OmniStream: Mastering Perception, Reconstruction and Action in Continuous Streams

Yibin Yan[1,2*], Jilan Xu[3*], Shangzhe Di[1], Haoning Wu[1], and Weidi Xie[1]

[1] School of Artificial Intelligence, Shanghai Jiaotong University
[2] Shanghai Innovation Institute
[3] Visual Geometry Group, University of Oxford
https://go2heart.github.io/omnistream/

Abstract. Autonomous agents require representations that are general, causal, and physically structured to operate in real-time streaming environments. However, current vision foundation models remain fragmented, specializing narrowly in image semantic perception, offline temporal modeling, or spatial geometry. This paper introduces **OmniStream**, a unified streaming visual backbone that effectively perceives, reconstructs, and acts from diverse visual inputs. By incorporating causal spatiotemporal attention and 3D rotary positional embeddings (3D-RoPE), our model supports efficient, frame-by-frame online processing of video streams via a persistent KV-cache. We pre-train OmniStream using a synergistic multi-task framework coupling static and temporal representation learning, streaming geometric reconstruction, and vision-language alignment on 29 datasets. Extensive evaluations show that, even with a strictly frozen backbone, OmniStream achieves consistently competitive performance with specialized experts across image and video probing, streaming geometric reconstruction, complex video and spatial reasoning, and robotic manipulation. Rather than pursuing benchmark-specific dominance, our work demonstrates the viability of training a single, versatile vision backbone that generalizes across semantic, spatial, and temporal reasoning, *i.e.*, a meaningful step towards general-purpose visual understanding for interactive and embodied agents.

Keywords: Visual Representation· 3D Reconstruction · Embodied AI

1 Introduction

Modern visual agents are increasingly deployed in *streaming* settings: a robot observes the world through a camera, an assistant watches a video, or an AR device processes an egocentric feed. In such cases, the agent must update its beliefs online from a continuous stream, under tight latency and memory budgets. A useful representation must therefore be (i) *general*, supporting recognition, reasoning, and interaction; (ii) *causal*, depending only on past and present frames; and (iii) *structured*, capturing not only appearance but also geometry and motion.

* Equal Contribution.

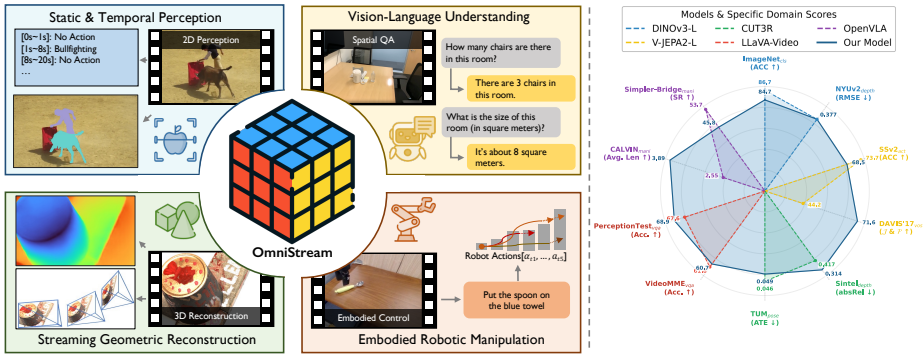

Fig. 1: Left: **OmniStream** supports a wide spectrum of tasks, including 2D/3D perception, vision-language understanding, and embodied robotic manipulation. Right: the frozen features of our single backbone achieve highly competitive or superior performance compared to leading domain-specific experts.

In language community, models achieve impressive generalisation with a single autoregressive backbone trained under next-token prediction [1, 16, 91]. This works not merely due to scale, but because the diverse tasks can share a common representation and training interface. While vision, however, remains far less unified. The tasks differ not only in supervision but also in the nature of their outputs: discrete labels [25], segmentation masks [41], dense depth [103], 3D geometry [84], and temporally evolving predictions. As a result, the field has converged on specialized foundation models: strong image encoders (*e.g.*, DINO [58, 74], SigLIP [82, 109]), video models (*e.g.*, V-JEPA [6, 10], VideoMAE [81, 85]), and geometric experts (*e.g.*, DepthAnything [47, 103, 104], VGGT [84]). While effective in-domain, these models typically learn representations tailored to a narrow objective (semantic invariances, temporal dynamics, or spatial geometry), and do not directly yield a single backbone that transfers across *static semantics*, *temporal dynamics*, and *3D structure*, especially in an online, causal regime.

A natural direction is to "unify" vision by converting everything into text-like token generation, as explored by Florence [96, 108], OFA [86], and Unified-IO [53, 54]. These models provide a convenient interface, but the unification still largely happens at the *output* level. In practice, accommodating new tasks often requires expensive re-training, re-tokenization of outputs, or architectural adjustments to the generative head. This leaves an open question that is more representation-centric than interface-centric: Can we learn a *single streaming visual backbone* whose representations are sufficiently universal that many downstream tasks can be solved on top of it, without modifying or fine-tuning the backbone?

In this paper, we propose **OmniStream**, a unified streaming visual backbone that turns a strong pre-trained image model into an online, temporally causal model while preserving its spatial priors. OmniStream is built around two design choices. First, we introduce **causal spatiotemporal attention**, which enforces strict temporal causality and enables efficient frame-by-frame inference via a persistent KV-cache, avoiding re-computation over past frames. Second, building

on video RoPE formulations [92], we adapt 3D rotary positional embeddings (3D-RoPE) to a causal ViT with DINOv3 initialization. This extends DINOv3's 2D spatial RoPE to a consistent spatiotemporal relative encoding while preserving the pretrained spatial priors.

Architecture alone, however, does not guarantee universality: the representation needs to be *trained* to encode what downstream tasks require. To this end, we adopt a unified pre-training recipe that couples three complementary signals, each targeting a different axis of *understanding*: **i) static and temporal representation learning**: a self-supervised student-teacher distillation objective that unifies image representation learning with causal video modeling. By distilling both global (image/video-level) and local (patch-level) features, the backbone learns semantic invariances and motion-sensitive dynamics while respecting causality; **ii) streaming geometric reconstruction**: lightweight feedforward dual Dense Prediction Transformer (DPT) and camera heads predict depth maps, ray maps, and camera poses from the stream, injecting explicit 3D constraints so that the representation reflects physical scene structure rather than appearance alone; **iii) vision-language alignment**: a lightweight autoregressive language decoder trained on captioning, Optical Character Recognition (OCR), and visual grounding connects visual tokens with linguistic concepts, encouraging fine-grained details and semantic grounding that are useful for reasoning-heavy tasks. Together, these objectives encourage OmniStream to learn representations that are simultaneously temporally coherent, geometrically grounded, and language-aligned, as illustrated in Fig 1.

In summary, we propose **OmniStream**, a unified *streaming* visual backbone that extends a pre-trained image encoder with causal spatiotemporal attention and 3D-RoPE, enabling strictly causal, efficient online inference. Our extensive multi-task pre-training across 29 diverse datasets demonstrate that scaling diverse objectives is highly synergistic rather than merely additive: (i) causal video modeling is essential for capturing dynamic motions; (ii) explicit geometric pre-training is prerequisite for downstream spatial intelligence and embodied control; and (iii) early vision-language alignment within the backbone is critical to prevent catastrophic failures during VLM integration. Lastly, we demonstrate that OmniStream serves as a strong **vision backbone**: its features perform well on standard image/video probing; it natively supports competitive streaming geometric reconstruction; and it enables downstream language models and policy to solve general and complex spatial video question answering, and robotic manipulation without fine-tuning the backbone, matching or surpassing task-specific baselines. The code is available at https://github.com/Go2Heart/OmniStream.

2 Related Work

Vision foundation models. The pursuit of a unified vision foundation model capable of generalizing across diverse perceptual dimensions remains a central objective in computer vision. For static images, vision-language contrastive learning frameworks such as CLIP [66] and SigLIP [82, 109] have demonstrated im-

pressive semantic generalization, while self-supervised paradigms like DINO [20, 58, 74], MAE [31], and I-JEPA [5] excel in extracting low-level visual features. In the video domain, research has pivoted toward capturing complex spatiotemporal dependencies. Supervised models [3, 12, 98, 100] leverage large-scale video-text corpora, whereas self-supervised approaches focus on reconstructing masked spatiotemporal patches at the pixel level [81, 85] or feature level [6, 10]. Despite these advancements, the landscape remains fragmented: most models are either restricted to static perception or rely on non-causal, offline temporal processing. Our OmniStream bridges this gap by unifying image and video representation learning within a strictly causal, streaming framework.

Feed-forward 3D reconstruction models. DUSt3R [88] has led a paradigm shift in multi-view geometry, moving from traditional optimization-based methods like Structure-from-Motion (SfM) to learning-based feed-forward neural networks. Building upon this, subsequent works have extended this paradigm to support multi-view and video inputs beyond simple image pairs [47, 84, 90, 101, 110]. Parallel to these developments, the research focus has transitioned from offline sequence reconstruction to the more challenging *online* setting [42, 83, 87, 94, 118], which requires real-time geometric reasoning. Although these specialized 3D experts demonstrate remarkable precision in geometric tasks, they often lack high-level semantic reasoning capabilities. Our method not only achieves competitive performance in streaming geometric estimation but also exhibits superior generalization across vision tasks that require deep semantic understanding, effectively bridging the gap between spatial structure and semantic abstraction.

Visual representation for VLMs and VLAs. Advanced Vision-Language Models (VLMs) [7, 9, 43, 49, 50, 112] rely on robust visual encoders to ground textual reasoning. While contrastive models like CLIP [66] and SigLIP [109] serve as the *de facto* standard for global semantics, they typically fall short in fine-grained visual perceptions [51, 80]. Recent studies have sought to enhance visual representations by adopting more powerful backbones [2, 8, 82], leveraging ensembles of multiple specialized encoders [36, 52, 72, 79], or incorporating specialized geometric experts to bolster spatial understanding [26, 113]. Building on general VLMs, Vision-Language-Action (VLA) models [14, 39, 119] demand both high-level semantics and low-level geometric precision. However, a significant gap exists between current VLMs and the demands of VLA embodied tasks [44], especially in visual representation [71, 111]. In contrast, our OmniStream bridges this gap by unifying semantic, dynamic and geometric representations into a single, efficient streaming backbone.

3 Method

This section presents **OmniStream**, a unified *streaming* visual backbone. We formalize causal streaming perception in Sec. 3.1, describe the backbone that turns a pre-trained image ViT into an efficient online encoder in Sec. 3.2, introduce our multi-task training framework in Sec. 3.3, and detail how the learned representation supports unified perception, reasoning, and action in Sec. 3.4.

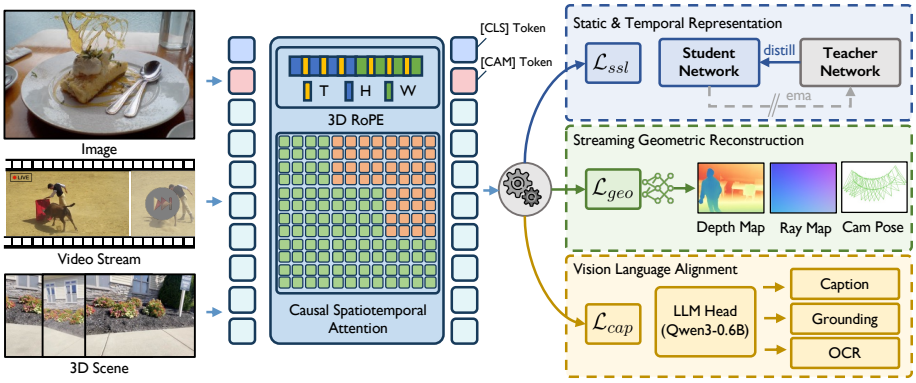

Fig. 2: Overall framework of OmniStream. Equipped with 3D-RoPE and Causal Spatiotemporal Attention, our unified backbone is trained via a multi-task framework that couples static and temporal representation learning, streaming geometric reconstruction, and vision-language alignment.

3.1 Problem Formulation

We aim to train a **unified streaming visual backbone** that produces a general-purpose representation at each time step of a video stream and can be reused by diverse downstream tasks with minimal task-specific modeling.

Formally, let a continuous video stream be $\mathcal{V}^T = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_T\}$, where $\mathcal{I}_t \in \mathbb{R}^{H \times W \times 3}$. At time step t, the backbone processes the current frame and historical context to produce a composite output state $\mathbf{O}_t$:

$$\mathbf{O}_t = f_\theta(\mathcal{I}_t \mid \mathcal{I}_1, \ldots, \mathcal{I}_{t-1}), \quad t \in \{1, \ldots, T\}. \tag{1}$$

To support real-time deployment, we enforce a **strict causality constraint**: $\mathbf{O}_t$ must not depend on future frames $\{\mathcal{I}_{t+1}, \ldots, \mathcal{I}_T\}$. The same backbone is expected to support multiple regimes:

i) **static perception** ($T = 1$): the stream degenerates to a single image; the model acts as an image encoder capturing semantics and layout;

ii) **dynamic understanding** ($T > 1$): the backbone captures temporal evolution and motion cues from a continuous stream;

iii) **geometric reasoning** ($T > 1$): by aggregating multi-frame evidence causally, the backbone supports online reconstruction of 3D/4D scene structure;

iv) **embodied control** ($T \geq 1$): the backbone supplies real-time, action-oriented representation to drive closed-loop policies for robotic manipulation.

3.2 Streaming Visual Backbone

Overview. We instantiate **OmniStream** as a vision transformer from DINOv3 [74], leveraging its strong spatial representation as a starting point. As depicted in Fig. 2, to evolve an image encoder into a **unified streaming** backbone, we apply two modifications: (i) **causal spatiotemporal attention** to

enable online inference with temporal causality and a persistent KV-cache; (ii) **3D rotary positional embeddings (3D-RoPE)**, extending the 2D positional priors to the spatiotemporal domain. These changes allow efficient processing of long streams while preserving the pre-trained spatial inductive bias.

Tokenization and outputs. For a video stream $\mathcal{V}^T \in \mathbb{R}^{T\times H\times W\times 3}$, we split each frame into non-overlapping $p \times p$ patches, producing $h \times w$ patches per frame with $h = H/p$ and $w = W/p$. After linear projection, we prepend per-frame special tokens: one global `[CLS]` token, four register tokens [24], and an optional `[CAM]` token used by the camera head. Let N_s be the number of special tokens per frame; the input sequence is $\mathbf{z}^0 \in \mathbb{R}^{T\times(N_s+hw)\times d}$.

After multiple Transformer blocks, we extract (i) dense spatiotemporal feature maps from selected intermediate layers $\mathbf{L}$, and (ii) final-layer special tokens for global semantics and camera prediction:

$$\mathbf{O} = \{\mathcal{Z}, \mathbf{z}_{\text{cls}}, \mathbf{z}_{\text{cam}}\} = f_\theta\left(\mathbf{z}^0\right), \tag{2}$$

where $\mathcal{Z}$ denotes selected layers' dense patch feature $\left\{\mathbf{z}^{(\ell)} \in \mathbb{R}^{T\times h\times w\times d}\right\}_{\ell\in\mathbf{L}}$, the last-layer `[CLS]` token $\mathbf{z}_{\text{cls}} \in \mathbb{R}^{T\times d}$ provides an image/video-level summary, and the `[CAM]` token $\mathbf{z}_{\text{cam}} \in \mathbb{R}^{T\times d}$ is routed to the camera head for pose estimation.

Causal spatiotemporal attention. A naive Transformer that attends over all frames violates causality and is inefficient for streaming. We therefore apply spatiotemporal self-attention with a **causal temporal mask**, so tokens at time t may only attend to tokens from times $\leq t$.

Let the total number of tokens be $N_{\text{total}} = T \cdot (N_s + hw)$, the attention mask as $\mathbf{M} \in \{0, -\infty\}^{N_{\text{total}}\times N_{\text{total}}}$. The masked attention is formalized as:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{head}}} + \mathbf{M}\right)\mathbf{V}, \tag{3}$$

For a query token with global index u and a key token with global index v, the mask is defined via a frame-index mapping $\tau(\cdot)$:

$$M_{u,v} = \begin{cases} 0, & \text{if } \tau(u) \geq \tau(v) \\ -\infty, & \text{if } \tau(u) < \tau(v) \end{cases}, \tag{4}$$

where $\tau(u)$ returns the time step of token u. This ensures every patch in frame t can attend to all patches in frames $\leq t$, but never to future frames.

Streaming inference with KV-cache. Under the above causality, inference can be performed incrementally: when frame t arrives, we compute its queries and reuse cached keys/values from frames $\leq t-1$. This yields online processing without recomputing attention over the full past stream, while producing the same output as full causal attention.

3D rotary positional embeddings (3D-RoPE). The original DINOv3 [74] uses 2D RoPE to encode relative spatial offsets. We extend RoPE to the spatiotemporal domain by re-purposing the per-head feature dimension d_{head} while

preserving as much pre-trained spatial structure as possible. Concretely, we adopt an allocation strategy with a **2:3:3** split across (t, y, x), namely, time, height, and width. We interleave temporal components into the original 2D RoPE such that indices $i \equiv 3 \pmod 4$ encode time t [8], while the remaining indices follow the original DINOv3 pattern for y and x. We further apply RoPE-box jittering [74] to improve robustness of the positional embeddings.

3.3 Unified Multi-task Learning Objectives

A unified streaming backbone should ideally produce representations that transfer seamlessly across semantics, dynamics, geometry, and language. We therefore train OmniStream with a unified multi-task framework that combines three complementary objectives: **static & temporal representation learning**, **streaming geometric reconstruction**, and **vision-language alignment**. These signals are designed to be compatible with causal streaming environments and to encourage representations that are simultaneously discriminative, temporally coherent, and physically grounded.

Static & temporal representation learning. We treat an image as a degenerate stream with $T = 1$, enabling a single objective to cover both images and videos. We employ a DINOv3-style student-teacher distillation objective [74]:

$$\mathcal{L}_{\text{ssl}} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + 0.1 \times \mathcal{L}_{\text{KoLeo}} + \mathcal{L}_{\text{gram}}, \tag{5}$$

Given a set of views (global/local crops from an image or a video clip), the four terms encourage (i) global semantic consistency ($\mathcal{L}_{\text{DINO}}$) [20], (ii) patch-level discriminative features ($\mathcal{L}_{\text{iBOT}}$) [116], (iii) a well-spread feature space ($\mathcal{L}_{\text{KoLeo}}$) [70], and (iv) a consistent patch-level feature during training ($\mathcal{L}_{\text{gram}}$) [74]. We provide full formulations in the supplementary material.

Streaming geometric reconstruction. To inject explicit 3D constraints, we attach feed-forward geometric heads that reconstruct monocular streams into a coherent dynamic representation. We use: (i) a depth head implemented as a dual-DPT module [47] to predict dense depth maps and ray maps; and (ii) a lightweight MLP camera head to predict camera poses.

Using the selected backbone features $\{\mathbf{z}^{(\ell)} \in \mathbb{R}^{T\times h\times w\times d}\}_{\ell\in\mathbf{L}}$, the depth head predicts depth maps $\hat{D} \in \mathbb{R}^{T\times H\times W\times 1}$ and ray maps $\hat{R} \in \mathbb{R}^{T\times H\times W\times 6}$. The camera head consumes camera tokens $\mathbf{z}_{\text{cam}}$ and predicts per-frame pose parameters $\hat{g} \in \mathbb{R}^{T\times 9}$, which is the concatenation of the rotation quaternion $\mathbf{q} \in \mathbb{R}^{T\times 4}$, the translation vector $\mathbf{t} \in \mathbb{R}^{T\times 3}$, and the field of view $\mathbf{f} \in \mathbb{R}^{T\times 2}$, with the camera's principal point assumed to be at the image center. Following [84], we supervise these outputs against normalized ground-truth depth maps D, ray maps R, point maps P, and camera poses g and perform a multi-task geometric loss:

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{ray}} + \mathcal{L}_{\text{points}} + \mathcal{L}_{\text{camera}}, \tag{6}$$

Depth with confidence. We use an L1 regression loss with an additional gradient term and a learned confidence c_t:

$$\mathcal{L}_{\text{depth}} = \sum_{t=1}^{T} \left(\left\| c_t(\hat{D}_t - D_t) \right\|_1 + \left\| c_t(\nabla \hat{D}_t - \nabla D_t) \right\|_1 - \alpha \log c_t \right), \tag{7}$$

where ∇ is the spatial gradient operator and α is a hyper-parameter.

Rays, points, and camera. Similarly, we apply standard L1 regression to ray maps, point maps, and camera poses:

$$\mathcal{L}_{\text{ray}} = \sum_{t=1}^{T} \left\| \hat{R}_t - R_t \right\|_1, \quad \mathcal{L}_{\text{points}} = \sum_{t=1}^{T} \left\| \hat{P}_t - P_t \right\|_1, \quad \mathcal{L}_{\text{camera}} = \sum_{t=1}^{T} \|\hat{g}_t - g_t\|_1 \tag{8}$$

where the camera ray is defined by concatenating the ray origin $\mathbf{o} \in \mathbb{R}^3$ and direction $\mathbf{d} \in \mathbb{R}^3$. And point maps can thus be derived as $\hat{P}_t = \hat{\mathbf{o}}_t + \hat{D}_t \odot \hat{\mathbf{d}}_t$.

Vision-language alignment. To align visual tokens with linguistic concepts and enable reasoning-oriented downstream tasks, we attach an MLP projector and lightweight autoregressive language decoder (Qwen3-0.6B [99]) to the vision backbone. This branch is trained on a mixture of vision-language tasks including dense captioning, OCR, and object grounding.

Given the visual tokens from the last layer $\mathbf{z}^L$ and an instruction prompt $\mathbf{x}_{\text{inst}}$, we train with the standard language modeling objective:

$$\mathcal{L}_{\text{cap}} = -\sum_{n=1}^{L_{\text{text}}} \log P_{\text{text}}\left(y_n \mid \mathbf{z}^L, \mathbf{x}_{\text{inst}}, \mathbf{y}_{<n}\right), \tag{9}$$

where P_{text} is the token distribution predicted by the decoder and $\mathbf{y}_{<n}$ denotes previously generated tokens. During training, gradients are back-propagated through the language decoder into the vision backbone, injecting fine-grained semantic and spatial supervision, which proves to be essential capabilities for VLM and VLA integration in our experiments.

Total objective. The overall training loss is a weighted combination:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ssl}} \cdot \mathcal{L}_{\text{ssl}} + \lambda_{\text{geo}} \cdot \mathcal{L}_{\text{geo}} + \lambda_{\text{cap}} \cdot \mathcal{L}_{\text{cap}}, \tag{10}$$

where we set $\lambda_{\text{ssl}} = 0.1$ and $\lambda_{\text{geo}} = \lambda_{\text{cap}} = 1$ to balance the loss magnitudes.

3.4 Universal Representation for Downstream Applications

A foundation vision backbone should be reusable without expensive full-model adaptation. Accordingly, we evaluate OmniStream in a **strictly frozen** setting, where the backbone parameters are kept fixed and only specific task modules are trained. We organize downstream tasks into three increasing levels of complexity:

Perception: image & video. For discriminative perception tasks, we attach lightweight heads on top of frozen vision features. For static image tasks (*e.g.*,

Table 1: Overview of pre-training datasets. We leverage a diverse collection of images, videos, and 3D/4D data to train our OmniStream.

Task	Pre-training Dataset	Frames
Image SSL	DataComp-100M [28], ImageNet-21K [68]	∼113M
Video SSL	Kinetics [21], SSv2 [30], PE-Videos [15]	∼20M
3D/4D Scenes	Co3Dv2 [67], WildRGBD [95], BlendedMVS [106], MegaDepth [46], MVS-Synth [34], PointOdyssey [114], HyperSim [69], TartanAir [89], DL3DV [48], ScanNet [23], ScanNet++ [107], ArkitScenes [11], Virtual Kitti 2 [19], MapFree [4], Dynamic Replica [37], Spring [57], OmniWorld-Games [117], Aria Synthetic Environments [61], Waymo-Open [77]	∼18M
Captioning	GRIT [18, 63], RefCOCO Series [38], Blip3-OCR [97], SA1B-Caption [22]	∼50M
Total	-	∼200M

segmentation), we apply linear decoders on spatial tokens; for video tasks, we use attentive pooling modules on the frozen spatiotemporal features.

Reasoning: vision-language models (VLM). To enable spatial and temporal reasoning, we integrate OmniStream with an LLM by using an MLP projector that maps frozen visual tokens into the language embedding space. The LLM then generates text conditioned on the projected visual context and textual instructions. Throughout, the vision backbone remains frozen, forcing the LLM to interpret the intrinsic visual representation.

Action: vision-language-action (VLA). For embodied control, we extend the VLM setup by attaching an MLP-based action expert to the LLM outputs to predict robot actions a_t from the frozen visual observations and language instructions. Crucially, because our pre-training objectives ($\mathcal{L}_{\text{ssl}}, \mathcal{L}_{\text{geo}}$) explicitly encode dynamics and geometry, the frozen features provide a strong bridge from perception to action without domain-specific visual fine-tuning.

4 Experiments

4.1 Pre-training Data

We train our model on a wide range of datasets featuring static images, dynamic videos, and geometric 3D/4D Scenes. Totally, our training pipeline consumes approximately 200M frames from 29 datasets, providing a dense and diverse supervisory signal that ensures the model's robustness across various streaming scenarios. A detailed breakdown of the pre-training data is provided in Tab. 1.

4.2 Implementation Details

We initialize our model using the pre-trained weights of DINOv3 [74] ViT-L. We employ a multi-task learning strategy combined with gradient accumulation, where we sequentially interleave all tasks within each training step and perform a single model parameter update after traversing the batches from all tasks. The training is conducted on 64× NVIDIA H200 GPUs and is divided into two stages: the first trained for 60K steps at a resolution of 224×224, followed by a second

Table 2: Holistic evaluation. We compare OmniStream across 5 domains. Here, "-" indicates that a specialized baseline is not natively applicable to or lacks the capability for the given task. Subscripts denote the specific task (*i.e.*, cls: classification, monodepth: monocular depth estimation, seg: semantic segmentation, act: video action recognition, vos: video object segmentation, videodepth: video depth estimation, pose: camera pose estimation, vqa: visual question answering, mani: robotic manipulation). Detailed evaluation protocols are discussed in the supplementary material.

	Benchmark	Ours	DINOv3-L	V-JEPA2-L	CUT3R	LLaVA-Video	OpenVLA
Image	**ImageNet**$_{\text{cls}}\uparrow$	84.7	86.7	-	-	-	-
	NYUv2$_{\text{monodepth}}\downarrow$	0.377	0.377	-	-	-	-
	ADE20K$_{\text{seg}}\uparrow$	49.1	51.5	-	-	-	-
Video	**SSv2**$_{\text{act}}\uparrow$	68.5	54.0	73.7	-	-	-
	K400$_{\text{act}}\uparrow$	85.7	83.6	85.1	-	-	-
	DAVIS'17$_{\text{vos}}\uparrow$	71.6	73.2	44.2	-	-	-
3D Geom.	**Sintel**$_{\text{videodepth}}\downarrow$	0.314	-	-	0.417	-	-
	BONN$_{\text{videodepth}}\downarrow$	0.072	-	-	0.078	-	-
	KITTI$_{\text{videodepth}}\downarrow$	0.136	-	-	0.122	-	-
	Sintel$_{\text{pose}}\downarrow$	0.227	-	-	0.220	-	-
	TUM$_{\text{pose}}\downarrow$	0.049	-	-	0.046	-	-
	ScanNet$_{\text{pose}}\downarrow$	0.076	-	-	0.099	-	-
VLM	**VideoMME**$_{\text{vqa}}\uparrow$	60.7	-	-	-	61.8	-
	VideoMMMU$_{\text{vqa}}\uparrow$	40.0	-	-	-	38.7	-
	PerceptionTest$_{\text{vqa}}\uparrow$	68.9	-	-	-	67.6	-
	EgoSchema$_{\text{vqa}}\uparrow$	60.9	-	-	-	57.3	-
	VSI-Bench$_{\text{vqa}}\uparrow$	70.6	-	-	-	35.6	-
VLA	**CALVIN**$_{\text{mani}}\uparrow$	3.89	-	-	-	-	2.55
	Simpler-Bridge$_{\text{mani}}\uparrow$	45.8	-	-	-	-	53.7

stage trained for 120K steps at a resolution of 512×512. For optimization, we use Adam optimizer [40] with a peak learning rate of 1×10^{-4}, employing a 4K-step warmup followed by a cosine annealing schedule. Weight decay is disabled in both stages. All multi-frame sequences, including video and geometric scenes, are sampled with a sequence length of $T = 16$ frames per sample. Benefiting from our causal mask design, each training sample effectively provides distinct supervision signals for varying temporal contexts ranging from 1 to T frames.

4.3 Downstream Evaluations

To validate the universality and robustness of our learned representations, we structure our downstream experiments into four phases, as illustrated in Tab. 2: (i) **image & video probing**, which assesses the quality of global and dense features on standard image and video benchmarks; (ii) **streaming geometric reconstruction**, which evaluates the model's capability to estimate continuous camera poses and depth maps in an online fashion; (iii) **visual backbone for VLMs**, which incorporates the representation into Vision-Language Models (VLMs) for complex video and spatial reasoning; and (iv) **visual backbone for VLA policies**, where the model serves as the visual engine for Vision-Language-Action (VLA) policies in robotic manipulation. Crucially, to demonstrate its capacity as a truly universal foundational vision model, **we keep the backbone strictly frozen** across all downstream tasks.

Table 3: Comparison on online video depth estimation.

Method	Param.	Sintel		BONN		KITTI	
		Abs Rel↓	$\delta < 1.25$↑	Abs Rel↓	$\delta < 1.25$↑	Abs Rel↓	$\delta < 1.25$↑
Span3R [83]	600M	0.597	0.384	0.072	0.953	0.251	0.876
Cut3R [87]	600M	<u>0.417</u>	<u>0.507</u>	0.078	0.937	<u>0.118</u>	<u>0.876</u>
Point3R [94]	600M	0.481	0.448	**0.066**	<u>0.958</u>	**0.093**	**0.935**
Ours	400M	**0.314**	**0.583**	<u>0.072</u>	**0.967**	0.136	0.818

Table 4: Comparison on online video camera pose estimation.

Method	Param.	Sintel			TUM-dynamics			ScanNet		
		ATE↓	RPE_T ↓	RPE_R ↓	ATE↓	RPE_T ↓	RPE_R ↓	ATE↓	RPE_T ↓	RPE_R ↓
Span3R [83]	600M	0.329	0.110	4.471	0.056	0.021	<u>0.591</u>	<u>0.096</u>	0.023	0.661
Cut3R [87]	600M	**0.213**	<u>0.066</u>	**0.621**	**0.046**	<u>0.015</u>	**0.473**	0.099	<u>0.022</u>	**0.600**
Point3R [94]	600M	0.442	0.154	1.897	0.058	0.031	0.758	0.097	0.035	2.791
Ours	400M	<u>0.227</u>	**0.051**	<u>0.695</u>	<u>0.049</u>	**0.012**	0.601	**0.076**	**0.017**	<u>0.659</u>

Image & video probing. We first evaluate the versatility of the frozen features extracted by OmniStream across both static and dynamic domains. Following the DINOv3 [74] protocol for image tasks, we train task-specific linear heads on top of the frozen backbone to evaluate ImageNet-1K [25] classification, ADE20K [115] semantic segmentation, and NYUv2 [73] monocular depth estimation. For video understanding, we probe action recognition on Kinetics-400 (K400) [21] and Something-Something V2 (SSv2) [30] using attentive pooling from the V-JEPA 2 [6] setup. Furthermore, we evaluate dense spatiotemporal tracking on the DAVIS'17 [64] Video Object Segmentation (VOS) benchmark. Since V-JEPA 2 does not report DAVIS'17, we evaluate V-JEPA 2 ourselves using the same VOS probing pipeline for a fair comparison.

As presented in Tab. 2, OmniStream successfully bridges the gap between static and dynamic perception. On image benchmarks, it preserves robust spatial discrimination, achieving competitive results on dense prediction tasks (NYUv2 and ADE20K) comparable to the image specialist DINOv3. In contrast, OmniStream demonstrates superior temporal understanding ability, significantly outperforming DINOv3 on the motion-intensive SSv2 dataset (68.5% vs. 54.0%). Unlike typical video backbones that exhibit poor spatial alignment (*e.g.*, V-JEPA 2 at 44.2 $\mathcal{J}\&F$), OmniStream leverages the KV-cache to enable long-range extraction of the entire video. Our approach yields a $\mathcal{J}\&F$ mean of 71.6, performing on par with leading image-based per-frame extractors (DINOv3 at 73.2). Qualitative mask propagation results are shown in Fig. 3. This confirms that our unified multi-task training effectively injects temporal motion dynamics into the representations, without compromising the fine-grained spatial priors required for dense image tasks.

Streaming geometric reconstruction. To evaluate the geometric ability of our streaming representations, we benchmark OmniStream on online 3D reconstruction tasks following CUT3R [87]. We assess online video depth estimation across Sintel [17], BONN [60], and KITTI [29] (Tab. 3), and online camera pose estimation on Sintel, TUM Dynamics [76], and ScanNet [23] (Tab. 4).

Table 5: Comparison on VSI-Bench.

Methods	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist	Rel. Dir.	Route Plan	Appr. Order
GPT-4o [35]	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5-Pro [78]	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
LLaVA-OneVision-7B [43]	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-Video-7B [112]	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
Qwen2.5-VL-7B [9]	32.7	34.5	19.4	47.6	40.8	32.8	24.5	32.5	29.4
SpaceR-7B [59]	43.5	61.9	28.6	60.9	35.2	38.2	46.0	31.4	45.6
VILASR-7B [93]	45.4	63.5	34.4	60.6	30.9	48.9	45.2	30.4	49.2
VLM-3R-7B [26]	60.9	70.2	49.4	69.2	67.1	65.4	80.5	**45.4**	40.1
VST-7B [105]	60.6	72.0	44.4	74.3	68.3	59.7	55.8	44.9	65.2
SpaceMind [113]	69.6	**73.3**	**61.4**	**77.3**	74.2	67.2	**88.4**	44.3	70.5
OmniStream-7B	**70.6**	73.2	55.7	76.9	**74.8**	**72.3**	82.1	**45.4**	**84.6**

Across both tasks, our unified encoder consistently establishes highly competitive or superior results against specialized online 3D models, validating its robustness in real-time spatial perception. Notably, benefiting from our causal temporal attention mechanism, the architecture natively supports KV-caching during inference. This allows the model to process incoming video streams frame-by-frame with $O(T)$ temporal complexity per step, avoiding the redundant recomputation typical of bidirectional global attention. Furthermore, by leveraging the geometric extrapolation properties inherent to our 3D-RoPE, OmniStream can seamlessly generalize to much longer, unseen sequences that significantly exceed the training horizon. For instance, although OmniStream is pre-trained with a fixed temporal window of only $T = 16$ frames, it demonstrates impressive zero-shot length extrapolation capabilities, effortlessly processing continuous streams of up to 110 frames during evaluation.

Visual backbone for VLMs. Adhering to the framework of LLaVA-Video [112], we evaluate the efficacy of OmniStream as a general-purpose visual encoder for Video LLMs. We integrate our backbone with Qwen2.5-7B-Instruct [65] to perform Video and Spatial Question Answering. Further training details are provided in the supplementary material.

As shown in the *VLM* section of Tab. 2, despite a frozen visual encoder and an identical general video data recipe, OmniStream achieves better performance than LLaVA-Video on general Video QA benchmarks. Although marginally behind on VideoMME [27], it consistently achieves superior results across other major benchmarks like VideoMMMU [32], PerceptionTest [62], and EgoSchema [55], highlighting its effectiveness. Our model further demonstrates exceptional spatial reasoning capabilities. As shown in Tab. 5, on the **VSI-Bench** [102], a challenging benchmark for spatial intelligence, our model achieves state-of-the-art performance. Remarkably, it surpasses specialized geometry-aware baselines equipping extra geometry encoders, such as VLM-3R [26] and SpaceMind [113]. This result strongly validates that our unified streaming pre-training naturally engenders rich, actionable spatial understanding within the visual features themselves, eliminating the need for auxiliary geometric modules.

Table 6: Comparison on CALVIN ABC-D. All entries are evaluated under the VLM4VLA framework [111].

Model (VLM Backbone)	Task-1	Task-2	Task-3	Task-4	Task-5	Calvin↑
Expert Vision-Language-Action Models						
OpenVLA (Llama-2) [39]	0.792	0.644	0.499	0.368	0.245	2.548
pi0 (Paligemma-1) [14]	0.896	0.785	0.786	0.610	0.532	3.509
VLM with VLM4VLA Models (Full Fine-tuning)						
Qwen2.5VL-7B [9]	0.935	0.864	0.807	0.758	0.693	4.057
Qwen3VL-8B [8]	0.940	0.868	0.797	0.746	0.684	4.035
Paligemma-1-3B [13]	0.914	0.813	0.692	0.599	0.488	3.506
Paligemma-2-3B [75]	0.901	0.775	0.669	0.575	0.486	3.406
KosMos-2-1.7B [63]	0.878	0.721	0.591	0.498	0.408	3.096
VLM with VLM4VLA Models (Frozen Vision)						
Qwen2.5VL-7B	0.901	0.700	0.536	0.433	0.334	2.905
LLaVA-Video-7B [112]	0.876	0.701	0.542	0.438	0.340	2.898
OmniStream-7B	**0.931**	**0.848**	**0.768**	**0.703**	**0.634**	**3.885**

Visual backbone for VLA policies. Finally, we extend our evaluation to Embodied AI, assessing OmniStream as the "perception brain" for VLA policies. Following VLM4VLA [111], all VLA experiments follow the re-implantation evaluation, and we adapt OmniStream-7B into a Vision-Language-Action model by attaching a lightweight MLP action head, trained on two simulated robotic manipulation benchmarks: **CALVIN** (ABC-D split) [56] for long-horizon instruction following, and **SIMPLER-ENV** (Bridge V2 dataset) [45] for real-to-sim generalization.

Crucially, the visual encoder remains **fully frozen** during adaptation. This "frozen" setting differs from that of VLM4VLA, which freezes both the vision encoder and the MLP projector. Our protocol is therefore designed to fairly probe the quality of the visual backbone itself. As shown in Tab. 6 and Tab. 7, strong general-purpose VLMs like Qwen2.5-VL struggle significantly in this setting, failing to translate high-level semantic knowledge into precise low-level control signals, highlighting a critical *misalignment* between general visual features and the requirements of embodied interaction, such as depth perception and dynamics prediction. In contrast, OmniStream achieves strong performance even with a frozen backbone, reaching success rates of 3.885 on CALVIN and 45.8% on SIMPLER-ENV. This underscores the universality of our representation: by explicitly encoding 3D geometry and temporal dynamics during pre-training, OmniStream bridges the gap between perception and action. These results indicate that OmniStream can serve as a competitive general-purpose frozen visual backbone for VLA adaptation, paving the way for more efficient and generalizable embodied agents.

4.4 Ablation Study

To investigate the effects of multitask learning, we conduct experiments with different task configurations. For computational efficiency, all models in this study are evaluated using the checkpoints of Stage-1 pre-training at 224×224

Table 7: Comparison on SimplerEnv-Bridge. All entries are evaluated under the VLM4VLA framework [111].

Model (VLM Backbone)	Carrot	Eggplant	Spoon	Cube	Simpler↑
Expert Vision-Language-Action Models					
OpenVLA (Llama-2) [39]	4.2	0.0	0.0	12.5	4.2
pi0 (Paligemma-1) [14]	62.5	100.0	54.2	25.0	60.4
ThinkAct (Qwen2.5VL-7B) [33]	37.5	70.8	58.3	8.7	43.8
VLM with VLM4VLA Models (Full Fine-tuing)					
Qwen2.5VL-7B [9]	12.5	100.0	75.0	0.0	46.8
Qwen3VL-8B [8]	58.3	95.8	79.2	0.0	58.3
Paligemma-1-3B [13]	50.0	91.7	75.0	4.2	55.3
Paligemma-2-3B [75]	75.0	75.0	79.2	0.0	57.3
KosMos-2-1.7B [63]	37.5	100.0	75.0	29.2	60.4
VLM with VLM4VLA Models (Frozen Vision)					
Qwen2.5VL-7B	4.2	66.6	4.2	0.0	18.5
LLaVA-Video-7B [112]	0.0	**87.5**	33.3	0.0	30.2
OmniStream-7B	**33.3**	83.3	**41.7**	**25.0**	**45.8**

Table 8: Ablation study for pre-training tasks.

Method	*Video Probing*		*Image Probing*			*VLM & VLA Probing*		
	SSv2	**DAVIS'17**	**ImageNet**	**NYUv2**	**ADE20k**	**VSIBench**	**VideoMME**	**CALVIN**
	ACC@1↑	$\mathcal{J}\&F$-Mean↑	ACC@1↑	RMSE↓	mIoU↑	Acc↑	Acc↑	Seq Len↑
Ours	**69.3**	**71.6**	85.2	**0.379**	**49.6**	57.3	54.1	**3.80**
w/o. VideoSSL	63.0	67.7	**85.4**	0.420	47.2	**57.9**	**55.8**	3.42
w/o. 3D Geometry	68.4	69.7	85.0	0.471	42.3	52.5	53.8	3.34
w/o. Captioning	67.4	71.0	84.4	0.395	46.9	44.9	45.0	2.38

resolution. Tab. 8 demonstrates the indispensable role of each pre-training objective: **(i) video modeling (*w/o VideoSSL*):** Omitting video data from $\mathcal{L}_{\text{ssl}}$ severely degrades dynamic perception (SSv2 drops from 69.3% to 63.0%; DAVIS declines from 71.6 to 67.7) and embodied control (CALVIN drops from 3.80 to 3.42), confirming its necessity for capturing dynamic motions. **(ii) geometric reconstruction (*w/o 3D Geometry*):** Disabling the streaming geometric reconstruction objective collapses spatial perception (NYUv2 RMSE worsens to 0.471; ADE20K drops 7.3 mIoU). Crucially, this geometric deprivation causes a sharp decline in downstream spatial intelligence (VSI-Bench drops 4.8%) and VLA performance (CALVIN drops 0.46), validating that explicit 3D priors are prerequisites for Embodied AI. **(iii) vision-language alignment (*w/o Captioning*):** While pure vision probing remains stable, omitting the generative captioning task causes catastrophic failures in VLM integration (VideoMME and VSI-Bench plummet by 9.1% and 12.4%, respectively) and further impacts VLA policy performance. This highlights that early vision-language alignment is critical to bridging the semantic gap.

In conclusion, our unified multitask formulation is not merely a concatenation of independent losses, but a synergistic framework. The semantic (Captioning), dynamic (VideoSSL), and geometric (3D Reconstruction) objectives mutually reinforce the backbone, culminating in a robust representation that excels across both perceptual and embodied tasks.

Fig. 3: DAVIS'17 VOS visualization example.

5 Conclusion

In this paper, we introduce OmniStream, a unified vision foundation model that integrates causal spatiotemporal attention and 3D-RoPE into a pre-trained Vision Transformer. Through multi-task training on large-scale datasets from diverse sources, our model achieves competitive performance across a broad spectrum of tasks, including image and video probing, streaming geometric reconstruction, general and spatial video question answering, and robotic manipulation. While OmniStream does not uniformly surpass specialized state-of-the-art methods on every benchmark, its consistent competitiveness across such diverse tasks underscores the viability of training a single, versatile vision backbone that generalizes across semantic, spatial, and temporal reasoning. We believe this versatility represents a more meaningful step toward general-purpose visual understanding than benchmark-specific dominance. As our study focuses on validating this unified paradigm, we leave model scaling as a promising future direction that we expect will further close the remaining gaps with task-specific approaches. We hope this work establishes a solid foundation for streaming visual perception, paving the way for more capable interactive and embodied agents.

Acknowledgement

Weidi would like to acknowledge the funding from Scientific Research Innovation Capability Support Project for Young Faculty (ZY-GXQNJSKYCXNLZCXM-I22) and Shanghai Special Fund for Promoting High-Quality Industrial Development: Pilot Industry Innovation and Development Program (Artificial Intelligence Track, ID: 2025-GZL-RGZN-02020).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: ICCV (2021)
4. Arnold, E., Wynn, J., Vicente, S., Garcia-Hernando, G., Monszpart, Á., Prisacariu, V.A., Turmukhambetov, D., Brachmann, E.: Map-free visual relocalization: Metric pose relative to a single image. In: ECCV (2022)
5. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: CVPR (2023)
6. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
7. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
8. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
9. 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)
10. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
11. Baruch, G., Chen, Z., Dehghan, A., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In: NeurIPS (2021)
12. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML (2021)
13. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
14. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
15. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Bangalath, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. In: NeurIPS (2026)
16. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
17. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV (2012)

18. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022)
19. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
20. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
21. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
22. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024)
23. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
24. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: ICLR (2024)
25. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
26. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Wang, P., Qu, H., Zhou, S., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In: CVPR (2026)
27. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR (2025)
28. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. In: NeurIPS-DB (2023)
29. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR (2013)
30. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV (2017)
31. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
32. Hu, K., Wu, P., Pu, F., Xiao, W., Yue, X., Li, B., Zhang, Y., Liu, Z.: Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In: ACL (2026)
33. Huang, C.P., Wu, Y.H., Chen, M.H., Wang, F., Yang, F.E.: Thinkact: Vision-language-action reasoning via reinforced visual latent planning. In: NeurIPS (2026)
34. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: CVPR (2018)
35. 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)
36. Kar, O.F., Tonioni, A., Poklukar, P., Kulshrestha, A., Zamir, A., Tombari, F.: Brave: Broadening the visual encoding of vision-language models. In: ECCV (2024)
37. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Dynamicstereo: Consistent dynamic depth from stereo videos. In: CVPR (2023)

38. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
39. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., et al.: Openvla: An open-source vision-language-action model. In: CoRL (2025)
40. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
41. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023)
42. Lan, Y., Luo, Y., Hong, F., Zhou, S., Chen, H., Lyu, Z., Yang, S., Dai, B., Loy, C.C., Pan, X.: STream3R: Scalable sequential 3D reconstruction with causal transformer. In: ICLR (2026)
43. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. TMLR (2024)
44. Li, X., Li, P., Liu, M., Wang, D., Liu, J., Kang, B., Ma, X., Kong, T., Zhang, H., Liu, H.: Towards generalist robot policies: What matters in building vision-language-action models. arXiv preprint arXiv:2412.14058 (2024)
45. Li, X., Hsu, K., Gu, J., Mees, O., Pertsch, K., Walke, H.R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., et al.: Evaluating real-world robot manipulation policies in simulation. In: CoRL (2024)
46. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: CVPR (2018)
47. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Zhao, Y., Peng, S., Guo, H., Zhou, X., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: ICLR (2026)
48. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: CVPR (2024)
49. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
50. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
51. Liu, Y., Liu, Y., Di, S., Wang, H., Zhao, Z., Tian, L., Zhou, X., Zhou, J., Yao, J., Wang, Y., et al.: Versavit: Enhancing mllm vision backbones via task-guided optimization. arXiv preprint arXiv:2602.09934 (2026)
52. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
53. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In: CVPR (2024)
54. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In: ICLR (2023)
55. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: NeurIPS (2023)
56. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE-RAL (2022)
57. Mehl, L., Schmalfuss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: CVPR (2023)

58. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
59. 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)
60. Palazzolo, E., Behley, J., Lottes, P., Giguere, P., Stachniss, C.: Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In: IROS (2019)
61. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, Y.C.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: ICCV (2023)
62. Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. In: NeurIPS (2023)
63. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Ye, Q., Wei, F.: Grounding multimodal large language models to the world. In: ICLR (2024)
64. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
65. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115
66. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
67. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: ICCV (2021)
68. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. In: NeurIPS DB (2021)
69. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021)
70. Sablayrolles, A., Douze, M., Schmid, C., Jégou, H.: Spreading vectors for similarity search. In: ICLR (2019)
71. Shang, J., Schmeckpeper, K., May, B.B., Minniti, M.V., Kelestemur, T., Watkins, D., Herlant, L.: Theia: Distilling diverse vision foundation models for robot learning. In: CoRL (2024)
72. Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Zhao, Y., Huang, D.A., Yin, H., Sapra, K., Yacoob, Y., et al.: Eagle: Exploring the design space for multimodal llms with mixture of encoders. In: ICLR (2025)
73. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)

74. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
75. Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., et al.: Paligemma 2: A family of versatile vlms for transfer. arXiv preprint arXiv:2412.03555 (2024)
76. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: IROS (2012)
77. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
78. 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)
79. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In: NeurIPS (2024)
80. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: CVPR (2024)
81. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NeurIPS (2022)
82. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
83. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. In: 3DV (2025)
84. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
85. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: CVPR (2023)
86. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: ICML (2022)
87. Wang*, Q., Zhang*, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: CVPR (2025)
88. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
89. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: IROS (2020)
90. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. In: ICLR (2026)
91. Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A.S., Arunkumar, A., Stap, D., et al.: Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In: EMNLP (2022)

92. Wei, X., Liu, X., Zang, Y., Dong, X., Zhang, P., Cao, Y., Tong, J., Duan, H., Guo, Q., Wang, J., et al.: Videorope: What makes for good video rotary position embedding? In: ICML (2025)
93. 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. In: NeurIPS (2026)
94. Wu, Y., Zheng, W., Zhou, J., Lu, J.: Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. In: NeurIPS (2026)
95. Xia, H., Fu, Y., Liu, S., Wang, X.: Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In: CVPR (2024)
96. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: CVPR (2024)
97. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., et al.: xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872 (2024)
98. Yan, Y., Xu, J., Di, S., Liu, Y., Shi, Y., Chen, Q., Li, Z., Huang, Y., Xie, W.: Learning streaming video representation via multitask training. In: ICCV (2025)
99. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
100. Yang, H., Rao, J., Wu, H., Xie, W.: Soccermaster: A vision foundation model for soccer understanding. In: CVPR (2026)
101. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: CVPR (2025)
102. 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: CVPR (2025)
103. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
104. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS (2024)
105. 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)
106. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. CVPR (2020)
107. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023)
108. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
109. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
110. Zhang, C., Le Moing, G., Koppula, S., Rocco, I., Momeni, L., Xie, J., Sun, S., Sukthankar, R., Barral, J.K., Hadsell, R., Ghahramani, Z., Zisserman, A., Zhang, J., Sajjadi, M.S.M.: Efficiently reconstructing dynamic scenes one d4rt at a time. In: CVPR (2026)
111. Zhang, J., Chen, X., Guo, Y., Hu, Y., Chen, J.: VLM4VLA: Revisiting vision-language-models in vision-language-action models. In: ICLR (2026)

112. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
113. Zhao, R., Zhang, Z., Xu, J., Chang, J., Chen, D., Li, L., Sun, W., Wei, Z.: Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. arXiv preprint arXiv:2511.23075 (2025)
114. Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G., Guibas, L.J.: Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In: ICCV (2023)
115. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)
116. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image bert pre-training with online tokenizer. In: ICLR (2022)
117. Zhou, Y., Wang, Y., Zhou, J., Chang, W., Guo, H., Li, Z., Ma, K., Li, X., Wang, Y., Zhu, H., Liu, M., Liu, D., Yang, J., Fu, Z., Chen, J., Shen, C., Pang, J., Zhang, K., He, T.: Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. In: ICLR (2026)
118. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025)
119. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023)