Beyond 3D VQAs: Injecting 3D Spatial Priors into Vision-Language Models for Enhanced Geometric Reasoning

Chun-Hsiao Yeh[1,2]*, Shengyi Qian[1], Manchen Wang[1], Yi Ma[2,3], Joseph Tighe[1], Fanyi Xiao[1]
[1]FAIR at Meta [2]UC Berkeley [3]HKU

Abstract

Vision-Language Models (VLMs) often struggle with robust 3D spatial reasoning. Prevailing methods that rely on fine-tuning with 3D visual question-answering (VQA) datasets may overfit dataset-specific biases, while integrating specialized 3D visual encoders is often inflexible and cumbersome. In this paper, we argue that genuine spatial understanding should emerge from learning fundamental geometric priors, not only from high-level VQA supervision. We propose ***GASP (Geometric-Aware Spatial Priors)****, a framework that injects these priors directly into the LLM's transformer layers. GASP employs a small correspondence head, applied as a deep supervision signal across all layers, and is trained with a dual objective leveraging ground-truth geometry from large-scale video scenes: a contrastive loss on ground-truth point correspondences enforces 2D view-invariance, while a depth consistency supervision resolves 3D geometric ambiguities. Our analysis first provides a diagnostic showing that standard VLMs' internal correspondence matching accuracy is very low (often below 5%). We then demonstrate that our training substantially improves this behavior, boosting peak layer-wise correspondence to over 70% and maintaining over 85% temporal robustness while baselines remain below 5%. These internal improvements translate to significant gains on downstream spatial benchmarks—including* ***+18.2%*** *on All-Angles Bench and* ***+29.0%*** *on VSI-Bench—all without training on any 3D VQA data. Our findings indicate that learning from fundamental geometric priors is a promising and generalizable pathway towards VLMs with more reliable 3D spatial reasoning.*

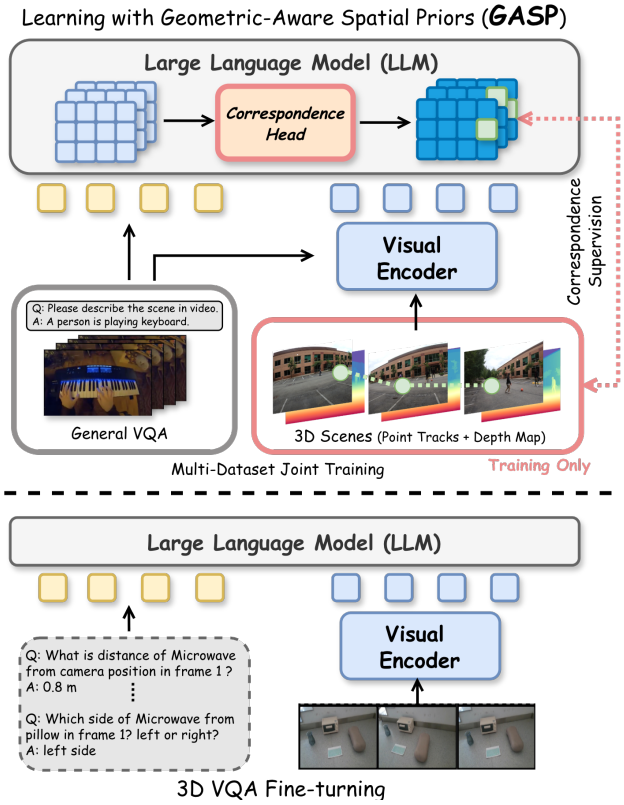

Figure 1. **Top**: Our proposed framework (**GASP**) learns geometric consistency by injecting the correspondence head into the LLM, supervised by 3D spatial priors. **Bottom**: Standard spatial VLMs rely on fine-tuning with 3D VQA datasets, which often leads to memorizing data-specific biases. *Note that our GASP requires no 3D prior input and processes as a standard VLM during inference.*

1. Introduction

The ability to perform robust spatial reasoning is a cornerstone of artificial intelligence, enabling agents to understand, navigate, and interact with the complex real world [37, 38]. In recent years, Vision-Language Models (VLMs) have demonstrated remarkable capabilities in multimodal understanding and reasoning [1, 2, 5, 17, 24, 40], yet their grasp of spatial concepts remains a significant challenge [7, 16, 22, 27, 32, 57]. A dominant paradigm to address this limitation involves fine-tuning these models on

*The work was done during CHY's internship at FAIR.

extensive 3D visual question-answering (VQA) datasets [6, 33, 39, 45, 47, 48, 58, 63]. Although effective to some extent, post-training strategies such as supervised fine-tuning (SFT) and reinforcement learning (RL) on these VQA pairs often encourages models to learn superficial correlations and memorize dataset-specific biases, leading to poor generalization on unseen scenarios. For example, as the experiments shown in [39], specialized models like VILASR [48], SpatialMLLM [47], and VG-LLM [61] show huge performance boosts on in-domain benchmark like VSI-Bench [54] after fine-tuning. However, these models show a consistent performance drop on out-of-domain spatial benchmarks such as MMSI-Bench [55], STI-Bench [25], and Space-Vista [39]. An alternative line of works [8, 47, 61] seek to extract 3D spatial information by integrating specialized visual encoders such as the VGGT [43] model, or by using direct 3D inputs like point clouds [4], pre-segmented objects [15, 46] or BEV maps [35]. However, this path presents significant practical limitations. These pre-trained spatial encoders are cumbersome, increasing model size and inference latency. Furthermore, they must typically be used "as is" (i.e., with frozen weights) because their specialized 3D training data and pipelines are incompatible with standard VLM training. This creates a challenging integration problem, forcing the model to align its native visual representations with these rigid, pre-computed 3D features.

In this work, we depart from both of these prevailing paradigms. *We argue that robust spatial intelligence emerges from learning the fundamental perceptual signals of 3D geometry.* We hypothesize that true spatial understanding is underpinned by the ability to establish visual correspondences across changing viewpoints. Rather than teaching a model to associate text with visual patterns, our goal is to teach it the underlying geometric consistency of the world itself. Learning this object constancy encourages the model to build an internal, view-invariant representation [23, 30, 34], providing a more generalizable foundation for downstream spatial reasoning tasks.

To this end, we propose **GASP (Geometric-Aware Spatial Priors)**, a novel training framework designed to directly inject geometric priors into the LLM transformer layers of the VLM's backbone shown in Figure 1. Our method introduces a lightweight correspondence head inserted across all transformer layers to receive deep supervision signal. This forces geometric consistency to be maintained at every stage of the model's feature representation. This head is trained with a dual objective leveraging ground-truth geometric priors from the large-scale video scenes [26]: First, a contrastive learning objective on point correspondence data across frames teaches the model the core principle of *object constancy*, forcing it to learn view-invariant 2D representations. Second, a depth consistency loss leverages ground-truth depth maps as a crucial geometric regularizer to resolve 3D ambiguities (i.e., foreground-background matching confusion) through matching depth values. Crucially, this correspondence head is only active during the training phase and is discarded entirely for inference.

We validate the effectiveness of our approach through extensive ablation experiments. We provide a novel visual correspondence matching analysis for VLM's backbones, and reveal that GASP dramatically improves the VLM's internal geometric representations, in terms of both the significantly improved correspondence matching scores, as well as its capability to maintain robust matching across long temporal range. Moreover, we also demonstrate that these internal improvements generalize to high-level reasoning. Our GASP achieves significant performance gains on downstream spatial reasoning benchmarks by improving camera pose estimation by +18.2% on the All-Angles Bench [56], object counting by +29.0% on VSI-Bench [54], and multi-view reasoning by +15.0% on BLINK [11]. Our contributions are summarized as follows:

- We introduce **GASP**, a novel framework that injects geometric priors directly into the LLM's transformer layers. GASP uses a deep supervision signal across all layers and is trained with a dual point correspondence and depth consistency to resolve 3D ambiguities.
- We provide a detailed correspondence analysis of VLM backbones including Qwen2.5-VL-7B [2], LLaVA-NeXT-Video-7B [59], revealing that our GASP framework boosts peak layer-wise correspondence matching accuracy from very low values (below 5%) to over 70% and maintains over 85% temporal robustness, while baselines remain under 5%.
- We demonstrate that our geometrically grounded model, trained **without any 3D VQA data**, improves internal visual correspondence with strong temporal robustness and achieves substantial gains over baselines on downstream spatial reasoning benchmarks, with only minor changes in general video QA performance.

Our findings suggest that learning from visual correspondence is a promising and generalizable path towards VLMs with more reliable 3D spatial reasoning.

2. Related Works

3D-Aware VLMs. Recent efforts have focused on enabling MLLMs to understand 3D scenes [4, 10, 13–15, 19, 35, 46, 53, 62]. A dominant approach processes explicit 3D data, such as point cloud features [4] or pre-segmented 3D objects [15, 46]. Another strategy projects multi-view images [50] into explicit spatial representations, like voxel space [62] or BEV maps [35]. Other work uses dual-encoder architectures or grounding agents to fuse 3D geometry features with 2D semantic features [8, 52, 61, 64]. A common thread is their reliance on explicit 3D data, which

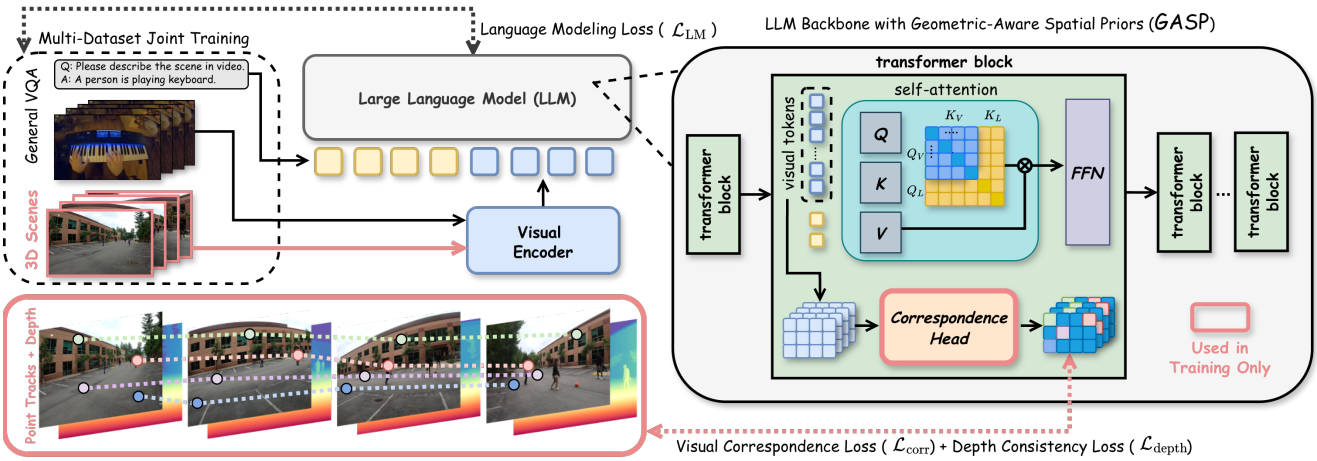

Figure 2. **Injecting the Geometric-Aware Spatial Priors (GASP) into VLMs.** Standard approaches rely on fine-tuning with 3D VQA datasets, which may encourage memorizing dataset-specific biases. We instead insert a small correspondence head into the intermediate layers of the LLM backbone. During the training phase, this head is supervised by visual correspondence and depth consistency signals derived from ground-truth point tracks and depth maps. At inference, the head is discarded and the model processes inputs (e.g., VQA) as a standard VLM, without any auxiliary 3D input. Note that the 3D scene example shown is from EgoHumans [20] for illustration; our training data is sourced from DL3DV [26].

poses a significant alignment challenge, as the LLM must integrate a new, rigid feature stream. In contrast, our work proposes a more lightweight alternative, avoiding explicit 3D data inputs and dual-encoder fusion. We instead inject geometric priors directly into the intermediate layers of the existing LLM backbone to find 3D consistency within its own representations.

Spatial Reasoning in VLMs. VLMs face significant challenges in complex spatial reasoning [3, 6, 8, 33, 39, 45, 47, 48, 61, 63]. Catalyzed by benchmarks like VSI-Bench [54], a dominant paradigm emerged: creating large-scale, 3D-related VQA datasets [8, 33, 39, 45, 58] to fuel specialized models [8, 47, 48, 61] via fine-tuning. This reliance may encourage VLMs to learn superficial correlations and memorize dataset-specific biases, leading to poor generalization. In contrast, our work departs from this VQA-based supervision, instead injecting fundamental geometric priors—correspondence and depth consistency—directly into the VLM's internal representations.

3. Preliminaries: Self-Attention in VLMs

Modern VLMs process a sequence of visual tokens, $V \in \mathbb{R}^{N\times d}$, and language tokens, $L \in \mathbb{R}^{M\times d}$, by concatenating them into a unified input sequence $X = \text{Concat}(V, L) \in \mathbb{R}^{(N+M)\times d}$, which is fed into the LLM backbone. Within each transformer layer, this sequence X is projected into queries Q, keys K, and values V. The core scaled dot-product attention mechanism computes an output Z:

$$Z = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Here, $Q, K, V \in \mathbb{R}^{(N+M)\times d_k}$, and QK^T is the similarity matrix representing scores between all token pairs. To analyze spatial reasoning, we partition the query and key matrices based on their origin:

$$Q = \text{Concat}(Q_V, Q_L), \quad K = \text{Concat}(K_V, K_L) \quad (2)$$

where Q_V, K_V are projections of visual tokens and Q_L, K_L are projections of language tokens. Consequently, the attention similarity matrix $S = QK^T$ deconstructs into four distinct quadrants:

$$S = QK^T = \begin{pmatrix} Q_V \\ Q_L \end{pmatrix} \begin{pmatrix} K_V^T & K_L^T \end{pmatrix} = \begin{pmatrix} Q_V K_V^T & Q_V K_L^T \\ Q_L K_V^T & Q_L K_L^T \end{pmatrix} \quad (3)$$

These quadrants represent visual self-attention ($Q_V K_V^T$), language self-attention, and cross-modal attention. We are primarily interested in the **visual self-attention** matrix, $Q_V K_V^T$, as analyzing this QK-matching provides a direct window into the model's learned spatio-temporal correspondence which is most relevant to geometric reasoning.

To this end, we pose a direct hypothesis: *genuine high-level spatial understanding in VLMs can be unlocked by explicitly learning their internal visual self-attention representations ($Q_V K_V^T$) to be geometrically consistent.* This mirrors findings in video diffusion models, where QK-matching is a key metric for temporal consistency [18, 31]. Therefore, we posit that by explicitly training the $Q_V K_V^T$ representations to be geometrically aware, we can inject a robust inductive bias that is essential for high-level spatial understanding.

4. Learning Geometric Correspondence

Building on our hypothesis, we posit this $Q_V K_V^T$ deficiency does not stem from the visual encoder alone, but from the core LLM, which lacks a robust 3D geometric inductive bias from its pre-training with the web-scale text corpora that lack fine-grained 3D geometric information. We argue that 3D VQA fine-tuning encourages memorizing superficial correlations rather than learning geometric principles, leading to poor generalization (Figure 3). To address this, we depart from QA-based supervision and instead directly inject a geometric inductive bias into the LLM transformer layers. Our core idea is to teach the model object permanence by supervising its internal visual representations, using a correspondence head trained with both ground-truth point correspondence and depth supervision.

4.1. View-Invariant Visual Correspondence

We augment a standard VLM, denoted by the function Φ, by attaching a lightweight correspondence head, $\mathcal{H}_c$, to the output of an intermediate LLM transformer block at layer l. This head takes as input the sequence of visual tokens from that layer, $V^{(l)} = \{\mathbf{v}_i^{(l)}\}_{i=1}^N \in \mathbb{R}^{N\times d}$. The correspondence head, implemented as a 2-layers MLP, projects these general-purpose features into a new, lower-dimensional embedding space specifically optimized for correspondence matching:

$$\mathbf{E} = \mathcal{H}_c(V^{(l)}) \tag{4}$$

where $\mathbf{E} = \{\mathbf{e}_i\}_{i=1}^N \in \mathbb{R}^{N\times d_{emb}}$ is the set of correspondence-aware embeddings. This design minimally alters the base VLM architecture while enabling direct supervision of its internal geometric understanding.

We leverage ground-truth point correspondences [26] as our supervisory signal. For an anchor point $\mathbf{p}_i^a$ in a source frame a, its corresponding point $\mathbf{p}_i^b$ in a target frame b provides the positive sample. All other points $\{\mathbf{p}_k^b\}_{k\neq i}$ in the target frame form the negative set. We employ the InfoNCE contrastive loss [21] to train the correspondence head. Following standard practice, we use $\langle \mathbf{u}, \mathbf{v}\rangle$ to denote the dot product between two L2-normalized embeddings (i.e., their cosine similarity). The loss for a single anchor embedding $\mathbf{e}_i^a$ is defined as:

$$\mathcal{L}_i = -\log \frac{\exp(\langle \mathbf{e}_i^a, \mathbf{e}_i^b\rangle/\tau)}{\exp(\langle \mathbf{e}_i^a, \mathbf{e}_i^b\rangle/\tau) + \sum_{k\neq i}\exp(\langle \mathbf{e}_i^a, \mathbf{e}_k^b\rangle/\tau)} \tag{5}$$

where τ is temperature hyperparameter. The full correspondence loss, $\mathcal{L}_{\text{corr}}$, is the average over all anchor points.

4.2. Depth-Aware 3D Consistency

Beyond 2D visual correspondence, we incorporate 3D geometric supervision. Our objective is not to train a high-fidelity depth prediction head [43, 44], but to learn robust *depth consistency* across frames. We therefore do not regress depth values directly; instead, we use depth as a supervisory signal to align geometrically valid correspondences and enforce 3D consistency.

Concretely, for each anchor point $\mathbf{p}_i^a$ in frame a, we derive a soft matching distribution $\mathbf{A}_{ij}$ over candidate patches in frame b by normalizing the similarity scores from the contrastive loss:

$$\mathbf{A}_{ij} = \frac{\exp(\langle \mathbf{e}_i^a, \mathbf{e}_j^b\rangle/\tau)}{\sum_{k=1}^{N_{\text{cand}}}\exp(\langle \mathbf{e}_i^a, \mathbf{e}_k^b\rangle/\tau)} \tag{6}$$

where $\mathbf{A}_{ij}$ represents the model's belief that anchor point i corresponds to candidate patch j, and N_{cand} denotes the total number of candidate patches. Note that we directly reuse the similarity computations from Equation 5 to ensure computational efficiency.

Using these soft matching weights, we compute the *expected depth* for the anchor point in the target frame as a weighted average over all candidate patches:

$$\hat{d}_i^b = \sum_{j=1}^{N_{\text{cand}}} \mathbf{A}_{ij}\cdot d_j^b \tag{7}$$

where d_j^b is the depth value at candidate patch j in frame b. Note that this weighted summation is a standard *Soft-Argmax* formulation that computes the expected depth $\mathbb{E}[d]$ under the matching distribution. This makes the index selection differentiable. To obtain robust depth estimates, we apply average pooling over the spatial region corresponding to each patch in the depth map.

The depth consistency loss then measures the discrepancy between this expected depth and the ground-truth depth at the corresponding point in frame b:

$$\mathcal{L}_{\text{depth}} = \frac{1}{N_{\text{valid}}}\sum_{i\in\text{valid}} \frac{\left|d_i^b - \hat{d}_i^b\right|}{d_i^b + \hat{d}_i^b + \epsilon} \tag{8}$$

where d_i^b is the ground-truth depth of point i at its corresponding location $\mathbf{p}_i^b$ in frame b (obtained from the point correspondence annotation), and the summation is over points with sufficient visibility and confidence scores. The relative formulation makes the loss scale-invariant to enable it to handle scenes with varying depth ranges without requiring per-scene normalization.

The gradient from this loss flows back through the soft matching weights $\mathbf{A}$ to the correspondence embeddings $\mathbf{E}$. $\mathcal{L}_{depth}$ acts as a **discriminative regularizer** rather than a depth estimator: This provides a crucial inductive bias: visually similar patches that reside at different depths in the 3D scene (e.g., a foreground object and a background wall) are forced to have *lower* feature similarity, as they are not valid correspondences. This geometric regularization complements the contrastive loss, resolving ambiguities in scenarios with repetitive textures or foreground-background confusion.

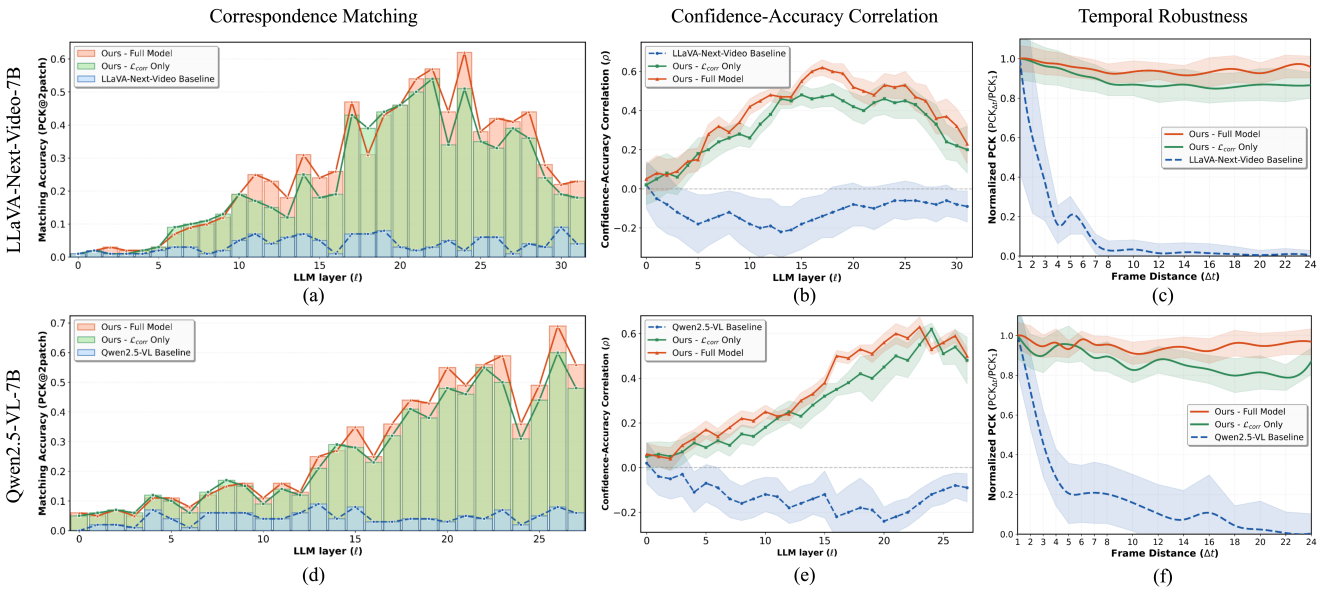

Figure 3. **Analysis of visual correspondence learning.** On LLaVA-NeXT-Video-7B (**top row**) and Qwen2.5-VL-7B (**bottom row**). We compare **(a, d)** layer-wise correspondence matching accuracy (PCK), **(b, e)** confidence-accuracy correlation (ρ), and **(c, f)** temporal robustness for our proposed GASP models against the baselines. Shaded regions indicate standard deviation across 200 test sequences.

Our final training objective combines the LLM loss $\mathcal{L}_{\text{LM}}$ with these dual geometric supervision signals:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda_c \mathcal{L}_{\text{corr}} + \lambda_d \mathcal{L}_{\text{depth}} \tag{9}$$

where λ_c and λ_d are weighting coefficients. This multi-task formulation enables the VLM to jointly optimize for language, 2D correspondence, and 3D depth consistency. By explicitly injecting these complementary geometric priors, we teach the model to develop geometrically-grounded visual representations without relying on 3D VQA datasets.

5. Experiments

In this section, we detail our experiments including implementation specifics, training dataset, correspondence analysis, and compares our method to state-of-the-art approaches across multiple spatial reasoning benchmarks.

5.1. Implementation Details

Our model is initialized from the pre-trained Qwen2.5-VL-7B [2] and LLaVA-NeXT-Video-7B [59]. We attach our correspondence head, $\mathcal{H}_c$, to all 28 or 32 layers of their LLM backbones, initializing its weights from the pre-trained query projection weights via SVD. The entire model is then fine-tuned with a LoRA rank of 512. We train using the AdamW optimizer with a cosine learning rate schedule (peak `1e-4`) and a 4x higher differential rate for the $\mathcal{H}_c$ head's contrastive loss. For stability and efficiency, we use a gradient norm clipping of 1.0, `bfloat16` mixed-precision, and gradient checkpointing. For our contrastive loss, we adopt negative patches from all frames except the anchor patch to maximize diversity. Training requires approximately 10 hours on 32 H200 GPUs.

5.2. Training Datasets

Our model was trained using DL3DV [26] and LLaVA-Video-178K [60], to inject geometric awareness while preserving foundational language capabilities. Geometric supervision comes from a large-scale point correspondence dataset curated from the VGGT [43] training collection. To generate diverse sequences with rich motion parallax, we first sample an anchor frame index t_a from a video $\mathcal{V} = \{I_t\}_{t=1}^{T_{max}}$. Subsequently, a full sequence of F frames is constructed by sampling the remaining $F-1$ frame indices, $\{t_k\}_{k=2}^{F}$, uniformly from a local temporal window $[t_a - R, t_a + R]$ around the anchor. The sequence length F is randomly chosen from 8 to 24, and the window radius R is set to 48. This strategy results in ≈1.75M sequences. We generate ground-truth correspondences on both coarse (8×8) and fine (24×24) grids for each sequence. To prevent catastrophic forgetting, we interleave this geometric data with the LLaVA-Video-178K instruction tuning dataset for joint training.

5.3. Visual Correspondence Evaluation

To validate our core hypothesis – *the baseline VLMs fail due to the lack of a strong internal geometric representation*, we conduct a detailed internal analysis. We first move beyond downstream VQA scores to evaluate the model's internal representations along three critical dimensions: (1) layer-

wise correspondence matching, (2) confidence-accuracy correlation, and (3) temporal robustness. These analysis compare our GASP - full model ($\mathcal{L}_{corr}$ + $\mathcal{L}_{depth}$) and GASP - correspondence-only ($\mathcal{L}_{corr}$) against pre-trained baselines: Qwen2.5-VL-7B and LLaVA-NeXT-Video-7B. Results are summarized in Figure 3.

Evaluation Setup and Metrics. We curate a held-out test set by randomly sampling 200 video sequences from DL3DV [26], explicitly excluded from training. Each sequence is annotated with dense ground-truth point correspondences on 8×8 grids across 8 frames. We design three evaluation metrics as follows:

1) Layer-wise Correspondence Matching. Inspired from DiffTrack [31], we evaluate matching precision using the percentage of correct keypoints (PCK) metric. We extract motion tracks from the internal similarity matrices (Section 3). From a given LLM layer l, let $\mathbf{F}_Q^1 = \{\mathbf{f}_{i,Q}^1\}_{i=1}^{HW} \in \mathbb{R}^{HW \times d}$ be the flattened query descriptors from frame 1, and $\mathbf{F}_K^k = \{\mathbf{f}_{j,K}^k\}_{j=1}^{HW} \in \mathbb{R}^{HW \times d}$ be the flattened key descriptors from frame k. We compute the pairwise cosine similarity matrix $\mathbf{S}^{1,k}$:

$$S_{ij}^{1,k} = \frac{\mathbf{f}_{i,Q}^1 \cdot (\mathbf{f}_{j,K}^k)^T}{\|\mathbf{f}_{i,Q}^1\|_2 \|\mathbf{f}_{j,K}^k\|_2} \tag{10}$$

Given N_{pts} query points $\{\mathbf{p}_i^1\}_{i=1}^{N_{pts}}$, we find their corresponding locations $\{\mathbf{p}_i^k\}$ in frame k by applying an `argmax` operation over the similarity matrix:

$$\mathbf{p}_i^k = \underset{\mathbf{p} \in \Omega_k}{\text{argmax}} \left(S^{1,k}(\mathbf{p}_i^1, \mathbf{p}) \right) \tag{11}$$

where $\mathbf{p}_i^1$ are the query coordinates and Ω_k is the feature grid's spatial domain. The full predicted track $\mathcal{T}_i$ is constructed by concatenating and upscaling these positions:

$$\mathcal{T}_i = \text{Interpolate}\left(\text{Concat}(\mathbf{p}_i^1, \mathbf{p}_i^2, \ldots, \mathbf{p}_i^F)\right) \tag{12}$$

A predicted point $\mathbf{p}_i^k$ is correct if its Euclidean distance to the ground-truth $\mathbf{p}_i^{\text{gt},k}$ is within an error threshold of $\delta = 2$ feature patches. We compute PCK for each LLM layer to identify which layers encode geometric correspondences.

2) Confidence-Accuracy Correlation. We assess whether the model's confidence is *calibrated* with its actual correctness by computing the Pearson correlation coefficient ρ. For a given layer ℓ with N predictions, we correlate the confidence scores $\{c_i\}_{i=1}^N$ (maximum attention probability) with the binary correctness labels $\{y_i\}_{i=1}^N$ ($y_i = 1$ if PCK@2, 0 otherwise):

$$\rho_\ell = \frac{\sum_{i=1}^N (c_i - \bar{c})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (c_i - \bar{c})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}. \tag{13}$$

A positive correlation ($\rho > 0$) indicates well-calibrated predictions where higher confidence corresponds to higher accuracy. Negative correlation ($\rho < 0$) is a statistical signature of systematic miscalibration, providing a quantitative diagnosis of positional bias where the model confidently predicts incorrect matches.

3) Temporal Robustness. To measure robustness across temporal gaps, we evaluate PCK at varying frame distances $\Delta t \in \{1, 2, \ldots, 24\}$, where Δt denotes the temporal offset between matched frames. We plot normalized performance: $Y(\Delta t) = \text{PCK}(\Delta t)/\text{PCK}(\Delta t = 1)$. This normalization anchors all models to 1.0 at the shortest distance, enabling fair comparison of degradation rates.

Comparative Analysis. We summarized our findings of three metrics in analysis (Figure 3) as follows:

(i) Layer-wise Matching (Figure 3a, d): Baseline models (blue) achieve near-zero PCK, confirming their lack of geometric awareness. Our methods (green, red) significantly improve matching accuracy globally, peaking in middle-to-deep layers (20–25 for LLaVA, 25–28 for Qwen2.5-VL). The GASP full model (red) consistently outperforms the correspondence-only model (green), validating the effectiveness of our depth consistency supervision.

(ii) Confidence-Accuracy Correlation (Figure 3b, e): This analysis diagnoses *why* baselines fail. Baselines exhibit a negative correlation ($\rho \approx -0.22$), a statistical signature of positional bias where higher confidence predicts *incorrect* matches. In stark contrast, our full model achieves strong positive correlation ($\rho \approx +0.62$), demonstrating a well-calibrated model that learns genuine geometric reasoning.

(iii) Temporal Robustness (Figure 3c, f): The baseline's matching ability (blue) collapses, retaining $< 5\%$ of its performance beyond an 8-frame gap. In contrast, our full model (red) exhibits graceful degradation, maintaining over 85% performance even at 24-frame distances, demonstrating it learns temporal-invariant geometric features.

5.4. Spatial Reasoning Benchmarks Evaluation

Our analysis (Section 5.3) confirmed GASP improves *internal* geometric representations. The ultimate goal is translating this to superior *high-level* reasoning. We thus evaluate on downstream spatial VQA benchmarks to assess generalization for complex, language-based spatial tasks.

Evaluation Benchmarks. Our primary evaluation uses benchmarks designed for 3D and multi-view spatial understanding: All-Angles Bench [56] (varying viewpoints), VSI-Bench [54] (object permanence, relational reasoning), and the spatial subset of BLINK [11] (relative depth, multi-view perception). We select these benchmarks as they are designed to isolate 'pure' geometric reasoning (e.g., viewpoint consistency) from the high-level semantic reasoning often confounded in 3D VQA datasets. The evaluation follows its standard protocol [54]. Second, to ensure our specialized training does not cause catastrophic forgetting, we evaluate on a suite of broad VQA benchmarks. This includes CV-Bench [42] (2D/3D tasks), as well as Video-MME [9], TempCompass [29], and NextQA [49] for holis-

Table 1. **Comparison with state-of-the-art VLMs on spatial reasoning benchmarks.** We evaluate models on All-Angles Bench, VSI-Bench, and BLINK. Our GASP framework shows strong performance in spatial relation understanding and relative depth estimation.

Methods	All-Angles Bench			VSI-Bench				BLINK		
	Cam. Pose Est.	Manip.	Rel. Dir.	Obj. Count	Route Plan	Rel. Dir.	Appear. Order	Spa. Rela.	Rel. Depth	Multi-View
General VLMs										
GPT-4o	27.3	41.4	40.9	46.2	31.5	41.3	28.5	76.9	64.5	60.2
Gemini-1.5-Pro	25.0	40.3	29.8	56.2	36.0	46.3	34.6	67.1	50.0	41.3
Cambrain-8B	8.5	30.7	30.4	7.0	29.9	30.9	26.2	74.8	51.6	36.8
InternVL2.5-4B	36.9	40.1	33.2	29.1	30.9	41.4	32.5	83.9	66.9	44.4
InternVL2.5-8B	31.8	43.7	34.1	16.9	28.8	41.1	34.7	89.5	77.4	44.4
InternVL2.5-78B	38.6	42.2	38.6	26.6	31.9	40.3	29.9	93.0	82.3	54.1
Qwen2.5-VL-32B	32.4	49.8	40.9	17.4	34.5	30.4	31.1	86.0	75.0	44.3
Qwen2.5-VL-72B	34.1	45.0	48.3	14.3	28.4	27.6	31.4	88.8	81.5	53.4
LLaVA-Onevision-7B	20.5	42.4	36.4	47.7	29.4	35.2	24.4	83.9	75.0	55.6
LLaVA-Onevision-72B	20.5	47.7	33.8	43.5	32.5	39.9	44.6	78.3	78.2	53.4
3D Spatial Reasoning VLMs (Finetuning on 3D Related VQAs)										
VG-LLM	16.5	30.0	26.9	67.9	32.4	40.7	59.2	84.3	77.2	50.8
AoTD	32.4	37.6	26.7	23.5	28.8	41.4	23.3	61.5	49.2	45.1
VLM-3R	22.7	35.9	30.9	70.2	45.4	80.5	40.1	48.3	47.6	50.1
LLaVA-NeXT-Video-7B										
SFT on LLaVA-Video 178K (Baseline)	22.7	39.9	24.7	23.5	24.7	32.4	11.5	53.1	44.4	42.1
+ SFT on DL3DV VQA dataset	19.8	38.1	28.2	21.4	25.1	31.8	9.2	54.5	44.0	42.5
+ GASP - Correspondence ($\mathcal{L}_{corr}$)	34.7	44.3	26.4	39.8	31.4	30.5	17.6	49.2	46.7	44.4
+ GASP - Full Model ($\mathcal{L}_{corr}$ + $\mathcal{L}_{depth}$)	40.9	43.5	29.8	52.5	32.5	41.2	22.0	47.6	48.4	57.1
Δ *Improvement*	↑18.2	↑3.6	↑5.1	↑29.0	↑7.8	↑8.8	↑10.5	↓5.5	↑4.0	↑15.0
Qwen2.5-VL-7B										
SFT on LLaVA-Video 178K (Baseline)	34.1	41.3	36.9	33.8	26.8	34.3	26.5	80.2	78.9	41.5
+ SFT on DL3DV VQA dataset	31.5	41.5	36.2	33.2	27.1	34.3	25.3	81.0	78.1	42.0
+ GASP - Correspondence ($\mathcal{L}_{corr}$)	50.0	39.3	37.8	39.6	30.9	36.7	34.6	88.1	79.0	54.9
+ GASP - Full Model ($\mathcal{L}_{corr}$ + $\mathcal{L}_{depth}$)	52.8	40.1	37.2	41.6	30.4	40.6	35.0	88.8	80.7	53.4
Δ *Improvement*	↑18.7	↓1.2	↑0.3	↑7.8	↑3.6	↑6.3	↑8.5	↑8.6	↑1.8	↑11.9

Table 2. **Comparison with VLMs on CV-Bench.** We show progressive improvements of our GASP framework built on Qwen2.5-VL-7B. The **best** score is marked in bold in each column.

Methods	CV-Bench				
	Overall	2D-count	2D-relation	3D-depth	3D-distance
SpaceQwen2.5VL-3B	51.4	62.2	45.4	45.4	50.0
Kimi-VL-3B-Thinking	57.5	60.5	79.1	43.5	44.0
Qwen2.5-VL-3B-Instruct	68.5	62.6	70.3	78.0	64.7
InternVL2.5-4B	74.1	68.0	79.9	80.7	69.2
LLaVA-OneVision-7B	73.2	69.2	77.9	81.7	65.0
Qwen2.5-VL-7B-Instruct	76.6	63.7	87.7	85.5	72.7
Cambrain-8B	62.2	60.7	81.7	55.0	50.5
LLaMA-3.2V-11B	58.2	59.0	55.9	67.3	50.5
LLaMA-3.2V-11B-CoT	72.8	59.1	78.9	78.8	78.0
LLaVA-1.5-13B	58.2	56.1	57.2	66.8	53.3
SpaceLLaVA-13B	58.2	56.1	57.2	66.8	53.3
Qwen2.5-VL-32B	79.7	68.9	80.8	86.5	**85.8**
LLaVA-OneVision-72B	79.7	**70.2**	**89.2**	82.5	79.0
Qwen2.5-VL-7B					
+ GASP - Correspondence ($\mathcal{L}_{corr}$)	79.4	68.0	88.1	86.6	78.6
+ GASP - Full Model ($\mathcal{L}_{corr}$ + $\mathcal{L}_{depth}$)	**79.8**	68.2	88.3	**87.3**	79.2

tic video and temporal understanding.

Comparison Baselines. We compare our models against a wide range of state-of-the-art models including (1) *General VLMs* (e.g., GPT-4o, InternVL2.5) [2, 5, 12, 17, 24, 28, 40–42] and (2) *3D Spatial Reasoning VLMs* (e.g., VG-LLM, VLM-3R) [3, 8, 36, 61]. To isolate the specific contribution of our geometric-aware training, our primary comparison is a controlled ablation. We establish our control models by fine-tuning the LLaVA-NeXT-Video-7B [59] and Qwen2.5-VL-7B [2] models on the general-purpose LLaVA-Video 178K dataset [59]. Our models are trained on the exact same LLaVA-Video 178K dataset, but also include our proposed GASP geometric priors ($\mathcal{L}_{corr}$ and $\mathcal{L}_{depth}$). This setup ensures that all observed performance gains are directly attributable to our geometric training paradigm, not just to the effects of continued SFT. Furthermore, to disentangle the effect of data exposure from our training objective, we include a "Fairness Baseline." Using the same DL3DV point tracks as GASP, we construct a VQA dataset where the correspondence task is reformulated as explicit question-answer pairs by following [51, 58] (e.g., *Which labeled point in Image-2 corresponds to the marked location in Image-1?* or *Predict the* $[x, y]$ *coordinates of this point in Image-2.*). This baseline is fine-tuned on the identical data mix (LLaVA-Video 178K + DL3DV VQA) with the same data quantity, isolating whether gains arise from data content or from GASP's geometric objective.

5.5. Benchmark Results

Our GASP framework improves over the baselines on several spatial benchmarks in Table 1, with the largest gains on tasks directly related to the injected geometric priors. This provides evidence that learning from fundamental 3D priors can enhance high-level spatial reasoning. We highlight our findings as follows:

Analysis of 3D Geometric Consistency. As shown in Table 1, our gains are most pronounced on the specific sub-tasks that are directly related to the geometric priors we inject. On the All-Angles Bench, which tests geometric consistency, our full model nearly doubles the Camera Pose Estimation score (e.g., 34.1% to 52.8% for Qwen2.5-

Table 3. **Comparison on generic multimodal benchmarks.**

Methods	Video-MME w/o sub.	Video-MME w/ sub.	Temp Compass$_{MC}$	NextQA
LLaVA-NeXT-Video-7B				
(baseline)	40.8	40.3	50.1	59.4
+ GASP - Correspondence	42.3	41.6	53.7	62.8
+ GASP - Full Model	42.8	41.9	53.8	61.6
Qwen2.5-VL-7B				
(baseline)	60.6	59.3	68.4	76.6
+ GASP - Correspondence	62.6	61.2	71.5	78.4
+ GASP - Full Model	63.2	61.6	70.3	74.7

VL) and improves Relative Direction (e.g., 24.7% to 29.8% for LLaVA-NeXT). This directly validates that our training imbues the model with a robust understanding of viewpoint. Also, the DL3DV VQA baseline even degrades several key metrics compared to the SFT baseline (e.g., Camera Pose 22.7%→19.8%, Object Counting 23.5%→21.4% for LLaVA-NeXT). In contrast, GASP achieves 40.9% and 52.5% on these same metrics. This confirms that the improvements stem from the GASP geometric objective rather than data exposure, consistent with the overfitting patterns of VQA-based methods observed in Appendix.

Analysis of Relational & Temporal Capability. This geometric understanding also improves abstract reasoning. On VSI-Bench, our model shows a dramatic improvement in Object Counting (e.g., 23.5% to 52.5% on VSI-Bench). This suggests that the view-invariant features learned by our method help the model maintain object identity, preventing it from double-counting or missing objects across frames. This multi-view consistency is further validated on BLINK, where our framework provides the boost on the Multi-View task (e.g., 42.1% to 57.1% for LLaVA-NeXT), proving our model's superior ability to correlate information across different perspectives.

Analysis on General-Purpose Benchmarks. A natural concern is whether specialized geometric training harms general VQA capabilities. As presented in Table 3, we observe a modest trade-off: our Qwen2.5-VL-7B model with GASP loses 1.9% on NextQA (76.6% → 74.7%), but improves on temporal video benchmarks, including Video-MME (59.3% → 61.6%) and TempCompass (68.4% → 70.3%). We attribute these temporal gains to improved object permanence: video understanding requires maintaining persistent object identity across viewpoint changes (e.g., *Temporal Reasoning* and *Action Forecasting* in Video-MME), which directly benefits from our view-invariant geometric representations. Also, on the broad CV-Bench (Table 2), our model achieves an Overall score of 79.8%. Overall, GASP trades about 1–2% general VQA accuracy for consistent gains on spatial and temporal benchmarks.

5.6. Ablation Studies

We conduct ablation studies on two key hyperparameters for our GASP framework: the LoRA rank and the correspondence head ($\mathcal{H}_c$) injection strategy, reported in Table 4.

Table 4. **Ablation studies of the LoRA rank effect and correspondence head injection into LLM layers.**

GASP - Full Model ($\mathcal{L}_{corr}$ + $\mathcal{L}_{depth}$)	Avg. PCK	All-Angles Bench	VSI-Bench	BLINK
Impact of LoRA Rank on Performance				
LLaVA-NeXT-Video-7B				
LoRA Rank = 64	8.4	30.1	28.5	44.9
LoRA Rank = 128	13.7	32.6	30.6	45.8
LoRA Rank = 256	17.1	35.8	33.9	47.5
LoRA Rank = 512	26.2	**38.1**	**37.1**	**51.0**
LoRA Rank = 1024	**28.6**	37.2	34.8	48.7
Qwen2.5-VL-7B				
LoRA Rank = 64	18.2	38.5	**37.3**	70.2
LoRA Rank = 128	26.7	**43.4**	36.9	**74.3**
LoRA Rank = 256	28.8	41.8	35.5	73.5
LoRA Rank = 512	31.2	40.2	34.1	72.4
LoRA Rank = 1024	**32.5**	38.9	33.2	73.8
Correspondence Head Injection into LLM Layers				
LLaVA-NeXT-Video-7B				
Layer 10 - 18	21.7	34.8	35.9	47.7
Layer 18 - 25	25.1	37.5	35.2	49.5
Layer 25 - 32	25.8	**39.1**	36.5	49.3
All Layers (1 - 32)	**26.2**	38.1	**37.1**	**51.0**
Qwen2.5-VL-7B				
Layer 10 - 16	19.8	37.9	34.2	68.2
Layer 16 - 22	23.3	38.8	35.5	71.1
Layer 22 - 28	25.2	42.7	**37.4**	72.8
All Layers (1 - 28)	**26.7**	**43.4**	36.9	**74.3**

Impact of LoRA Rank. We analyze performance varying the LoRA rank from 64 to 1024. We find a clear trade-off: internal correspondence (Avg. PCK) generally scales with the rank, but downstream performance peaks earlier (512 for LLaVA, 128 for Qwen). This suggests that a higher Avg. PCK does not always lead to improved complex spatial benchmark reasoning. We hypothesize that with very high LoRA ranks, while it fits better for our geometric priors, may begin to harm foundational language capabilities.

Correspondence Head Injection into LLM Layers. We target LLM layers rather than the visual encoder because spatial reasoning fundamentally requires sequence-level temporal binding. We then ablate *where* in the LLM the correspondence head is injected. While injecting at later layers (e.g., LLaVA 'Layer 25-32'), which contain richer semantic information, outperforms earlier layers, our key finding is that applying supervision across all layers (1-32 for LLaVA, 1-28 for Qwen) yields the best and most consistent performance. This demonstrates that reinforcing geometric consistency across all feature transformation stages is most effective.

6. Conclusion

In this paper, we proposed GASP, a framework that injects fundamental geometric priors directly into the LLM's transformer layers. Our analysis showed GASP corrects VLMs' near-zero internal correspondence accuracy, boosting it to over 70%. These internal improvements generalize, achieving significant gains on downstream spatial benchmarks. We thus find learning from geometric priors is a promising and generalizable path toward spatially-intelligent VLMs.

References

[1] Anthropic. Claude, 2024. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 5, 7

[3] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 3, 7

[4] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26428–26438, 2024. 2

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 7

[6] Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025. 2, 3

[7] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023. 1

[8] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 2, 3, 7

[9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 6

[10] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 2

[11] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 2, 6

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7

[13] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 2

[14] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37: 113991–114017, 2024.

[15] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 2

[16] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022. 1

[17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 7

[18] Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7276–7287, 2025. 3

[19] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024. 2

[20] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An egocentric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19807–19819, 2023. 3

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 4

[22] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1

[23] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and

Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 7

[25] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025. 2

[26] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 2, 3, 4, 5, 6

[27] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 1

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 7

[29] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 6

[30] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Viewpoint rosetta stone: Unlocking unpaired ego-exo videos for view-invariant representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15802–15812, 2025. 2

[31] Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. *arXiv preprint arXiv:2506.17220*, 2025. 3, 6

[32] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. *arXiv preprint arXiv:2406.11815*, 2024. 1

[33] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 2, 3

[34] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bootstrap your own views: Masked ego-exo modeling for fine-grained view-invariant video representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13661–13670, 2025. 2

[35] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025. 2

[36] Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8523–8533, 2025. 7

[37] Chan Hee Song, Jihyung Kil, Tai-Yu Pan, Brian M Sadler, Wei-Lun Chao, and Yu Su. One step at a time: Long-horizon vision-and-language navigation with milestones. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15482–15491, 2022. 1

[38] Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav Sukhatme. Embodied bert: A transformer model for embodied, language-guided visual task completion. *arXiv preprint arXiv:2108.04927*, 2021. 1

[39] Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, et al. Spacevista: All-scale visual spatial reasoning from mm to km. *arXiv preprint arXiv:2510.09606*, 2025. 2, 3

[40] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 7

[41] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[42] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 6, 7

[43] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 4, 5

[44] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 4

[45] Wentao Wang, Heqing Zou, Tianze Luo, Rui Huang, Yutian Zhao, Zhuochen Wang, Hansheng Zhang, Chengwei Qin, Yan Wang, Lin Zhao, et al. Video-str: Reinforcing mllms in video spatio-temporal reasoning with relation graph. *arXiv preprint arXiv:2510.10976*, 2025. 2, 3

[46] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023. 2

[47] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 2, 3

[48] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025. 2, 3

[49] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining

temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 6

[50] Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024. 2

[51] Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025. 7

[52] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701. IEEE, 2024. 2

[53] Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29501–29512, 2025. 2

[54] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 2, 3, 6

[55] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 2

[56] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025. 2, 6

[57] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1

[58] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025. 2, 3, 7

[59] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 2, 5, 7

[60] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 5

[61] Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025. 2, 3, 7

[62] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 2

[63] Fangrui Zhu, Hanhui Wang, Yiming Xie, Jing Gu, Tianye Ding, Jianwei Yang, and Huaizu Jiang. Struct2d: A perception-guided framework for spatial reasoning in mllms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 3

[64] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 2