Contents

A. Extended Model Architecture

Vision Encoder The universal vision encoder is built upon the architecture proposed in [6], which adopts a redesigned Vision Transformer (ViT) backbone. Specifically, the height and width of the input images are resized to multiples of 28 before being fed into the ViT. Then, the images are split into patches with a stride of 14. For video data, every two consecutive frames are grouped into a single unit to reduce the number of visual tokens and alleviate computational overhead. To address the quadratic complexity of standard self-attention when processing images of varying resolutions, a windowed attention with a maximum window size of 112×112 (corresponding to 8×8 patches) is introduced in most layers, ensuring that computational cost scales linearly with the number of patches rather than quadratically. Additionally, the positional encoding scheme is extended from 2D to 3D patch partitioning, enabling the encoder to jointly model both spatial and temporal dimensions for unified image and video understanding. To inject 3D information into conventional 2D video frames, following [58, 79], we first calculate a set of global coordinates (x, y, z) of each pixel at the position:

$$[x\ y\ z\ 1] = [\mathbf{D}_{i,j}\ \cdot\ [j\ i\ 1]\ \cdot (\mathbf{K}^{-1})^{\top}\ 1]\ \cdot \mathbf{B}^{\top}, \qquad (12)$$

where $\mathbf{D} \in \mathbb{R}^{H\times W}$ denotes the depth maps, $\mathbf{B} \in \mathbb{R}^{4\times 4}$ the extrinsic matrix, and $\mathbf{K} \in \mathbb{R}^{3\times 3}$ the camera intrinsic matrix. A sinusoidal positional encoding is then applied to these 3D coordinates to obtain coordinate embeddings, which are subsequently added to the visual token embeddings, forming position-aware representations that effectively capture spatial and temporal structure.

MoE Layer. Each MoE layers consists of 4 parallel expert modules, implmented as the FFN-styled architecture in the LLM. These experts expand the feature dimension from 3584 to an intermediate dimension of 18944 via parallel gated projections (gate_proj and up_proj), followed by SiLU activations and dimensional reduction back to 3584 (down_proj). Tokens are dynamically routed to these experts using a learnable Top-K gating network (i.e., a linear layer), which adaptively selects the most suitable experts based on token-level semantic characteristics. In our implementation, we set K to 2.

Video Foundation Model In our work, we employ V-JEPA 2 [4] as the video foundation Model. Architecturally, the model is parameterized as a vision transformer [] with a tubelet size of $2 \times 16 \times 16$ ($T \times H \times W$) to patchify the input video frames. A 3D-RoPE is adopted to encode relative position information in the vision transformer. V-JEPA 2 has been pre-trained on more than 1 million hours of video via a self-supervised manner to comprehensively capture the broad temporal motion and action within the video.

3D Foundation Model This work employs the VGGT [59] as the 3D foundation teacher model to distill 3D geometric information, such as inter-frame correspondences within input frames. The VGGT is built based on a fairly standard large transformer, with an alternating-attention design. Specifically, the frame-wise self-attention attends to the tokens within each frame separately, and global self-attention attends to the tokens across all frames jointly. The alternating attention is devised to make the transformer focus within each frame and globally alternately. During training, the VGGT is optimized to predict a full set of 3D attributes, including camera parameters, depth maps, point maps, and 3D point tracks. Recent works [16, 38, 59, 78] have shown that the VGGT can serve as a 3D genmetric feature extractor to enhance the downstream tasks.

Alignment Layer for Coarse-grained Synergtic Learning. To ensure dimensional compatibility between the generated synergistic latent features and the modality-specific priors extracted from video and 3D foundation models, we introduce dedicated alignment layers that project the synergy representations into the corresponding feature spaces. These alignment modules are intentionally designed with slight architectural differences, reflecting the distinct dimension nature of temporal features in videos and geometric features in 3D scenes.

B. More Implementation Details

B.1. CSQA Data Construction

Here, we show the detailed cross-vision synergy question-answer (CSQA) pair construction pipeline.

Data Source. We construct the CSQA dataset using paired scene-graph annotations from image–video and image–3D sources, derived from two established multimodal SG datasets [62]:

- **Image-Video Scene Graph** is constructed based on PVSG [66], AG [31], where the first frame of each video serves as the image view and non-adjacent clips provide the corresponding video segments. Cross-modal object associations are directly taken from the existing annotations, as follows.
- **Image-3D Scene Graph** is derived from 3DSG [3] by sampling 2D views linked to 3D scenes. Furthermore, the 3D objects are ground-annotated into the sampled images, and relationship annotations are inherited to form complete paired SGs.

QA Generation. To capture fine-grained cross-vision synergy, we design two categories of questions: **object-level** (e.g., spatial consistency, motion continuity) and **relation-level** (e.g., interaction dynamics, viewpoint-dependent changes), as illustrated in Fig. 8. Given the paired scene graphs, we prompt GPT-4o [29] to generate QA pairs targeting these specific aspects. For quality control, each generated question is validated using Qwen25-VL-72B [6], which answers the question based on both the scene graph and the corresponding visual inputs. Only QA pairs whose predicted answers match the gold responses are retained; mismatched samples are discarded to ensure reliability of the final CSQA dataset. Finally, we obtain the 20K CSQAs, which contain diversified questions and a focus.

Prompt for Object-level CSQA

Instruction: You are given paired scene graphs, one from an image and one from a video. Your task is to generate object-level, cross-vision question–answer pairs that explicitly compare or link the image to the video. Your generated QAs must rely only on the information encoded in the scene graphs.
Each QA should:
- (1) Be object-focused: Ask about a specific object (e.g., child, toy, sofa, ball); Not general high-level scene understanding; Not reasoning beyond objects & relations
- (2) Be cross-vision: Questions must compare the image with the video appearance or disappearance of an object; change of position; object motion; object interaction; new objects entering the scene; object state change.
- (3) Be explicitly grounded. The question must be answerable only using the provided scene graph information.
- (4) Be natural, concise, human-like

Important Constraints:
- Do NOT hallucinate objects, attributes, or relations not found in the scene graphs.
- Avoid global reasoning (e.g., "What happens in the scene?").
- Do NOT describe entire events; stay object-focused.
- Each question must be answerable and concise.
- Each answer must be grounded strictly in the paired SG content.
- Output QAs must be in JSON format.

==========
Examples (DO NOT COPY; FOLLOW SIMILAR STYLE):
{"question": "What is the child doing in the video compared to the image?", "answer": "The baby continues pushing the toy walker, later tends and interacts with it."},
{"question": "How does the toy in the image move in the video?", "answer": "The toy moves from the left side of the frame to the right as the baby pushes it."}
{"question": "Does any new object appear in the video that is not in the image?", "answer": "Yes, another adult and a ball appear in the later frames. "}
==========
Input Data: [*paired image-3D scene graph*]
Output QAs:

Prompt for Relation-level CSQA

Instruction: You are given a paired Image Scene Graph (ISG) and Video Scene Graph (VSG). Your task is to generate relation-level, cross-vision, synergy-focused question–answer pairs, comparing relational changes between the image and the video.
Each QA must:
- be relation-centered, i.e., focusing on pairwise or ternary relations
- Connect image and video explicitly. Questions must compare: whether a relation persists or changes; when a relation appears / disappears; how a relation evolves over time; whether a relation becomes stronger/weaker or switches type; whether a new relation emerges only in the video.
- Stay grounded in the scene graphs, no hallucinated objects or relations.
- Produce natural, human-like questions

Important Constraints:
- Only use relations that appear explicitly in the scene graphs.
- No event hallucination.
- Keep questions short, clear, and relation-focused.
- Answers must be concise and grounded (no speculation).
- Avoid high-level reasoning beyond relations.
- Output Q&As must be in JSON format.

==========
Examples (DO NOT COPY; FOLLOW SIMILAR STYLE):

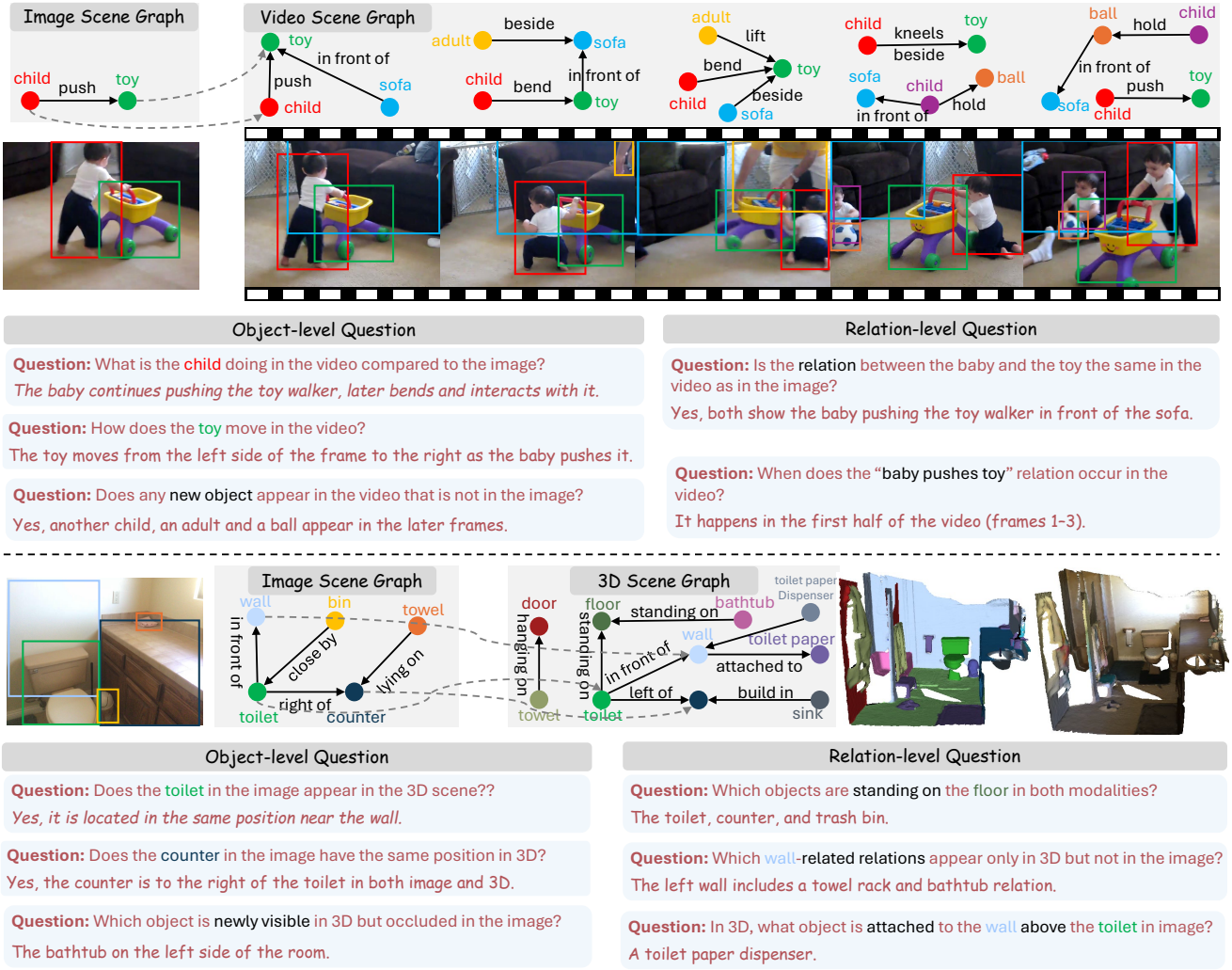

Figure 8. Illustration of CSQA examples, which is constructed based on a paired Image-Video scene graph and Image-3D scene graph.

{"question": "Is the relation between the baby and the toy the same in the video as in the image?", "answer": "Yes, both show the baby pushing the toy walker in front of the sofa."},
{"question": "Does any new relation appear in the video that is not present in the image?", "answer": "Yes. The man and another child appear in front of the sofa."}
==========
Input Data: textual captions
Output QAs:

B.2. Model Details.

We initialize the LLM backbone with Qwen2.5-VL-7B [6], which supports both image and video inputs. Following [79], we further incorporate 3D positional embeddings to enhance the model's spatial understanding. In the second training stage, MoE and dense layers are interleaved at four-layer intervals (i.e., layers 0, 4, 8, 12, 16, 20, 24, and 28) to balance efficiency and capacity. Each MoE layer consists of 4 experts, and the top-2 experts with the highest routing probabilities are dynamically activated for each token during inference. We show the detailed architecture and parameters number in Table 5.

B.3. Training Details

We provide additional details of the full training pipeline in this section. The complete set of hyperparameter configurations and training settings is summarized in Table 6.

Stage-1-1: Alignment Pre-training In this stage, we establish the initial alignment between the vision encoder and the backbone LLM. Since our model is initialized from Qwen2.5-VL-7B, which already supports image and video understanding, we focus exclusively on optimizing the model with 3D data to equip it with an initial capability

Table 5. Detailed Architecture of PolyV and Qwen2.5-VL(Base). "Width" represents the dimension of the hidden states. "FFN" denotes the dimension of the feed-forward network's intermediate layer. "FFN Factor" represents the quantity of linear layers in the FFN. "Activated" or "Total Param" refers to the activated or total number of parameters. We highlight the architecture of PolyV in blue.

Name	Experts	Top-k	MoE Layers	Embedding	Width	Layers	FFN	FFN Factor	Heads	Activated Param.	Total Param.
Base	-	-	-	152064	3584	28	18944	3	28	7.0B	7.0B
First Half	4	2	7	152064	3584	28	18944	3	28	8.4B	11.3B
Last Half	4	2	7	152064	3584	28	18944	3	28	8.4B	11.3B
Interval(4)	4	2	7	152064	3584	28	18944	3	28	8.4B	11.3B
Full	4	2	14	152064	3584	28	18944	3	28	9.9B	15.6B

Table 6. The hyperparameter settings in each stage.

Parameters	Stage-1-1	Stage-1-2	Stage-2-1	Stage-2-2
Optimizer	AdamW	AdamW	AdamW	AdamW
Scheduler	Cosine	Cosine	Cosine	Constant
Learning rate	2e-5	2e-6	1e-4	1e-5
Weight Decay	0.05	0.05	0.05	0.1
Warmup Ratio	0.03	0.03	0.05	0.05
Updated Module	Projector	FFN	Router FFN (Experts) Alignment	Router FFN (Experts)
Data Num.	40K	45K	50K	40K
Data Source	LLaVA-3D-Instruct-860K	LLaVA 1.5-558K LLaVA-Video-178K LLaVA-3D-Instruct-860K	ShareGPT4Video 3RScan Scan2Cap	CSQA LLaVA 1.5-Mix-665K LLaVA-Video-178K LLaVA-3D-Instruct-860K

for 3D comprehension. During this process, only the parameters of the projector are updated, while all other components remain frozen.

Stage-1-2: Instruction-Following Pre-training In this stage, we train separate model variants on modality-specific data, allowing the model to acquire specialized capabilities, such as temporal dynamics for videos and spatial geometry for 3D scenes, using richer vision–language instruction corpora. During this process, only the feed-forward network (FFN) parameters are updated, while all other components remain frozen. This design allows modality-specific expertise to be acquired efficiently without interfering with the shared cross-modal representations established in the previous stage.

Stage-2-1: Coarse-grained Synesthetic Learning This stage aims to equip the model with cross-vision synergy, enabling synesthetic reasoning across different visual modalities. Architecturally, the modality-specialized FFNs obtained from Stage-1-2 are used to initialize the MoE layers. During training, we first freeze all experts and update only the router and alignment layers to stabilize cross-modal routing behavior. Once the routing becomes stable, we subsequently unfreeze the experts and jointly optimize the experts, router, and alignment layers to achieve coarse-grained synesthetic fusion across modalities.

Stage-2-2: Fine-grained Synesthetic Learning In this stage, the model is further fine-tuned using the constructed CSQA dataset to enhance its fine-grained cross-vision synesthetic reasoning capabilities. To maintain the model's generality, we interleave a proportion of the modality-specific instruction data used in Stage 1-2 during training. Consequently, this stage primarily updates the MoE layers, enabling more precise and discriminative synergistic interactions across modalities.

Overall Parameters settings. Across all stages, the models are optimized using the AdamW optimizer [43] in conjunction with a cosine learning rate scheduler. A warm-up schedule with a warm-up ratio of 0.03 followed by cosine/-constant decay is applied independently for each stage. To enhance computational efficiency and reduce memory usage, we employ BF16 mixed-precision training. For video and 3D multi-view inputs, the maximum number of frames is set to 32 with 640×480 resolution.

B.4. Benchmarks

We detail the benchmarks we employed in our evaluation.

MMStar [11]. MMStar is a large-scale benchmark designed to evaluate truly vision-indispensable capabilities of multimodal models. The dataset contains 1,500 carefully curated human-verified samples across six core abilities and eighteen sub-dimensions. It mitigates two major issues in existing benchmarks—visual redundancy and potential data leakage, by ensuring that each question fundamentally requires visual understanding.

3DSRBench [45]. 3DSRBench focuses on 3D spatial reasoning, assessing whether models can understand object height, orientation, position, inter-object relations, and multi-view consistency under both common and uncommon viewpoints. The benchmark consists of 2,772 manually annotated VQA samples, including approximately 2,100 real-world images and 672 synthetic multi-view images.

MMSI-Bench [68]. MMSI-Bench evaluates multi-image spatial intelligence, requiring models to integrate information across multiple views rather than reasoning from a single image. The benchmark comprises 1,000 carefully constructed multiple-choice questions sourced from over 120,000 candidate images, each accompanied by human-designed step-by-step reasoning. The benchmark includes a taxonomy that covers grounding, scene reconstruction, contextual transformation, and spatial logic failures.

CV-Bench [57]. CV-Bench offers a vision-centric evaluation suite designed for 2D and 3D understanding in multimodal models. It contains 2,638 human-validated samples built upon datasets such as ADE20K [80], COCO [39], and Omni3D [9]. The benchmark focuses on determining whether LVLMs truly rely on visual information rather than linguistic priors, emphasizing core perception capabilities such as localization, recognition, and spatial reasoning.

VSI-Bench [67]. VSI-Bench evaluates visual spatial intelligence from videos, extending spatial understanding beyond static imagery. It includes over 5,000 QA pairs derived from 288 real videos. These videos are sourced from the validation sets of the public indoor 3D scene reconstruction datasets ScanNet [17], ScanNet++ [70], and ARKitScenes [8]. Tasks encompass configurational reasoning, metric estimation, and spatiotemporal reasoning, enabling a comprehensive assessment of spatial understanding.

Video-MME [21]. Video-MME is the first comprehensive full-spectrum video evaluation suite designed for multimodal LLMs, covering 900 videos (≈254 hours) and 2,700 manually aligned QA pairs across 6 domains and 30 fine-grained tasks. The benchmark integrates multimodal video inputs, including frames, audio, and subtitles, and spans a diverse range of durations from short clips to hour-long content. It is designed to evaluate a model's holistic video understanding, including temporal grounding, auditory reasoning, and long-range visual dependency modeling.

CVBench [83]. CVBench is a benchmark targeting cross-video relational reasoning, where multimodal models must jointly analyze multiple video streams to infer object associations, temporal relationships, and event-level dependencies. It contains approximately 1,000 carefully annotated QA pairs spanning object-level association, event matching, cross-video correspondence, and higher-order reasoning. The benchmark specifically evaluates a model's ability to integrate visual evidence across disjoint videos, an ability that current MLLMs often lack due to their limited temporal relation modeling capability.

STI-Bench [35]. STI-Bench focuses on fine-grained spatial-temporal intelligence in real-world scenarios such as robot manipulation or vehicle operation. It evaluates whether models can accurately reason about object appearance, pose, displacement, and dynamic state changes across desktop, indoor, and outdoor videos. The benchmark emphasizes physically grounded prediction, challenging models to infer where objects were, are, and will be and revealing substantial limitations in spatial-temporal consistency of current VLMs.

DSI-Bench [77] DSI-Bench is designed to systematically measure dynamic spatial intelligence by decomposing 3D spatial reasoning into independent motion factors. It contains nearly 1,000 dynamic videos and over 1,700 expert-verified QA pairs, covering nine fundamental motion patterns involving both moving observers and moving objects. By isolating each motion factor, such as ego-motion, object translation, rotation, and coordinated interaction, the benchmark provides a principled way to diagnose which dimensions of spatial reasoning current models fail to capture.

OpenEQA [47] OpenEQA is a large-scale benchmark designed for embodied question answering (EQA) in the era of foundation models, aiming to measure how well agents can perceive, navigate, and reason within 3D embodied environments. It provides more than 1,000 expert-verified EQA samples across diverse scenes and question types, covering perception, spatial reasoning, object grounding, action planning, and commonsense reasoning.

B.5. Baselines

Besides, the general VLM baselines (e.g., LLaVA-v1.5 [41], LLaVA-NeXT-Video [74], LLaVA-Video [75], LLaVA-Onevision [2, 33], InternVL [15, 82], Qwen2.5/3-VL [6]), we also compared our methods with baselines that

Table 7. Detailed evaluation results on **MMStar** [11]. Best results are marked in **bold**.

Model	Coarse Perception	Fine-grained Perception	Instance Reasoning	Logical Reasoning	Math	Science Technology	Average
LLaVA-NeXT-Video-7B [74]	52.4	25.6	41.6	28.0	29.2	25.2	33.7
LLaVA-OneVision-7B [33]	65.2	51.2	64.8	58.0	53.2	44.4	56.1
LLaVA-OneVision1.5-8B [2]	68.4	53.6	70.0	70.8	**77.6**	**59.2**	66.6
InternVL3-8B [15]	74.0	59.2	72.0	65.6	64.4	54.4	64.9
Qwen3-VL-7B [6]	75.2	58.0	69.2	51.2	34.8	41.6	55.0
SpaceR [49]	66.0	48.4	68.0	56.0	44.8	38.4	53.6
SpatialMLLM [60]	63.6	43.2	55.6	47.6	54.0	30.8	49.1
LLaVA3D [81]	54.8	22.8	41.6	28.0	19.2	24.0	31.7
Video3D-LLM [79]	23.6	3.2	18.8	11.6	14.8	18.0	15.0
Ross3D [58]	41.2	10.0	30.8	20.0	12.0	16.0	21.7
PloyV (Ours)	**82.0**	**70.0**	**78.0**	**74.0**	68.0	57.0	**71.4**

Table 8. Detailed evaluation results on **3DSRBench**$_{\text{real}}$ [45]. 'Loc.' denotes location, 'Orient.' means orientation, and 'Multi.' is multi-object.

Model	Overall	Height	Loc.	Orient.	Multi.
LLaVA-v1.5-7B [41]	38.1	39.1	46.9	28.7	34.7
LLaVA-NeXT-Video-7B [74]	49.4	52.6	55.3	42.6	48.5
LLaVA-OneVision-7B [33]	54.4	56.8	61.3	46.1	50.3
LLaVA-OneVision1.5-8B [2]	57.8	57.4	66.5	49.9	53.4
InternVL2.5-8B [15]	50.9	45.9	68.1	38.7	43.3
InternVL3-8B [82]	58.1	**61.3**	67.6	48.3	53.8
QWen2.5-VL-7B [6]	48.4	44.1	62.7	40.6	40.5
Qwen3-VL-8B [6]	60.0	53.8	74.0	51.5	53.6
LLaVA3D [81]	39.3	49.1	36.8	25.0	46.3
Ross3D [58]	39.6	36.5	40.2	35.9	42.4
SpaceR [49]	57.4	51.5	69.7	50.1	51.9
SpatialMLLM [60]	47.8	49.6	49.0	42.4	49.0
PloyV (Ours)	**63.4**	60.0	**78.0**	**57.0**	**58.0**

have been fine-tuned on the spatial-focused datasets, including

- **LLaVA3D** [81] is built upon the LLaVA-Video-7B [75], where the vision encoder is extended with a 3D-aware adapter that fuses RGB frames with 3D geometric cues. It is fine-tuned on large-scale 3D-annotated datasets such as LLaVA-3D-Instruct-86K [81], and MMScan QA [44].
- **SpaceR** [49] employs a Qwen-2.5-VL-7B-Instruct [6] as the backbone. The model is fine-tuned on a carefully cured dataset, i.e., SpaceR-151k.
- **SpatialMLLM** [60] adopts the Qwen2.5-VL-3B [6] as backbone and VGGT [59] as the spatial encoder. It is trained on the Spatial-MLLM-120k.
- **Video3D-LLM** [79] builds on the LLaVA-Video-7B [75]. The model is fine-tuned on 3D-awared datasets, including ScanRefer [10], Multi3DRefer [73], Scan2Cap [14], ScanQA [5] and SQA3D [46].
- **Ross3D** [58] is developed based on LLaVA-Video-7B [75], with cross-view and global-view reconstructions, enabling accurate spatial relationship modeling and comprehensive scene layout comprehension. It is fine-tuned on the combination of training sets of SQA3D [46], ScanQA [5], Scan2Cap [14], ScanRefer [10], and Multi3DRefer [73] datasets.

C. More Experimental Results

C.1. Detailed Benchmark Results

We provide more detailed results on the existing benchmark in Table 7, 8, 9, 10, 11, 13, and 12. Moreover, we also conduct experiments on more focus on cross-vision synergy benchmark, including CV-Bench [57], STI-Bench [35], and DSI-Bench [77]. As results demonstrated in Table 14, 16, and 15, the proposed PolyV consistently achieves superior average performance compared with all baselines, demonstrating its strong cross-vision synergistic reasoning capability.

C.2. Mitigating Catastrophic Forgetting Across Modalities

Our experiments reveal that models fine-tuned solely on spatial-centric data tend to suffer from catastrophic forgetting of previously acquired visual knowledge. For example, although SpaceR and Qwen2.5-VL-7B share the same backbone, SpaceR exhibits noticeably degraded performance on most benchmarks after training, indicating that its incremental learning strategy inadvertently erodes the image- and video-based competencies learned during pre-training. Similar degradation is observed for LLaVA-Video and other 3D-aware models when evaluated on VSI-Bench (cf. Table [67]), suggesting that current approaches struggle to balance the acquisition of new spatial capabilities with the retention of prior modality knowledge. In contrast, our proposed PolyV effectively avoids such forgetting.

Table 9. Detailed evaluation results on **MMSI-Bench** [68]. 'Cam.' denotes the camera, 'Obj.' means the object, 'Reg.' is the region, 'Meas.' is the measurement, and 'Appr.' is the appearance.

Model	Positional Relationship						Attribute		Motion		MSR	Average
	Cam.–Cam.	Obj.–Obj.	Reg.–Reg.	Cam.–Obj.	Obj.–Reg.	Cam.–Reg.	Meas.	Appr.	Cam.	Obj.		
LLaVA-v1.5-7B [42]	33.3	31.9	22.2	29.1	22.3	20.5	25.0	22.7	14.9	21.0	22.2	24.2
LLaVA-NeXT-Video-7B [74]	21.5	22.3	24.7	26.7	23.5	22.9	32.8	27.3	16.2	26.3	27.8	24.9
Qwen2.5-VL-7B [6]	24.7	24.5	24.7	25.6	29.4	26.5	25.0	18.2	20.3	**39.5**	25.8	25.9
Qwen3-VL-8B [6]	31.2	27.7	28.4	24.4	28.2	**34.9**	**35.9**	15.1	**27.0**	38.2	28.3	29.0
InternVL2.5-8B [15]	32.3	27.7	29.6	32.6	24.7	32.5	26.6	27.3	16.2	31.6	**30.3**	28.7
InternVL3-8B [82]	25.8	31.9	**37.0**	25.6	**35.3**	28.9	23.4	24.2	16.2	32.9	14.6	25.7
LLaVA-OneVision-7B [33]	20.4	**33.0**	29.6	29.1	25.9	30.1	29.7	25.8	18.9	34.2	11.6	24.5
LLaVA-OneVision1.5-8B [2]	**34.4**	26.6	33.3	33.7	34.1	25.3	25.0	28.8	22.9	30.3	29.3	29.6
LLaVA3D [81]	18.3	18.1	16.0	15.1	21.2	22.9	29.7	24.2	24.3	27.6	18.7	20.8
Ross3D [58]	12.9	12.8	13.6	11.6	23.5	25.3	31.2	6.06	5.4	23.7	16.2	16.4
SpaceR [49]	29.0	26.6	24.7	30.2	30.6	27.7	23.4	21.2	20.3	34.2	27.8	27.2
SpatialMLLM [60]	25.8	25.5	28.4	34.9	24.7	25.3	18.7	33.3	12.2	23.7	26.8	25.7
Video3D-LLM [79]	15.0	4.3	6.2	1.2	9.4	2.4	1.6	3.0	1.3	9.2	7.1	5.9
PolyV (Ours)	33.0	30.0	36.0	33.0	34.0	30.0	28.0	**30.0**	26.0	38.0	30.0	**31.7**

Table 10. Detailed evaluation results on **VSI-Bench** [67]. object count (Obj. Count), absolute distance (Abs. Dist.), object size (Obj. Size), room size, relative distance (Rel. Dist.), relative direction (Rel. Dir.), route plan, appearance order (Appr. Order)

Model	Average	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
LLaVA-Video-7B [75]	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-7B [74]	29.9	53.6	30.6	21.6	38.8	27.7	33.9	31.4	16.1
LLaVA-OneVision-7B [33]	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision1.5-8B [2]	46.3	71.2	35.1	**69.3**	60.6	38.3	39.1	29.9	25.2
InternVL3-8B [82]	50.7	**75.8**	52.2	65.2	**70.4**	36.8	36.4	37.3	**37.2**
Video3d-LLM [79]	15.8	17.5	0.0	0.0	0.0	29.6	33.4	26.3	24.8
SpaceR [49]	24.5	33.1	0.0	2.5	0.0	41.7	**44.7**	32.0	46.4
LLaVA3D [81]	9.4	2.5	0.0	0.1	0.0	31.8	6.1	27.8	18.9
PolyV (Ours)	**52.7**	74.6	**56.0**	65.0	70.0	**45.0**	42.0	**40.0**	29.0

Table 11. Detailed evaluation results on **CVBench** [83]. The tasks include: multi-view scene understanding (M. SU), multi-video temporal reasoning (M. TR), joint-video spatial navigation (J. SN), video difference captioning (VDC), cross-video counterfactual reasoning (C. CR), joint-video summarization (J.S), and cross-video procedural transfer (C. PT).

Model	Average	M. SU	M. TR	J. SN	VDC	C.CR	J.S	C.PT
LLaVA-NeXT-Video-7B [74]	30.9	5.3	14.6	14.6	20.0	17.3	5.8	7.8
Qwen2.5-VL-7B [6]	51.3	80.0	22.7	26.2	50.9	55.8	69.2	60.8
InternVL2.5-8B [15]	**59.4**	83.6	26.7	50.0	60.0	**69.2**	67.3	**68.6**
Qwen2.5-VL-8B [6]	51.1	80.0	31.1	47.6	**71.1**	66.7	71.1	68.0
LLaVA-OneVision-7B [33]	52.6	83.6	**40.0**	38.1	45.5	42.3	61.5	52.9
LLaVA-OneVision1.5-8B [2]	43.3	7.5	23.4	23.4	28.8	27.7	8.7	12.5
Video3D-LLM [79]	44.6	61.5	32.9	25.0	44.0	42.9	56.9	70.0
Spatial-MLLM [60]	38.2	74.5	29.7	33.3	41.8	46.0	57.7	46.0
SpaceR [49]	50.4	83.6	31.1	26.2	49.1	54.9	**76.9**	58.0
Ross3D [58]	42.1	73.7	30.4	35.9	41.2	62.5	72.4	48.0
PolyV (Ours)	59.1	**86.0**	38.0	**52.0**	58.0	63.0	68.0	49.0

After learning from diverse modality-specific data, PolyV not only maintains strong performance on original image- and video-based benchmarks but also successfully incorporates additional spatial understanding. This demonstrates its ability to achieve truly synergistic multi-modality learning without sacrificing previously learned capabilities.

Table 12. Detailed evaluation results on **OpenEQA**$_{\text{HM3D}}$ [47]. 'Obj. Rec.' denotes object recognition, 'Attr. Rec.' is attribute recognition, 'Spatial Und.' is the spatial understanding, 'Obj. State Rec.' is the object state recognition, 'Func. Rea.' is the function reasoning, 'World Know.' is the world knowledge, ' Obj. Loc.' is the object localization.

Model	Obj. Rec.	Attr. Rec.	Spatial Und.	Obj. State Rec.	Func. Rea.	World Know.	Obj. Loc.	Average
LLaVA-NeXT-Video-7B [74]	31.5	34.1	35.0	31.2	51.0	31.7	30.2	34.9
LLaVA-OneVision-7B [33]	35.1	42.5	33.8	35.0	53.7	35.9	34.6	38.7
LLaVA-OneVision1.5-8B [2]	37.5	45.6	36.8	37.2	58.9	38.0	36.4	41.2
InternVL2.5-8B [15]	42.3	51.0	40.0	41.3	68.0	41.5	40.5	46.7
InternVL3-8B [15]	44.4	51.1	42.5	43.5	**68.7**	43.6	42.5	48.0
Qwen2.5-VL-7B [6]	50.4	62.1	49.3	66.4	61.4	60.3	47.6	56.6
Spatial-MLLM [60]	33.7	9.7	15.5	1.7	31.2	21.2	17.0	18.0
Video3d-LLM [79]	33.2	41.2	47.0	60.0	45.5	43.7	35.7	43.7
Ross3D [58]	41.7	59.5	52.5	64.2	57.5	55.5	39.7	52.5
PolyV (Ours)	**60.8**	**66.7**	**57.9**	**70.4**	68.6	61.8	**58.2**	**63.4**

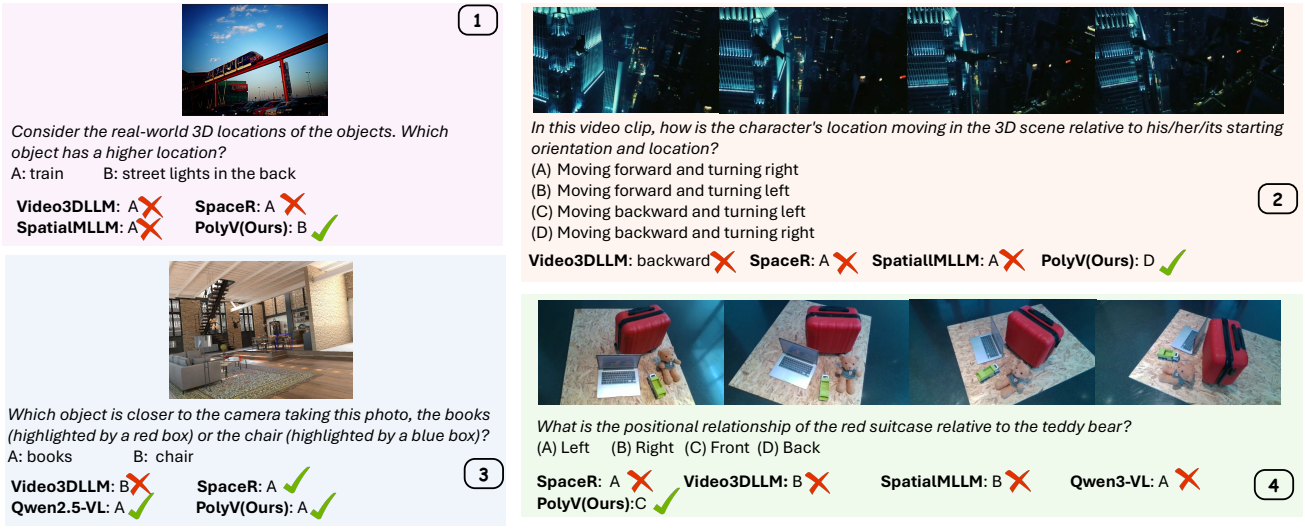

Figure 9. Qualitative comparisons between PolyV and existing models. Case 1 is from 3DSRBench [45], Case 2 from DSI-Bench [77], Case 3 from CV-Bench [57], and Case 4 from STI-Bench [35].

Table 13. Detailed evaluation results on **VideoMME**$_{\text{w/o sub}}$ [21].

Model	Short	Mid	Long	Average
LLaVA-NeXT-Video-34B [74]	61.7	50.1	44.3	52.0
Qwen3-VL-8B [6]	58.1	49.2	49.1	52.1
LLaVA-3D [81]	27.1	27.0	25.4	26.5
SpaceR [49]	67.0	55.4	48.2	56.9
Spatial-MLLM [60]	54.4	42.0	35.8	44.1
Video3d-LLM [79]	47.7	42.0	37.1	41.9
Ross3D [58]	44.1	39.9	38.9	40.9
PolyV (Ours)	**75.2**	**70.1**	**63.5**	**69.6**

Table 14. Detailed evaluation results on **CV-Bench** [57].

Model	2D		3D		Average
	Relation	Count	Depth	Distance	
Qwen2.5-VL-7B [6]	77.4	63.2	74.2	44.0	64.8
Qwen3-VL-8B [6]	92.5	71.7	95.0	58.7	79.1
SpaceR [49]	83.7	65.1	78.5	**66.0**	72.9
Spatial-MLLM [60]	62.3	53.7	60.5	65.7	60.1
Video3d-LLM [79]	53.2	18.4	56.2	60.5	45.1
Ross3D [58]	13.8	6.3	33.2	24.2	18.3
PolyV (Ours)	**96.0**	**82.0**	**98.0**	58.0	**83.5**

C.3. More Quantitation Results

We provide more quantitation results in Fig. 9.

D. Limitation and Future Work

Although we believe that synesthetic, cross-vision synergy is a fundamental capability that future vision–language models must possess—especially for downstream applica-

Table 15. Detailed evaluation results on **DSI-Bench** [77] across different video sources.

Model	CamerBench [40]	SynFMC [53]	internet	k700 [55]	llava178k [75]	Average
LLaVA-NeXT-Video-7B [74]	36.0	28.4	36.9	33.6	35.8	35.6
LLaVA-OneVision-7B [33]	46.0	33.0	48.1	47.5	45.3	46.4
LLaVA-OneVision1.5-8B [2]	50.4	48.6	50.3	57.0	**49.6**	51.1
InternVL2.5-8B [15]	47.6	48.6	50.4	57.4	46.8	50.4
InternVL3-8B [15]	52.8	43.1	45.9	47.5	47.6	47.2
Qwen2.5-VL-7B [6]	40.0	38.5	39.4	41.2	48.1	40.5
Qwen3-VL-8B [6]	38.3	37.6	39.5	37.4	40.3	39.0
SpaceR [49]	39.4	41.3	39.5	43.1	43.6	40.7
Spatial-MLLM [60]	27.5	24.8	30.7	29.8	31.2	29.8
Video3d-LLM [79]	39.9	34.8	38.9	49.6	38.2	40.0
Ross3D [58]	37.1	32.3	34.3	37.3	30.8	34.5
PolyV (Ours)	**60.8**	**57.4**	**59.3**	**67.9**	48.1	**58.7**

Table 16. Detailed evaluation results on **STI-Bench** [35]. 'Dim. Meas' is the Dimensional Measurement, 'Disp. & P.L.' means Displacement & Path Length, 'Speed & Acc.' denotes Speed & Acceleration, 'Ego Orient.' indicates Ego-Centric Orientation, 'Traj. Desc.' is Trajectory Description, and 'Pose Est.' is Pose Estimation.

Model	Static Understanding			Dynamic Understanding					Average
	Dim. Meas	Spatial Relation	3D Video Grounding	Disp. & P.L.	Speed & Acc.	Ego Orient.	Traj. Desc.	Pose Est.	
LLaVA-NeXT-Video-7B [74]	22.1	42.5	19.9	21.5	21.1	27.0	42.3	18.9	23.6
LLaVA-OneVision-7B [33]	25.6	30.8	24.6	25.1	30.8	41.6	51.3	**55.8**	34.2
LLaVA-OneVision1.5-8B [2]	**34.3**	29.4	33.1	24.9	**35.0**	63.8	44.9	55.6	39.0
InternVL2.5-8B [15]	26.3	50.7	31.5	20.9	29.3	**66.5**	46.1	56.7	38.0
InternVL3-8B [15]	26.3	52.0	27.1	21.5	33.2	0.2	25.6	40.6	30.4
Qwen2.5-VL-7B [6]	24.5	46.6	32.8	24.6	30.6	17.8	43.6	47.2	32.4
Qwen3-VL-8B [6]	25.9	50.7	37.8	19.6	26.7	55.7	39.7	56.1	37.0
SpaceR [49]	21.4	47.3	30.3	23.0	29.4	26.5	47.4	49.2	32.4
Spatial-MLLM [60]	23.5	42.5	17.0	17.9	12.1	51.3	28.2	26.9	24.3
Video3d-LLM [79]	22.5	42.5	20.5	19.3	23.9	13.5	39.7	17.2	22.2
PolyV (Ours)	33.4	**55.8**	**40.1**	**26.3**	34.7	59.2	**48.6**	36.3	**46.8**

tions such as robotics and autonomous driving—our current approach still faces several limitations. First, despite the use of a MoE architecture, which reduces computational cost during deployment by activating only a subset of experts, the overall model remains resource-intensive. Efficient scaling and lightweight deployment thus remain open challenges. Second, video and 3D processing inherently require handling large numbers of frames or point-cloud tokens. Current input-length constraints limit the model's ability to fully capture long-term dynamics or fine-grained spatial structures. Furthermore, our model acquires synesthetic ability primarily through supervised fine-tuning. While effective, this approach may not fully exploit the model's potential for self-thinking.

For future work, we plan to explore reinforcement learning–based cross-vision synergy. By encouraging the model to explicitly reason about spatial structure during its thinking process, e.g., inferring 3D geometry from 2D images or predicting spatial evolution from video cues, the model may develop more robust, self-refining synesthetic abilities. Such advances are expected to yield stronger performance across a wide spectrum of vision-centric downstream tasks.