT while the remaining 25% retain standard self-attention as anchor layers. The anchor layers maintain full attention access over the entire context, preserving the pretrained model's semantic reasoning ability. Meanwhile, TTT layers compress long-range temporal dependencies into adaptive fast weights W_t, achieving sublinear memory growth.

Within each TTT layer, conventional implementations use small chunks (*e.g.*, 16 or 64 tokens) for frequent fast-weight updates [51]. This works poorly for streaming visual observations: small chunks lead to low GPU utilization due to poor parallelism, and chunk boundaries artificially break the spatial structure of video frames. Inspired by LaCT [69], we instead adopt a large chunk size for visual tokens, roughly aligned with multiple video frames, to substantially improve parallelism and hardware efficiency and to keep spatially coherent visual content inside the same update unit. A key challenge with large chunks arises from the causal constraints in TTT updates: to prevent information leakage, the current chunk cannot interact with itself during fast-weight updates, which would otherwise allow earlier tokens to access later ones. This restriction precludes any intra-chunk token interactions, which is undesirable for maintaining spatiotemporal continuity in visual data. To address this, we incorporate sliding window attention (SWA) within each TTT layer, operating in parallel with TTT, sharing query, key, and value projections (with lightweight learnable scale and shift for TTT's queries and keys). The window size w is set to be no smaller than the chunk size b, so that the SWA fully covers the causal lower-triangular

attention matrix within each chunk, ensuring the completeness of the causal structure. The layer output combines both branches:

$$o_t = \text{WindowAttn}(q_t, K_{[t-w:t]}, V_{[t-w:t]}) + f_{W_t}(q_t), \tag{3}$$

where $K_{[t-w:t]}$ and $V_{[t-w:t]}$ denote the keys and values within the sliding window. For the fast-weight network f_W, we use a bias-free SwiGLU-MLP to increase the nonlinearity and expressiveness of the memory:

$$f_W(\mathbf{x}) = W_2\big[\text{SiLU}(W_1\mathbf{x}) \odot (W_3\mathbf{x})\big], \tag{4}$$

where $\mathbf{x}$ denotes the input (k_t for update, q_t for apply), $W = \{W_1, W_2, W_3\}$ are the learnable fast weights, and $\odot$ is element-wise multiplication.

Spatial-Predictive Mechanism. Streaming spatial understanding poses unique challenges: spatial information emerges from continuous visual observations, where the camera moves, viewpoints change, and objects are gradually revealed or occluded. Adjacent visual tokens describe a progressively unfolding 3D scene with strong geometric and temporal continuity. However, in conventional TTT designs [50, 69], Q, K, V are generated through point-wise linear projections: $k_t = W_k x_t$, $v_t = W_v x_t$, which ignore spatial-temporal structure among visual tokens. Since fast weights must compress the chunk into compact memory to represent 3D structure, using isolated token transforms for K, V without local geometric context forces them to learn spatial patterns from scratch—making compression harder and memory less structured. To address this, we introduce a spatial-predictive mechanism with lightweight depth-wise spatiotemporal convolution on the Q, K, V of the TTT branch. For visual tokens from videos, we reshape them into a spatiotemporal grid and aggregate neighborhood information through local aggregation. Formally, for the i-th channel of a token at spatiotemporal position (t, h, w), the conv-enhanced query, key and value are:

$$\tilde{x}^i_{t,h,w} = \sum_{\delta \in \mathcal{N}} \theta^i_\delta \cdot x^i_{t+\delta_t, h+\delta_h, w+\delta_w}, \quad x \in \{q, k, v\}, \tag{5}$$

where $\mathcal{N} = \{-\lfloor \kappa/2 \rfloor, ..., \lfloor \kappa/2 \rfloor\}^3$ is the local neighborhood with kernel size κ, and $\theta \in \mathbb{R}^{\kappa^3 \times d}$ are learnable kernel weights initialized with Dirac delta to preserve identity mapping at initialization. To further improve the stability and effectiveness of the TTT update, we adopt the Muon update rule [29, 69] instead of vanilla implementation (*i.e.,* Equation 1) for optimization:

$$G_t = \text{MuonUpdate}(G_{t-1}, \nabla_W \mathcal{L}(f_{W_{t-1}}(\tilde{k}_t), \tilde{v}_t)), \tag{6}$$

$$W_t \leftarrow \text{L2Norm}(W_{t-1} - \eta G_t), \tag{7}$$

where G_t is the orthogonalized gradient with momentum, $\text{Muon}(\cdot, \cdot)$ accumulates momentum and orthogonalizes via Newton-Schulz iterations, and L2Norm normalizes weights while preserving their original magnitude. Powered by the spatial-predictive mechanism, fast weights no longer learn prediction between isolated tokens, but rather predictive mapping between spatial-temporal contexts. This enables fast weights to implicitly capture geometric correspondence and temporal continuity in streaming spatial understanding.

3.3 Bridging Sparse Spatial QA with Dense Supervision

The effectiveness of TTT [49] depends on whether the model learns fast-weight update dynamics that retain information useful for future timesteps. However, supervision in existing spatial intelligence datasets [18, 58, 64] is typically sparse and local. For instance, a typical spatial QA task queries relations between two objects in a small region of the scene (often answerable from only a few frames), and the target answers are often short (*e.g.,* a multiple-choice option or an integer). Such target answers only supervise a small subset of the underlying 3D scene states, providing weak and low-coverage gradient signals for learning fast-weight update dynamics. Consequently, the model has limited incentive to construct a coherent and persistent global 3D memory over long video streams.

To bridge this gap, we construct a dense scene-description dataset from SceneVerse annotations [28]. For each training example, the model receives a spatial video stream and is required to generate a comprehensive description of the underlying 3D scene. Unlike short-form answers, the target output is formatted as a coherent scene walkthrough, which is derived from object-centric 3D scene graphs [28]. Specifically, the target description include following aspects: (1) *Global context:* identifying the scene type and functional setting, encouraging fast weights to encode global semantic descriptors beyond local visual cues; (2) *Objects and counts:* enumerating object categories and precise counts, which encourages retention of persistent instance-level evidence across time; and (3) *Object relations:* describing spatial layouts and pairwise relations, which promotes encoding of geometric structure and inter-object constraints. This dense scene description task provides rich and high-coverage supervision that complements sparse spatial QA. By training the model to generate detailed 3D scene descriptions, we encourage the fast-weight dynamics to capture global and persistent spatial representations that benefit downstream spatial reasoning.

3.4 Spatial-Aware Progressive Training

With the architecture and data designed above, the remaining question is how to train the model effectively. We meticulously design a two-stage spatial-aware progressive training strategy: (1) initializing fast weights with global 3D awareness using dense scene description data, and (2) tuning streaming spatial reasoning capability with large-scale spatial VQA data. Specifically, we first train the hybrid TTT architecture on the dense scene description dataset described above, with the goal of teaching fast weights to retain comprehensive scene-level information through chunk-by-chunk memory updates. To better preserve the pretrained visual knowledge during this process, we do not directly set the sliding window size equal to the chunk size. Instead, we employ a sliding window annealing strategy: the sliding window size w is linearly annealed from an initial value $w_{\max}$ to $w_{\min} = b$ (the chunk size) over the first stage. As the window size gradually decreases, TTT layers are forced to take over more responsibility for cross-chunk information propagation while fast weights progressively learn to encode global 3D scene structure during memory updates. However, spatial understanding

requires not only “remembering" spatial information but also effectively “recalling and reasoning" about spatial relations during streaming observations. To further promote streaming visual-based spatial intelligence, we fine-tune the model in the second stage using 2M spatial VQA samples covering diverse tasks such as object relative direction/distance estimation, spatial counting, route planning, and room size estimation. In this stage, the window size and chunk size are fixed at the same value ($w = b$), so TTT layers are fully focused on cross-chunk spatial information aggregation. Through this fine-tuning, the model learns to selectively retain task-relevant spatial evidence via fast weight updates and retrieve accumulated spatial knowledge when reasoning.

At inference time, we employ a dual KV cache mechanism for constant-memory streaming. The first is a sliding window KV cache of fixed length w for local context modeling in sliding window attention; when the cache exceeds the window size, the earliest entries are discarded. The second is a TTT pending KV cache that accumulates key-value pairs for fast weight updates: it starts empty and grows as new tokens arrive; whenever its length reaches the chunk size b, these KV pairs are used to perform one fast weight update, after which the pending cache is cleared.

4 Experiments

4.1 Experiment Setup

Implementation Details. Our model is initialized from Qwen3-VL-2B-Instruct [3]. In the hybrid TTT architecture design, we apply TTT layers to every three of every four attention layers. To preserve pretrained knowledge and facilitate convergence, we do not initialize new QKV projection matrices from scratch and make the TTT layers share the original attention layer's QKV projection matrix. To increase expressivity of TTT layers, we introduce lightweight learnable scale-and-shift parameters applied to the projected q and k. At the start of training, the gating projection is initialized to zero, the scale parameters are set to ones, and the shift parameters are set to zeros. In the spatial-predictive mechanism, we apply depthwise 3D convolutions with kernel size $3 \times 3 \times 3$ on the Q, K, V projections of TTT layers, which is Dirac-initialized, ensuring that the network initially behaves identically to the original full-attention model. The LaCT chunk size b is set to 2648, and the window size w is initialized to 5600—sufficient to cover the entire sequence—and annealed to 2648 over the first two epochs on the dense scene-description dataset, where 32 frames are uniformly sampled for training. We then perform continuous finetuning with large-scale spatial VQA data from 64 to 128 frames. We use a learning rate of 1e-6 for the pretrained backbone and 1e-5 for the newly introduced TTT-related parameters, with a warmup of 1k steps and a cosine learning-rate scheduler across all training stages.
Datasets. In the first stage, we train our model on the dense scene-description dataset (Sec. 3.3). This dataset contains approximately 16k samples, including 3.6k descriptions of ScanNet [14] scenes and 12.5k descriptions of ARKitScenes [5] scenes. In the second stage, we train the model on a 3M large-scale spatial

Table 1: Evaluation Results on VSI-Bench [62]. For Numerical Questions, we report MRA score. For Multiple-Choice Questions, we report ACC score, Avg. is the macro average across all tasks, following the original paper. For human, we directly use the reported results in VSI-Bench, which is on a subset of VSI-Bench with 400 samples. We use and to denote the best and second-best results within proprietary models and open-source models, respectively.

Models	Numerical Question				Multiple-Choice Question				Avg.
	Obj. Count	Abs. Dist	Obj. Size	Room Size	Rel. Dis	Rel. Dir	Route Plan	Appr. Order	
Human	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0	79.2
Random Choice	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2	34.0
Proprietary Models									
Seed-2.0 [9]	49.4	25.3	69.5	25.8	61.8	44.9	44.3	71.0	50.7
Grok-4 [60]	37.1	32.9	60.8	45.4	53.1	39.6	47.4	66.8	47.9
Gemini-2.5-pro [13]	46.0	37.3	68.7	54.3	61.9	43.9	47.4	68.7	53.5
Gemini-3-pro [21]	49.0	**42.8**	71.5	41.8	56.6	**57.5**	**61.9**	60.0	**56.0**
Kimi-K2.5 [53]	**57.2**	34.9	69.3	**54.4**	59.6	41.3	52.1	67.0	53.6
GPT-5 [42]	53.3	34.4	**73.3**	47.5	**63.7**	48.6	50.2	**68.9**	55.0
Open-source General Models									
LongVA-7B [68]	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7	29.2
LLaVA-OneVision-72B [31]	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2
LLaVA-Video-72B [39]	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
InternVL3-2B [70]	64.8	30.8	32.4	22.9	32.2	34.9	32.9	12.6	32.9
InternVL3-8B [70]	66.0	34.8	43.6	47.5	48.0	39.3	26.2	31.3	42.1
Qwen2.5-VL-3B-Instruct [4]	24.3	24.7	31.7	22.6	38.3	42.6	26.3	21.2	29.0
Qwen2.5-VL-7B-Instruct [4]	40.9	14.8	43.4	10.7	38.6	40.1	33.0	29.8	31.4
Qwen3-VL-2B-Instruct [3]	62.1	40.2	71.4	49.7	52.2	42.0	30.4	54.5	50.3
Qwen3-VL-8B-Instruct [3]	67.5	47.0	**76.3**	61.9	58.0	50.9	35.0	66.3	57.9
Open-source Spatial Intelligence Models									
MindCube-3B [66]	12.8	22.7	4.3	23.4	20.2	15.7	15.9	22.4	17.2
SpatialLadder-3B [32]	62.1	35.3	61.9	41.4	45.6	46.4	27.3	38.5	44.8
SpaceR-7B [43]	44.5	24.7	53.5	37.3	41.9	46.1	29.3	54.8	41.5
ViLaSR-7B [59]	58.1	33.8	61.4	28.8	45.0	46.5	29.9	53.2	44.6
VST-3B-SFT [63]	69.3	45.4	71.8	62.4	59.0	46.0	38.7	70.2	57.9
VST-7B-SFT [63]	**72.0**	44.4	74.3	**68.3**	59.7	55.8	44.9	65.2	60.6
Cambrian-S-3B [64]	70.7	40.6	68.0	46.3	**64.8**	61.9	27.3	**78.8**	57.3
Spatial-MLLM-4B [58]	65.3	34.8	63.1	45.1	41.3	46.9	33.5	46.3	47.0
Ours									
Spatial-TTT-2B	70.8	**47.8**	71.7	65.9	61.8	**73.0**	**47.4**	77.0	**64.4**

question-answering dataset, which includes open-sourced spatial data and self-constructed data. Detailed statistics and curation procedures are provided in the Supplementary Materials.

Baselines. We compare our model against a broad set of baselines, covering proprietary multimodal models (including GPT-5 [42], Gemini-3-Pro [21], Seed-2.0 [9], Kimi-K2.5 [53] etc.), open-source general-purpose MLLMs (LLaVA-OneVision [31], LLaVA-Video [39], InternVL3 series [70], Qwen2.5-VL series [4], Qwen3-VL series [3], etc.), and open-source spatial-intelligence models (VST [63], Cambrian-S [64], Spatial-MLLM [58], etc.). For long-horizon recall and counting on VSI-SUPER [64], we additionally include long-video understanding baselines (MovieChat [47] and Flash-VStream [67]). Additionally, we report Human and Random Choice scores for VSI-Bench [62] and MindCube [66] as references.

4.2 General Spatial Understanding Results

Results on VSI-Bench. To assess the models' general spatial understanding capabilities, we evaluate our model and the baseline models on VSI-Bench [62], a benchmark that measures multimodal visual-spatial intelligence through egocentric video understanding. VSI-Bench contains over 5,000 question–answer pairs constructed from 288 real-world indoor videos drawn from the validation splits of ScanNet [14], ScanNet++ [65], and ARKitScenes [5], covering diverse environments (*e.g.,* homes, offices, labs, and factories) across multiple geographic regions.

Following the original paper [62], we use Accuracy (ACC) for multiple-choice questions and Mean Relative Accuracy (MRA) for numerical questions to quantify how closely predictions align with ground-truth values. Avg. is the macro average across these metrics. The results are reported in Table 1.

As shown, our Spatial-TTT-2B achieves the best overall performance on VSI-Bench with an Avg. of 64.4, surpassing both proprietary and open-source baselines despite its compact 2B scale. In particular, our model exhibits strong advantages on multiple-choice spatial reasoning tasks that require consistent geometric understanding across egocentric views, including Relative Direction and Route Plan, indicating improved capability in direction reasoning and navigation planning. Meanwhile, on numerical questions, our model attains the best score on Absolute Distance and remains highly competitive on Room Size and Object Count, suggesting stronger metric grounding and scene-scale estimation than other baselines.

Table 2: Evaluation results on MindCube-Tiny [66]. Avg. is micro average across all tasks, following the original paper. We use and to denote the best and second-best results within proprietary models and open-source models, respectively.

Models	Avg.	Rotation	Among	Around
Human	94.5	-	-	-
Random Choice	33.0	33.3	31.8	35.7
Proprietary Models				
Seed-2.0 [9]	54.8	89.0	45.2	52.0
Grok-4 [60]	63.5	93.0	54.4	61.6
Gemini-2.5-pro [13]	57.6	88.0	44.9	63.2
Gemini-3-pro [21]	**63.9**	73.5	**59.3**	66.0
Kimi-K2.5 [53]	57.7	82.5	50.2	56.5
GPT-5 [42]	56.3	**94.5**	38.2	**68.4**
Open-source General Models				
InternVL3-2B [70]	37.5	28.9	36.9	45.6
InternVL3-8B [70]	41.5	36.5	38.1	53.6
Qwen2.5-VL-3B [4]	37.6	33.5	35.9	44.8
Qwen2.5-VL-7B [4]	36.0	37.0	32.3	44.0
Qwen3-VL-2B [3]	34.5	32.5	31.6	42.8
Qwen3-VL-8B [3]	29.4	29.5	28.6	31.2
Open-source Spatial Intelligence Models				
MindCube-3B [66]	51.7	34.0	51.0	67.6
SpatialLadder-3B [32]	43.4	35.0	43.2	50.8
SpaceR-7B [43]	37.9	35.0	34.2	49.2
ViLaSR-7B [59]	35.0	35.5	31.0	44.4
VST-3B-SFT [63]	35.9	32.0	34.9	41.6
VST-7B-SFT [63]	39.7	37.0	35.9	50.8
Cambrian-S-3B [64]	32.5	27.0	33.2	35.2
Spatial-MLLM-4B [58]	33.4	39.0	30.5	36.0
Ours				
Spatial-TTT-2B	**76.2**	**55.5**	**74.0**	**89.8**

Results on MindCube Benchmark. We further evaluate our model and the baseline models on MindCube [66], a benchmark that pairs multi-view image groups with spatial reasoning questions to assess fine-grained spatial capabilities. MindCube specifically tests (1) cross-view object consistency and (2) reasoning about occluded or unseen elements under changing camera viewpoints. Following common practice [10], we use MindCube-Tiny as evaluation set, which is a diagnostic subset sampled from MindCube. Mindcube-tiny contains 1,050 questions in total (600 AMONG, 250 AROUND, and 200 ROTATION). We report accuracy (ACC) for all question types.

The results are summarized in Table 2. As shown, Spatial-TTT achieves 76.2 ACC score on MindCube-Tiny, outperforming all baseline models. By question type, it attains the best performance in ACROSS and AMONG subsets. Relative to the strongest proprietary baseline (Gemini-3-pro; 63.9% average) and the strongest open-source spatial model (MindCube-3B; 51.7% average), Spatial-TTT improves by 12.3 and 24.5 percentage points, respectively. These results indicate that Spatial-TTT provides stronger spatial reasoning under viewpoint changes and occlusions.

4.3 Streaming Spatial Sensing Results

To assess models' continual spatial sensing under long-horizon, streaming-style video, we benchmark our model and baselines on VSI-SUPER-Recall (VSR) and VSI-SUPER-Count (VSC) [64]. Both benchmarks target spatial supersensing in arbitrarily long egocentric videos by requiring models to continuously accumulate evidence as new objects appear under changing viewpoints and to maintain long-term spatiotemporal memory for subsequent queries. Specifically, VSC requires models to count objects across extended sequences, whereas VSR uses a multiple-choice setting to test whether models can recall the temporal order of inserted objects over even longer videos. Following the benchmark protocol, we report mean recall accuracy (MRA) for VSC and accuracy (ACC) for VSR. As shown in Table 3, Spatial-TTT achieves competitive performance on VSI-SUPER-Recall across different video lengths, while significantly outperforming all baselines on VSI-SUPER-Count. Long-video understanding models (*e.g.*, MovieChat [47] and Flash-VStream [67]) consistently underperform across all video durations, suggesting that strong long-context modeling alone is insufficient for continual spatial sensing without robust spatial reasoning capabilities. General purpose and spatial intelligence models (*e.g.*, Qwen3-VL [3] and Cambrian-S [64]) achieve relatively competitive results on shorter videos, but their performance collapses on longer sequences. This degradation is largely attributable to practical limitations in processing extended inputs, such as exceeding context budgets or running into out-of-memory, which prevents them from continuously integrating new observations and retaining long-term spatiotemporal evidence. In contrast, Spatial-TTT maintains stable performance as video length increases by performing online updates that incrementally integrate new observations, enabling continual accumulation and retention of long-term spatiotemporal evidence.

Table 3: Results on VSI-SUPER-Recall (VSR) and VSI-SUPER-Count (VSC) [64]. We report the average score on the 10-120 minute subset of VSR and the 10–120 minute subset of VSC. For all subsets, videos are sampled at 1 fps. Qwen3-VL-2B and Cambrian-S-7B run out of memory (OOM) on an 80GB-GPU for videos longer than 120 minutes; therefore, their scores are reported as 0.

Models	VSI-SUPER-Recall				VSI-SUPER-Count			
	10min	30min	60min	120min	10min	30min	60min	120min
General Models and Spatial Intelligence Models								
Qwen3-VL-2B [3]	35.0	30.0	**28.3**	0.0	0.8	0.0	0.0	0.0
Cambrian-S-7B [64]	**38.3**	**35.0**	6.0	0.0	0.6	0.0	0.0	0.0
Long-video Understanding Models								
MovieChat [47]	18.3	21.7	16.7	26.7	0.0	0.0	0.0	0.0
Flash-VStream [67]	28.3	33.3	23.3	28.3	0.0	0.0	0.0	0.0
Ours								
Spatial-TTT-2B	**38.3**	**35.0**	**28.3**	**30.0**	**31.8**	**45.6**	**36.2**	**38.4**

4.4 Ablation Study and Analysis

In Table 4, we report ablation results of Spatial-TTT on VSI-Bench [62]. As shown, the full model achieves the best overall performance (64.4 Avg.), outperforming all ablated variants on both the Numerical and Multiple-Choice subsets, validating the effectiveness of our proposed components. Following, we provide a component-wise analysis.

Spatial-predictive mechanism. As shown, replacing the depth-wise 3D spatiotemporal convolution with identity projections (*w/o SP-Mechanism*) drops avg. from 64.4 to 62.1, with a larger decline on numerical questions (from 64.0 to 60.7). This supports that injecting spatiotemporal inductive bias stabilizes fast-weight updates and benefits metric-level reasoning.

Table 4: Ablations of Spatial-TTT on VSI-Bench [62]. "w/o SP-Mechanism " denotes replacing spatial-predictive conv with identity projections. "w/o Dense Data" denotes training without dense scene-description data. "w/o Hybrid Arch" denotes pure TTT architecture.

Setting	Numerical	Multiple-Choice	Avg.
Spatial-TTT	**64.0**	**64.8**	**64.4**
w/o SP-Mechanism	60.7	63.4	62.1
w/o Dense Data	61.0	61.5	61.3
w/o Hybrid Arch	55.4	52.4	53.9

Dense scene-description supervision. As shown, removing dense scene description data (*w/o Dense Data*) reduces the overall Avg. from 64.4 to 61.3. On the Numerical and Multiple-Choice subsets, the scores drop by 3.0 and 3.3 points, respectively. These results indicate that dense, scene-level supervision provides stronger training signals for learning effective online update dynamics and for retaining globally useful 3D evidence over long video streams.

Table 5: Peak memory usage (GB) and TFLOPs per forward pass for varying input lengths. The results are measured at 352×480. For fair comparison, we select representative models with comparable base model sizes: Qwen3-VL-2B [3] as the general-purpose MLLM baseline, and Spatial-MLLM-4B [58] as the geometry-augmented spatial VLM baseline. Spatial-MLLM-4B ran out of GPU memory at 512 and 1024 frames on our test hardware (reported as OOM), TFLOPs are therefore not reported.

Models	128 frames		256 frames		512 frames		1024 frames	
	MEM	TFLOPs	MEM	TFLOPs	MEM	TFLOPs	MEM	TFLOPs
Qwen3-VL-2B [3]	6.2	75.9	8.3	179.9	12.6	473.9	21.2	1403.1
Spatial-MLLM-4B [58]	25.9	1698.8	41.8	6002.1	OOM	N/A	OOM	N/A
Spatial-TTT-2B (Ours)	6.2	74.3	7.0	156.2	8.4	341.9	11.9	799.4

Hybrid TTT architecture. Eliminating self-attention anchor layers (*w/o Hybrid Arch*) causes the largest degradation on average score (from 64.4 to 53.9), especially on Multiple-Choice (from 64.8 to 52.4). These results highlights the necessity of hybrid interleaving to retain cross-modal alignment and global-context reasoning while scaling to long-horizon videos.

Analysis for memory usage and theoretical TFLOPs. We further analyze decoding memory usage and theoretical TFLOPs across varying input lengths. As shown in Table 5, Spatial-MLLM [58], which introduces an explicit geometry encoder, exhibits prohibitive scaling—its computational cost grows super-linearly with frame count, and it runs out of memory (OOM) beyond 256 frames, making it impractical for streaming video scenarios. For the general-purpose baseline Qwen3-VL-2B [3], while starting with comparable resource usage at short contexts, both its memory and TFLOPs scale with the quadratic complexity inherent to standard Transformer attention. In contrast, our Spatial-TTT benefits from linear-complexity attention, where doubling the input length results in approximately doubled computation—closely tracking theoretical linear scaling. At 1024 frames, Spatial-TTT achieves over 40% reduction in both TFLOPs and memory compared to Qwen3-VL-2B. Crucially, this efficiency gap widens as context length grows, making our approach particularly well-suited for streaming spatial intelligence.

5 Conclusion

In this paper, we present Spatial-TTT, a novel framework for streaming visual-based spatial intelligence that leverages test-time training to maintain adaptive fast weights as compact memory for accumulating 3D evidence from long-horizon video streams. We employ a hybrid architecture that interleaves TTT layers with self-attention anchor layers to preserve pretrained knowledge while enabling efficient long-context processing, further enhanced by large-chunk updates and sliding-window attention for practical scalability. Beyond efficient architecture design, we further promote spatial awareness by introducing a spatial-predictive mechanism with 3D spatiotemporal convolutions to inject local neighborhood

inductive bias, and construct a dense scene-description dataset to provide rich supervision for learning effective fast-weight update dynamics. Extensive experiments demonstrate that Spatial-TTT achieves state-of-the-art performance across spatial benchmarks. We hope that Spatial-TTT provides a promising direction for building MLLMs with persistent spatial memory, enabling more robust and scalable spatial intelligence in real-world applications.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62576185, the Beijing Natural Science Foundation under Grant L252011, and the 2025 Tencent Rhino-Bird Focused Research Program.

References

1. Akyürek, E., Damani, M., Zweiger, A., Qiu, L., Guo, H., Pari, J., Kim, Y., Andreas, J.: The surprising effectiveness of test-time training for few-shot learning. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 942–963. PMLR (2025)
2. Ba, J., Hinton, G.E., Mnih, V., et al.: Using fast weights to attend to the recent past. In: NeurIPS (2016)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. arXiv preprint arXiv:2111.08897 (2021)
6. Behrouz, A., Li, Z., Kacham, P., Daliri, M., Deng, Y., Zhong, P., Razaviyayn, M., Mirrokni, V.: Atlas: Learning to optimally memorize the context at test time (2025), https://arxiv.org/abs/2505.23735
7. Behrouz, A., Razaviyayn, M., Zhong, P., Mirrokni, V.: Nested learning: The illusion of deep learning architectures. arXiv preprint arXiv:2512.24695 (2025)
8. Behrouz, A., Zhong, P., Mirrokni, V.: Titans: Learning to memorize at test time. arXiv preprint arXiv:2501.00663 (2024)
9. Bytedance Seed: Seed2.0 model card: Towards intelligence frontier for real-world complexity. Tech. rep., Bytedance (2025), https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2/0214/Seed2.0%20Model%20Card.pdf/, accessed: 29 June 2026
10. Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., Zhou, T., Li, J., Pang, H.E., Qian, O., Wei, Y., Lin, Z., Shi, X., Deng, K., Han, X., Chen, Z., Fan, X., Deng, H., Lu, L., Pan, L., Li, B., Liu, Z., Wang, Q.,

Lin, D., Yang, L.: Scaling spatial intelligence with multimodal foundation models (2025), https://arxiv.org/abs/2511.13719
11. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Driess, D., Florence, P., Sadigh, D., Guibas, L., Fei-Fei, L.: Spatial-vlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
12. Chen, Z., Zhang, M., Yu, X., Luo, X., Sun, M., Pan, Z., Feng, Y., Pei, P., Cai, X., Huang, R.: Think with 3d: Geometric imagination grounded spatial reasoning from limited views. arXiv preprint arXiv:2510.18632 (2025)
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
14. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
15. Daxberger, E., Wenzel, N., Griffiths, D., Gang, H., Lazarow, J., Kohavi, G., Kang, K., Eichner, M., Yang, Y., Dehghan, A., Grasch, P.: Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7395–7408 (October 2025)
16. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Nature **645**, 633–638 (2025). https://doi.org/10.1038/s41586-025-09422-z
17. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., et al.: Palm-e: An embodied multimodal language model. In: International Conference on Machine Learning (ICML) (2023)
18. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279 (2025)
19. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
20. Gandelsman, Y., Sun, Y., Chen, X., Efros, A.A.: Test-time training with masked autoencoders. In: NeurIPS (2022)
21. Google: A new era of intelligence with gemini 3, https://blog.google/products-and-platforms/products/gemini/gemini-3, accessed: 29 June 2026
22. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
23. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
24. Hong, Y., Zhen, H., et al.: 3D-LLM: Injecting the 3d world into large language models. In: Proceedings of the Neural Information Processing Systems (NeurIPS) (2023)
25. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

26. Huang, W., Wang, C., Li, Y., Zhang, R., Fei-Fei, L.: Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. arXiv preprint arXiv:2409.01652 (2024)
27. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
28. Jia, B., Chen, Y., Yu, H., Wang, Y., Niu, X., Liu, T., Li, Q., Huang, S.: Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In: European Conference on Computer Vision. pp. 289–310. Springer (2024)
29. Jordan, K., Jin, Y., Boza, V., You, J., Cesista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for hidden layers in neural networks (2024), https://kellerjordan.github.io/posts/muon/, accessed: 29 June 2026
30. Karami, M., Mirrokni, V.: Lattice: Learning to efficiently compress the memory. arXiv preprint arXiv:2504.05646 (2025)
31. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
32. Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: Spatialladder: Progressive training for spatial reasoning in vision-language models. arXiv preprint arXiv:2510.08531 (2025)
33. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
34. Li, K., Wang, Y., He, Y., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR (2024)
35. Li, M., Wang, Y.X., Ramanan, D.: Towards streaming perception. In: European Conference on Computer Vision (ECCV) (2020)
36. Li, X., Zhang, M., et al.: ManipLLM: Embodied multimodal large language model for object-centric robotic manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
37. Li, Y., Zhang, Y., Lin, T., Liu, X., Cai, W., Liu, Z., Zhao, B.: Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 25332–25342 (October 2025)
38. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (Jul 2024). https://doi.org/10.1007/s11263-024-02181-w
39. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: arXiv preprint arXiv:2311.10122 (2023)
40. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of the Neural Information Processing Systems (NeurIPS) (2023)
41. OpenAI: Openai o1 system card. Tech. rep., OpenAI (2024), https://openai.com/index/openai-o1-system-card/, accessed: 29 June 2026
42. OpenAI: GPT-5 System Card. Technical report, OpenAI (Aug 2025), accessed: 2025-08-10
43. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)

44. Schlag, I., Irie, K., Schmidhuber, J.: Linear transformers are secretly fast weight programmers. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 9355–9366. PMLR (18–24 Jul 2021)
45. Shu, M., Nie, W., Huang, D.A., et al.: Test-time prompt tuning for zero-shot generalization in vision-language models. In: NeurIPS (2022)
46. Snell, C., Jaeckle, L., Li, R., Kumar, A., Levine, S.: Scaling llm test-time compute optimally can be more effective than scaling parameters. In: Proceedings of the 41st International Conference on Machine Learning (ICML) (2024)
47. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., 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. pp. 18221–18232 (2024)
48. Sun, Y., Dong, X., Menon, S., Schwab, D., Kolter, J.Z.: Learning to (learn at test time): Rnns with expressive hidden states. In: arXiv preprint arXiv:2407.04620 (2024)
49. Sun, Y., Li, X., Dalal, K., Xu, J., Vikram, A., Zhang, G., Dubois, Y., Chen, X., Wang, X., Koyejo, S., Hashimoto, T., Guestrin, C.: Learning to (learn at test time): RNNs with expressive hidden states. In: Proceedings of the 41st International Conference on Machine Learning (ICML) (2024), https://arxiv.org/abs/2407.04620
50. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning. pp. 9229–9248. PMLR (2020)
51. Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J., Wei, F.: Retentive network: A successor to transformer for large language models. arXiv preprint arXiv:2307.08621 (2023)
52. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
53. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S.H., Cao, Y., Charles, Y., Che, H.S., Chen, C., Chen, G., Chen, H., Chen, J., Chen, J., Chen, J., Chen, J., Chen, K., Chen, L., Chen, R., Chen, X., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Y., Chen, Z., Chen, Z., Cheng, D., Chu, M., Cui, J., Deng, J., Diao, M., Ding, H., Dong, M., Dong, M., Dong, Y., Dong, Y., Du, A., Du, C., Du, D., Du, L., Du, Y., Fan, Y., Fang, S., Feng, Q., Feng, Y., Fu, G., Fu, K., Gao, H., Gao, T., Ge, Y., Geng, S., Gong, C., Gong, X., Gongque, Z., Gu, Q., Gu, X., Gu, Y., Guan, L., Guo, Y., Hao, X., He, W., He, W., He, Y., Hong, C., Hu, H., Hu, J., Hu, Y., Hu, Z., Huang, K., Huang, R., Huang, W., Huang, Z., Jiang, T., Jiang, Z., Jin, X., Jing, Y., Lai, G., Li, A., Li, C., Li, C., Li, F., Li, G., Li, G., Li, H., Li, H., Li, J., Li, J., Li, J., Li, L., Li, M., Li, W., Li, W., Li, X., Li, X., Li, Y., Li, Y., Li, Y., Li, Y., Li, Z., Li, Z., Liao, W., Lin, J., Lin, X., Lin, Z., Lin, Z., Liu, C., Liu, C., Liu, H., Liu, L., Liu, S., Liu, S., Liu, S., Liu, T., Liu, T., Liu, W., Liu, X., Liu, Y., Liu, Y., Liu, Y., Liu, Y., Liu, Y., Liu, Z., Liu, Z., Lu, E., Lu, H., Lu, Z., Luo, J., Luo, T., Luo, Y., Ma, L., Ma, Y., Mao, S., Mei, Y., Men, X., Meng, F., Meng, Z., Miao, Y., Ni, M., Ouyang, K., Pan, S., Pang, B., Qian, Y., Qin, R., Qin, Z., Qiu, J., Qu, B., Shang, Z., Shao, Y., Shen, T., Shen, Z., Shi, J., Shi, L., Shi, S., Song, F., Song, P., Song, T., Song, X., Su, H., Su, J., Su, Z., Sui, L., Sun, J., Sun, J., Sun, T., Sung, F., Tai, Y., Tang, C., Tang, H., Tang, X., Tang, Z., Tao, J., Teng, S., Tian, C., Tian, P., Wang, A., Wang, B., Wang, C., Wang, C., Wang, C., Wang, D., Wang, D., Wang, D., Wang, F., Wang, H., Wang, H., Wang, H., Wang, H.,

Wang, H., Wang, J., Wang, J., Wang, J., Wang, K., Wang, L., Wang, Q., Wang, S., Wang, S., Wang, S., Wang, W., Wang, X., Wang, X., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Z., Wang, Z., Wang, Z., Wang, Z., Wang, Z., Wang, Z., Wei, C., Wei, M., Wen, C., Wen, Z., Wu, C., Wu, H., Wu, J., Wu, R., Wu, W., Wu, Y., Wu, Y., Wu, Y., Wu, Z., Xiao, C., Xie, J., Xie, X., Xie, Y., Xin, Y., Xing, B., Xu, B., Xu, J., Xu, J., Xu, J., Xu, L.H., Xu, L., Xu, S., Xu, W., Xu, X., Xu, X., Xu, Y., Xu, Y., Xu, Y., Xu, Z., Xu, Z., Yan, J., Yan, Y., Yang, G., Yang, H., Yang, J., Yang, K., Yang, N., Yang, R., Yang, X., Yang, X., Yang, Y., Yang, Y., Yang, Y., Yang, Z., Yang, Z., Yang, Z., Yao, H., Ye, D., Ye, W., Ye, Z., Yin, B., Yu, C., Yu, L., Yu, T., Yu, T., Yuan, E., Yuan, M., Yuan, X., Yue, Y., Zeng, W., Zha, D., Zhan, H., Zhang, D., Zhang, H., Zhang, J., Zhang, P., Zhang, Q., Zhang, R., Zhang, X., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Z., Zhao, C., Zhao, F., Zhao, J., Zhao, S., Zhao, X., Zhao, Y., Zhao, Z., Zheng, H., Zheng, R., Zheng, S., Zheng, T., Zhong, J., Zhong, L., Zhong, W., Zhou, M., Zhou, R., Zhou, X., Zhou, Z., Zhu, J., Zhu, L., Zhu, X., Zhu, Y., Zhu, Z., Zhuang, J., Zhuang, W., Zou, Y., Zu, X.: Kimi k2.5: Visual agentic intelligence (2026), https://arxiv.org/abs/2602.02276

54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
55. Wang, D., Shelhamer, E., Liu, S., et al.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
56. Wang, K.A., Shi, J., Fox, E.B.: Test-time regression: a unifying framework for designing sequence models with associative memory. arXiv preprint arXiv:2501.12352 (2025)
57. Wei, W., Luo, Z., Feng, L., Liong, V.E.: Spatial-aware vision language model for autonomous driving. arXiv preprint arXiv:2512.24331 (2025)
58. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
59. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv preprint arXiv:2506.09965 (2025)
60. xAI: Grok 4, https://x.ai/news/grok-4, accessed: 29 June 2026
61. Xu, R., Wang, X., Zhang, T., et al.: Pointllm: Empowering large language models to understand point clouds. In: CVPR (2024)
62. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10632–10643 (June 2025)
63. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025)
64. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Lu, D., Fergus, R., LeCun, Y., Fei-Fei, L., Xie, S.: Cambrian-s: Towards spatial supersensing in video (2025), https://arxiv.org/abs/2511.04670
65. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
66. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025)

67. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Dai, J., Jin, X.: Flash-vstream: Memory-based real-time understanding for long video streams. arXiv preprint arXiv:2406.08085 (2024)
68. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024)
69. Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W.T., Tan, H.: Test-time training done right. arXiv preprint arXiv:2505.23884 (2025), https://arxiv.org/abs/2505.23884
70. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)