VISION AS UNIFIED MULTIMODAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We formulate computer vision as unified multimodal generation, where heterogeneous visual tasks are expressed through the native text and image generation spaces of a unified multimodal model (UMM), without task-specific architectures. With this formulation, the single model **VisionUMM** matches leading task-specialized systems across structured visual understanding, dense geometric prediction, segmentation, and multi-view visual geometry. Natural-language instructions and optional visual prompts specify the task, target regions or views, and decoding convention. Responses are then generated as text for symbolic records, images for dense spatial targets, or mixed outputs for compositional tasks. To enable large-scale training, we convert heterogeneous computer vision annotations into instruction-response examples compatible with these native generation spaces. This conversion yields **VisionCorpus**, a computer-vision instruction-response corpus spanning text, image, and mixed text-and-image targets. Starting from an off-the-shelf pretrained UMM, VisionUMM is trained primarily on the VisionCorpus, using auxiliary multimodal data as a capability-preserving mixture and requiring no task-specific prediction heads or architectural changes. The resulting model covers detection, OCR, keypoints, segmentation, depth, surface normals, point maps, and camera pose estimation, and can follow language-defined variants that combine category, color, region, and other visual cues. These results suggest unified multimodal generation as a scalable route for integrating computer vision into general-purpose foundation models. The VisionUMM model and VisionCorpus are publicly available.

1 INTRODUCTION

Large language models Brown et al. (2020) have shown that diverse language tasks can be consolidated through prompting and generation, and unified multimodal models (UMMs) Chameleon Team (2024); Xie et al. (2025); Wu et al. (2025); Deng et al. (2025); Diao et al. (2026) further extend this paradigm to both text and image generation. We consider whether the entire spectrum of classical computer vision—from detection to multi-view 3D—can be expressed as unified multimodal generation within a single UMM, without task-specific heads. Computer vision has made remarkable progress through specialist systems Carion et al. (2020); He et al. (2017); Kirillov et al. (2023); Teed & Deng (2020); Yang et al. (2024a); Wang et al. (2025a) tailored to individual task families. Their outputs range from boxes and masks to motion fields, dense maps, and 3D geometry, each typically paired with task-specific architectures, losses, decoding rules, and evaluation protocols. This task-specific organization makes visual supervision difficult to share, reuse, and compose across tasks, motivating a shared and scalable formulation for computer vision.

Prior efforts approach this goal in complementary ways: sequence-format unifications Wang et al. (2022); Chen et al. (2022b); Lu et al. (2024) bring diverse annotations into shared model interfaces, but different output types still require serialization and parsing conventions; representation-centric models Kirillov et al. (2023); Wang et al. (2023b); Yang et al. (2024a); Wang et al. (2025a) generalize within visually coherent task families yet offer no unified output space and limited language control. In parallel, generative foundation models offer a different substrate: diffusion and image-generation models Ho et al. (2020); Rombach et al. (2022) provide visual generative priors for spatially aligned outputs, while multimodal large language models (MLLMs) Alayrac et al. (2022); Liu et al. (2023) bring language instructions and reasoning into visual perception. Recent perception work adopts these strengths only partially: image-generative methods He et al. (2025b); Zhao et al. (2025);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

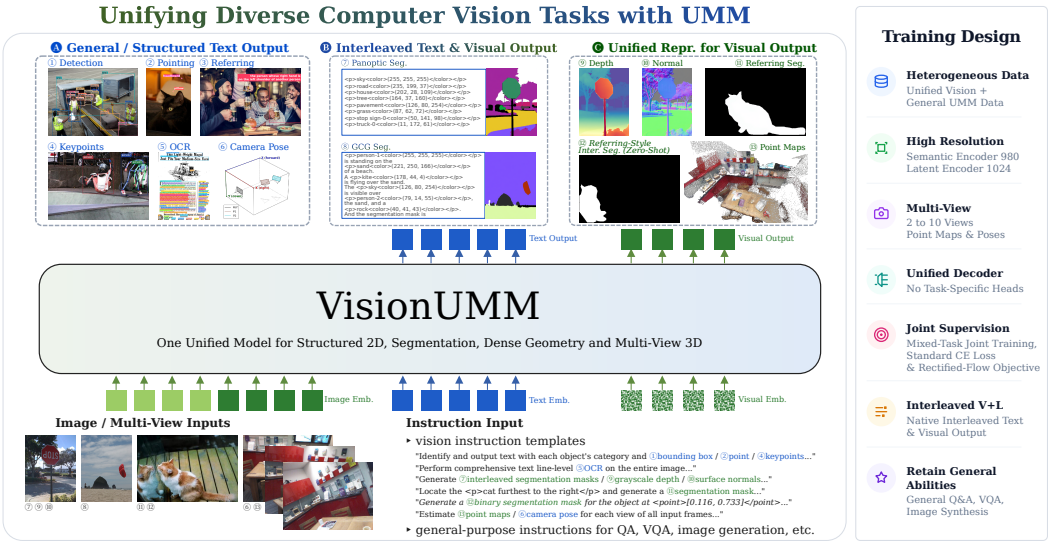


Figure 1: Overview of VisionUMM: heterogeneous computer vision annotations are converted into native text, image, and mixed text-image generation targets for joint training in a single UMM, without task-specific heads.

Gabeur et al. (2026) handle dense maps but struggle with symbolic records, while MLLM-based systems Lai et al. (2024); Wu et al. (2024a); Wang et al. (2026a) add language and reasoning yet still route dense outputs through task-specific decoders. Across these routes, symbolic records, dense spatial targets, and mixed outputs are still not expressed within one native multimodal generation framework.

To address this gap, we introduce VisionUMM, built on a simple formulation: heterogeneous computer vision tasks can be cast as unified multimodal generation. Natural-language instructions specify the task, target, output schema, and decoding convention, while text generation expresses symbolic visual answers such as categories, spatial references, OCR strings, and camera parameters. Image generation is natural for dense prediction because masks, depth maps, surface normals, and point maps are spatially aligned with the input image and can be represented on the same grid. Together, text and image generation provide complementary response spaces that can express a broad range of computer vision tasks. This formulation is, for visual perception, the analogue of what GPT Brown et al. (2020) did for NLP tasks: consolidating heterogeneous specialist supervision into a single generative interface.

Making this formulation trainable at scale requires converting heterogeneous computer vision annotations into instruction-response examples. We therefore construct VisionCorpus, spanning structured visual understanding, dense geometric prediction, segmentation, and multi-view visual geometry with decodable text, image, and mixed targets that can be recovered as boxes, masks, depth maps, surface normals, point maps, or camera poses. Built from the off-the-shelf UMM Bagel Deng et al. (2025), VisionUMM is trained on this corpus to express diverse visual task outputs through the unified multimodal generation interface illustrated in Figure 1. In quantitative evaluations reported in Section 4, VisionUMM leads on structured visual understanding, remains close to the strongest dense geometric prediction and segmentation baselines, and approaches the leading specialist in multi-view visual geometry.

VisionUMM is built on a simple premise: computer vision can become a native generative capability of unified foundation models, moving beyond a collection of isolated task-specific systems. First, we introduce a unified multimodal generation formulation that casts heterogeneous vision tasks into native input-output spaces of UMMs. Second, we construct VisionCorpus, a large-scale computer-vision instruction-response corpus with decodable text, image, and mixed targets. Third, we train VisionUMM and demonstrate that a single UMM can achieve competitive performance across four heterogeneous perception families and output modalities. Together, this work points toward a new era of computer vision, where perception is no longer engineered task by task, but becomes a pro-

grammable, generative, and extensible capability of unified foundation models, providing a reusable basis for future research.

2 Related Works

2.1 Early Unified Vision Models and Interfaces

Early unified vision models explore how diverse vision and vision-language tasks can be expressed through shared task formats. This line begins with casting structured visual predictions as sequences, as in Pix2Seq for object detection Chen et al. (2022a), and then extends to broader task families in Pix2Seq v2 Chen et al. (2022b) and text-box generation in UniTAB Yang et al. (2022). OFA Wang et al. (2022) and Uni-Perceiver Zhu et al. (2022); Li et al. (2023) further expand this idea to cross-modal and generalist vision-language modeling. Unified-IO Lu et al. (2023) and Unified-IO 2 Lu et al. (2024) broaden the direction by representing heterogeneous inputs and outputs in a common token space, while Florence-2 Xiao et al. (2024) scales prompt-based sequence generation to a wide range of vision tasks. Yet their unification still relies heavily on shared sequence formats and task-specific encoding rules, making the interface less natural for dense maps and structurally diverse visual outputs.

2.2 Task-Family Generalization with Visual Foundation Models

Another line of work seeks generalization with visual foundation models and pretrained visual representations. Representative pretraining methods such as MAE He et al. (2022) and DINOv2 Oquab et al. (2024) provide reusable visual features for recognition, dense prediction, and geometry transfer. SAM Kirillov et al. (2023) shows that promptable segmentation can cover a broad family of mask prediction tasks, while Painter Wang et al. (2023a) and SegGPT Wang et al. (2023b) use visual in-context examples to specify different image-to-image perception tasks. In geometry, Depth Anything Yang et al. (2024a) scales depth prediction with strong visual representations and large data, and VGGT Wang et al. (2025a) extends feed-forward visual modeling to cameras, point maps, reconstruction, and point tracking. These works generalize strongly within visually coherent task families, but the absence of a unified output space keeps their unification partial. They also provide limited support for understanding complex, compositional, or open-ended language instructions.

2.3 Image-Generative Models for Dense Perception

Diffusion and image generation models provide a generative alternative to conventional dense prediction by producing spatially aligned visual outputs. Marigold Ke et al. (2024) adapts diffusion priors to monocular depth estimation, while Lotus, Lotus-2, and FE2E He et al. (2025b;a); Wang et al. (2025b) adapt powerful image generative or editing models for geometric dense prediction such as depth and normal estimation. DICEPTION Zhao et al. (2025) casts multiple perceptual tasks as conditional image generation in a shared RGB output space, and Visual Bridge Gao et al. (2026) studies generative visual perception representations. Vision Banana Gabeur et al. (2026) further scales this direction with a stronger image-generation foundation model and lightweight instruction tuning, enabling richer language-conditioned visual predictions while preserving generation capability. These methods make dense targets compatible with the native image space of generative models, but their unification remains largely image-side. Since symbolic records, task schemas, camera parameters, and decoding rules are naturally textual, image generation alone cannot cover the symbolic and mixed outputs required by many tasks.

2.4 MLLM-based Perception and Task Modules

As powerful foundation models that integrate language and vision, MLLMs bring new opportunities to visual perception. One line extends MLLMs to structured visual grounding: Kosmos-2 Peng et al. (2023), Shikra Chen et al. (2023), and Ferret You et al. (2023) connect language generation with boxes, coordinates, and regions, while Rex-Omni Jiang et al. (2026) broadens this direction through next-point prediction. Another line represents dense spatial outputs through discrete token or logit spaces, as in Text4Seg Lan et al. (2025) and DenseMLLM Li et al. (2026), but preserving fine spatial details requires carefully designed tokenization and decoding schemes. Decoder-based

systems instead attach task-specific modules to MLLMs: LISA Lai et al. (2024), PixelLM Ren et al. (2024), and X-SAM Wang et al. (2026a) connect language reasoning to segmentation or mask decoders, while G^2VLM Hu et al. (2026) and VisionLLM v2 Wu et al. (2024a) introduce downstream modules or connections for geometry and broader perception tasks. This preserves useful inductive biases but keeps output spaces fragmented across task-specific components.

Recent UMMs Chameleon Team (2024); Xie et al. (2025); Wu et al. (2025); Deng et al. (2025); Diao et al. (2026) extend multimodal foundation models from understanding-centered MLLMs to models that natively support both text and image generation. They typically model language and visual content within shared or coupled generative spaces, allowing a single model to understand instructions, generate text, and synthesize images. This makes it timely to revisit the unification of computer vision tasks: heterogeneous visual supervision can be expressed as decodable targets in the same native spaces of a UMM. VisionUMM follows this route by aligning symbolic records, dense spatial targets, and mixed responses with the native text-and-image input-output spaces of a UMM, turning them into decodable training targets for a single model.

Complementary to this line of work, SenseNova-SI Cai et al. (2026) studies how curated spatial-intelligence supervision can scale spatial reasoning abilities in multimodal foundation models including UMM backbones. Our work builds on the same broader view that spatial and geometric abilities can be cultivated in foundation models, but focuses on a different question: how heterogeneous computer vision annotations can be converted into native text, image, and mixed generation targets for benchmark-decodable perception outputs.

3 DATA

3.1 DATA PROTOCOL

Training a UMM on unified computer vision tasks requires casting heterogeneous annotations into the model's native text-and-image generation spaces. We therefore convert each source annotation into an instruction-response example with one or more visual inputs, a language instruction, and a decodable target response. This conversion is organized through a data protocol covering four computer vision families: structured visual understanding, dense geometric prediction, segmentation, and multi-view visual geometry.

The instruction specifies the task intent, output schema, and decoding convention, while the response is represented as text, an image, or a mixed text-image output that can be recovered as benchmark-compatible labels, coordinates, masks, dense maps, or camera parameters. Figure 2 illustrates representative examples produced by this protocol across the four families. We use lightweight textual markers for structured and segmentation responses, and reserved special tokens for camera-pose records; Supplementary Tables S3, S9, and S12 summarize these conventions.

Structured Visual Understanding. This family covers structured visual understanding tasks whose outputs can be decoded as sparse symbolic records, including detection, referring, pointing, keypoint localization, OCR, layout understanding, and GUI grounding Jiang et al. (2026). We represent these annotations as text generation targets: labels, transcripts, and task-specific attributes remain ordinary text, while spatial fields are written as normalized image coordinates. Lightweight markers such as `<p>`, `<bbox>`, and `<point>` delimit phrases and coordinate fields. The generated response is parsed back into typed benchmark records, so detection, grounding, OCR, pointing, keypointing, layout, and GUI tasks share the same text-generation space while being separated by language instructions and textual schemas.

Dense Geometric Prediction. This family covers dense, pixel-aligned geometric targets, including depth maps and surface normals. These targets share the spatial layout of the input image, making conditional image generation a natural representation. The instruction specifies the requested geometric quantity and decoding convention, while the response image stores the corresponding dense signal in a deterministic visual encoding. For depth, valid metric values are converted to inverse depth and rendered as normalized grayscale images; surface normals are rendered as RGB maps whose channels encode the normal components. Generated images can then be decoded back into metric-compatible depth or normal maps for evaluation.



Figure 2: Representative data formats used by VisionUMM. The examples illustrate structured text outputs, segmentation masks, dense geometric maps, and multi-view geometric outputs under the unified instruction-response protocol.

Segmentation. Following the task taxonomy of X-SAM Wang et al. (2026a), this family covers single-target and multi-region segmentation. For single-target tasks such as referring, reasoning, and interactive segmentation, the instruction identifies the target region and the response is a binary mask image with fixed foreground and background colors. Interactive segmentation additionally provides visual prompts, such as points, boxes, scribbles, or masks, together with the input image. For multi-region tasks, such as generic segmentation (semantic and panoptic segmentation), we use a mixed text-image response: the text component lists regions and uses the `<color>` marker to specify RGB palette values in prompts or generated legends, while the image component renders the corresponding color-coded mask. Grounded conversation generation (GCG) segmentation further exercises this format: given an open-ended instruction, the model first produces region descriptions and color assignments, then renders the mask according to the generated legend. This design lets language handle category names, referring expressions, reasoning-derived regions, and generated region descriptions, while the image channel preserves dense pixel-level supervision.

Multi-View Visual Geometry. This family follows feed-forward visual geometry settings such as VGGT Wang et al. (2025a) and uses an ordered image set as visual input. The instruction specifies the view order, reference coordinate frame, and requested output types. Dense scene geometry is represented as image outputs in the form of per-view dense XYZ point maps; each point map stores aligned and normalized 3D coordinates in its RGB channels. Camera pose outputs are represented as structured sequences. Each target view is encoded relative to the reference frame as a quaternion rotation, a translation direction, and a scale, with reserved tokens such as `<frame>`, `<quat>`, `<offset>`, and `<scale>` delimiting the boundaries of view entries and pose fields. This mixed response format keeps dense point-map reconstruction in the image space and discrete geometric metadata in the text space.

3.2 Corpus Construction

Following this protocol, we construct **VisionCorpus**, a large-scale computer vision corpus built from public images. We directly convert available annotations when possible and generate or curate

additional targets to address incomplete supervision and improve data diversity. The corpus covers four source families: structured visual understanding, dense geometric prediction, segmentation, and multi-view visual geometry, providing respectively detection, OCR, layout, keypoint, depth, normal, mask, reconstruction, and camera-pose supervision. When the same source image appears in multiple tasks, we keep the corresponding converted examples as independent samples, since each task is defined by its own instruction and target response. For datasets that overlap with our evaluation benchmarks, we preserve the official benchmark splits and exclude the corresponding evaluation images and annotations from training. Supplementary Appendix A lists all datasets and construction details for each source family.

Some public sources nevertheless lack complete or protocol-compatible supervision, so we additionally generate or curate targets for these cases. For structured visual understanding, we draw on the data-construction pipeline of Rex-Omni Jiang et al. (2026) to construct part of the detection and OCR data. For dense geometric prediction, we use MoGe-2 Wang et al. (2025c) to densify incomplete supervision and expand data diversity by generating additional depth and normal targets, followed by validity and scene-content filtering. For segmentation, we curate mixed text-image targets such as grounded conversation generation segmentation, where region descriptions and color legends must be aligned with mask images. For multi-view visual geometry, we complete sparse depth with LingBot-Depth Tan et al. (2026) and filter examples with invalid depth, missing camera information, or inconsistent view metadata. We released this generated and curated subset as **VisionCorpus-50M**, a 50-million-example collection of converted computer vision supervision, and provide source lists, prompt templates, conversion rules, and examples to reproduce the remaining public-source portion of VisionCorpus. Supplementary Table S14 summarizes the released VisionCorpus-50M task families and frame counts.

4 Experiments

4.1 Training Setup

We adapt Bagel-7B-MoT Deng et al. (2025) through supervised fine-tuning on VisionCorpus together with auxiliary multimodal data spanning visual question answering, text-to-image, and image-to-image tasks. Textual targets are optimized with a cross-entropy loss in the native textual space, whereas visual targets are encoded into the VAE latent space and optimized with a rectified-flow objective. This mixed-task training aims to equip the base UMM with benchmark-compatible computer vision abilities while preserving its general multimodal capabilities. Detailed optimization settings are provided in Supplementary Appendix B.

4.2 Quantitative Evaluation

We evaluate whether VisionUMM covers four computer vision families through unified multimodal generation: structured visual understanding, dense geometric prediction, segmentation, and multi-view visual geometry. All tasks are formulated with natural-language instructions. Textual outputs are parsed into boxes, points, recognized text, keypoints, and camera parameters, while image outputs are decoded into masks, depth maps, normal maps, or 3D point maps using deterministic conversion rules. The resulting predictions are evaluated with the corresponding benchmark metrics.

Structured Visual Understanding. Structured visual understanding covers textual predictions such as bounding boxes, points, recognized text, and keypoint coordinates. As shown in Table 1, the evaluation covers box- and point-based localization and grounding on COCO-Common Lin et al. (2014), HumanRef Zhang et al. (2025), RefCOCOg val/test Yu et al. (2016), LVIS Gupta et al. (2019), Dense200 Jiang et al. (2026), and VisDrone Du et al. (2019), OCR localization on HierText Long et al. (2022; 2023) and ICDAR15 Karatzas et al. (2015), GUI grounding on ScreenSpot-V2 Wu et al. (2024b), and keypoint detection on COCO-Kpt. Lin et al. (2014). These tasks test coordinate-level prediction through a unified text interface, including long object lists and precise localization; VisionUMM achieves strong overall performance, especially on dense, long-tailed, small-object, referring, and OCR localization benchmarks.

Dense Geometric Prediction. Dense geometric prediction evaluates pixel-aligned geometric outputs, including monocular depth estimation and surface normal estimation. Depth maps are decoded

Table 1: Quantitative comparison of structured visual understanding. Performance is assessed using F1@mIoU for box-based detection, referring, and OCR localization tasks, F1@Point for VisDrone point localization, click accuracy for GUI grounding, and F1@mOKS for keypoint localization. Higher values indicate better performance for all metrics.

Method	Object Detection							OCR		GUI	Keypoint
	COCO-Com.	HR/RefCOCOg V/T		LVIS	Dense200	VisDrone		HierText	ICDAR15	ScreenSpot-V2	COCO-Kpt.
	bbox	bbox		bbox	bbox	bbox	point	bbox	bbox	bbox	point
Grounding DINO-Swin-T (Liu et al., 2024)	**56.6**	25.2 / 45.9 / 46.8		38.8	33.1	38.5	–	–	–	–	–
Bagel (Deng et al., 2025)	50.2	74.6 / 76.4 / 77.8		46.8	42.4	23.0	36.9	7.1	15.8	81.1	–
Qwen3-VL-8B-Instruct (Bai et al., 2025)	46.6	70.4 / 72.3 / 72.6		43.2	13.5	28.7	35.7	22.4	25.4	90.5	–
Qwen3.5-9B (Qwen Team, 2026)	49.3	71.7 / 72.1 / 72.6		43.2	27.5	26.8	41.7	19.6	11.4	**92.2**	–
LocateAnything (Wang et al., 2026b)	54.7	78.7 / 76.7 / 77.6		50.7	58.7	39.9	60.4	29.1	26.4	85.5	–
Rex-Omni (Jiang et al., 2026)	52.9	79.9 / 73.6 / 74.3		46.9	58.3	35.8	58.9	28.0	28.1	88.4	32.6
VisionUMM	**56.6**	**80.2 / 79.6 / 80.5**		**54.8**	**66.8**	**43.3**	**62.9**	**31.2**	**49.5**	85.9	**34.6**

Table 2: Quantitative comparison of dense geometric prediction. In each panel, the upper block reports geometry-specialized models, while the lower block compares generation-based methods. Methods denoted with an asterisk (∗) have been re-evaluated to ensure a direct and consistent comparison with our method.

Method	Depth abs rel. ↓ / δ_1 ↑					Normal mean err. ↓ / $\delta_{11.25}$ ↑		
	NYUv2	KITTI	ETH3D	ScanNet	DIODE	ScanNet	iBims-1	NYUv2
DSINE (Bae & Davison, 2024)	–	–	–	–	–	16.2 / 61.0	17.1 / 67.4	16.4 / 59.6
DepthAnything (Yang et al., 2024a)	4.3 / **98.1**	7.6 / 94.7	12.7 / 88.2	4.3 / 98.1	26.0 / 75.9	–	–	–
DepthAnything V2 (Yang et al., 2024b)	4.5 / 97.9	7.4 / 94.6	13.1 / 86.5	4.2 / 97.8	26.5 / 73.4	–	–	–
∗MoGe-2 (Wang et al., 2025c)	**3.5** / 98.0	**5.5 / 97.7**	**3.4 / 98.8**	**3.4 / 98.3**	**23.0 / 82.3**	**12.8 / 68.4**	**14.7 / 70.4**	**14.7 / 62.3**
Marigold (Ke et al., 2024)	5.5 / 96.4	9.9 / 91.6	6.5 / 95.9	6.4 / 95.2	30.8 / 77.3	21.3 / 45.6	18.5 / 64.7	20.9 / 50.5
DICEPTION (Zhao et al., 2025)	6.1 / 96.0	6.9 / 94.9	5.0 / 97.5	7.2 / 94.4	28.9 / 72.2	18.8 / 53.6	–	18.3 / 52.9
FE2E (Wang et al., 2025b)	4.1 / 97.7	6.6 / **96.0**	**3.8 / 98.7**	4.4 / 97.5	22.8 / **81.2**	13.8 / 67.2	**15.1 / 70.6**	16.2 / 59.6
Lotus-2 (He et al., 2025a)	4.1 / 97.6	6.7 / 94.5	4.6 / 98.1	4.2 / 97.6	22.1 / 75.2	14.2 / 66.8	15.4 / 70.4	16.9 / 59.0
VisionUMM	**4.0 / 98.1**	**5.9** / 95.9	4.3 / 97.4	**3.9 / 98.0**	**20.6** / 76.4	**12.8 / 68.9**	15.4 / 69.1	**14.4 / 62.7**

and evaluated using affine-invariant depth metrics, while color-coded surface normal maps are evaluated by angular error. As shown in Table 2, we report depth results on NYUv2 Silberman et al. (2012), KITTI Geiger et al. (2013), ETH3D Schöps et al. (2017), ScanNet Dai et al. (2017), and DIODE Vasiljevic et al. (2019), together with normal estimation results on ScanNet, iBims-1 Koch et al. (2018), and NYUv2. These results show that VisionUMM produces competitive dense geometric maps through image generation, outperforming recent generation-based methods on several benchmarks while remaining close to geometry-specialized models.

Segmentation. Segmentation evaluates mask prediction under semantic, referring, reasoning, grounded, and interactive guidance. Generated images are decoded into benchmark masks using the color palettes, visual prompts, or target specifications defined by each instruction. As shown in Table 3, we report generic, referring, reasoning, Grounded Conversation Generation (GCG) Rasheed et al. (2024), and interactive segmentation results using the corresponding benchmark metrics. VisionUMM achieves competitive overall performance among unified segmentation and multimodal baselines, with strong results on reasoning and GCG segmentation. Specialized segmentation models remain stronger on several generic and referring segmentation metrics, where segmentation-specific pretrained mask models such as SAM Kirillov et al. (2023) or Mask2Former Cheng et al. (2022) can provide strong mask priors.

Multi-View Visual Geometry. Multi-view visual geometry evaluates point map reconstruction and camera pose estimation from multiple input images, with results for both tasks reported in Table 4. For reconstruction on 7Scenes Shotton et al. (2013) and ETH3D Schöps et al. (2017), we report accuracy and completeness following VGGT Wang et al. (2025a), together with F1-score using the thresholds from Depth Anything 3 Lin et al. (2026); for camera pose estimation on RealEstate10K (Re10K) Zhou et al. (2018) and CO3Dv2 Reizenstein et al. (2021), we report relative rotation accuracy (RRA), relative translation accuracy (RTA), and AUC under the 30-degree threshold. All methods are re-evaluated for direct comparison. VisionUMM performs strongly among generalist geometric approaches, especially on ETH3D reconstruction and camera pose estimation, but remains behind geometry-specialized models on several metrics.

Table 3: Quantitative comparison of segmentation. For Gen. Seg., we report PQ for panoptic segmentation (Pan.) and mIoU for semantic segmentation (Sem.). For Ref. Seg., we report cIoU, defined as the ratio of total true positives to total union. For Rea. Seg., we report gIoU; for GCG Seg. and Inter. Seg., we report mIoU. Higher values indicate better performance for all metrics.

Method	Gen. Seg. Pan. / Sem.	Ref. Seg. RefCOCO / + / g	Rea. Seg. Val / Test	GCG Seg. Val / Test	Inter. Seg. Point / Box
LISA-7B (Lai et al., 2024)	–	74.9 / 65.1 / 67.9	52.9 / 47.3	62.0 / 61.7	–
PSALM (Zhang et al., 2024)	**55.9** / **66.6**	83.6 / 72.9 / 73.8	–	–	64.3 / 67.3
Text4Seg (Lan et al., 2025)	–	79.2 / 72.8 / 74.0	59.1 / 57.1	–	–
LENS (Zhu et al., 2026)	–	84.2 / **79.4** / 81.2	62.1 / 57.2	–	–
ConverSeg (Sahoo & Gkioxari, 2026)	–	79.4 / 74.3 / 74.9	61.9 / 57.0	–	–
X-SAM (Wang et al., 2026a)	54.7 / 66.5	**85.1** / 78.0 / **83.8**	56.6 / 57.8	**69.4** / **69.0**	**65.4** / 70.0
VisionUMM	48.8 / 64.0	81.3 / 76.0 / 80.3	**63.2** / **60.7**	65.7 / 66.2	60.9 / **73.9**

Table 4: Quantitative comparison of multi-view point map reconstruction and camera pose estimation. We evaluate feed-forward geometric models (top) alongside generalist geometric approaches (bottom). MapAnything is classified within the latter category, as it accepts images with optional geometric inputs and fuses the encoded features with a multi-view transformer.

Method	Multi-View Reconstruction Acc.↓ / Comp.↓ / F1↑		Camera Pose RRA@30↑ / RTA@30↑ / AUC@30↑	
	7Scenes	ETH3D	Re10K	CO3Dv2
DUSt3R (Wang et al., 2024)	0.026 / 0.034 / 87.1	0.359 / 0.531 / 66.6	99.8 / 84.9 / 67.6	97.7 / 93.4 / 78.3
DepthAnything3 (Lin et al., 2026)	**0.020** / **0.026** / **90.5**	0.228 / 0.212 / 76.6	**100.0** / **96.4** / **89.6**	**99.3** / **98.0** / **91.8**
VGGT (Wang et al., 2025a)	0.023 / 0.032 / 88.4	**0.177** / **0.155** / **80.9**	**100.0** / 93.5 / 79.3	98.3 / 96.6 / 89.2
MoRe (Fang et al., 2026)	0.038 / 0.039 / 77.1	0.348 / 0.318 / 62.7	**100.0** / 94.0 / 79.1	98.4 / 96.3 / 83.0
MapAnything (Keetha et al., 2026)	**0.027** / 0.029 / 87.8	0.400 / 0.524 / 67.0	**100.0** / 93.5 / **80.7**	95.5 / 91.6 / 70.9
G2VLM (Hu et al., 2026)	0.084 / 0.056 / 59.2	0.784 / 0.553 / 36.7	99.8 / 77.5 / 51.8	96.3 / 92.0 / 55.2
VisionUMM	0.028 / **0.026** / **87.9**	**0.301** / **0.175** / **72.2**	99.8 / **94.2** / 77.3	**97.4** / **95.4** / **80.1**

4.3 Comparison with Generalist Vision Models

We further compare VisionUMM with Youtu-VL Youtu-VL Team (2026) and Vision Banana Gabeur et al. (2026), two representative generalist visual models with complementary vision-language and image-generation-centered capabilities, as shown in Table 5. We therefore compare Youtu-VL on structured and semantic perception benchmarks, including detection, referring segmentation, semantic segmentation, and depth, and Vision Banana on image-space prediction benchmarks, including segmentation, depth, and surface normals. The comparison shows that recent generalist visual models extend beyond single-task specialization, but remain shaped by their native output modality. VisionUMM performs strongly against Youtu-VL on structured, semantic, and depth benchmarks, and remains competitive with Vision Banana on image-space segmentation and dense prediction benchmarks. This broader coverage comes from unified multimodal generation over text, image, and mixed outputs, which better matches the heterogeneous output forms required by computer vision tasks.

4.4 Qualitative Results

Figure 3 illustrates the outputs produced by VisionUMM across the four task families. Different instructions elicit structured localization, OCR and keypoint predictions, segmentation masks, and dense geometric maps; multi-view inputs additionally support reconstruction and camera pose estimation. These examples illustrate the range of task-specific representations expressed through one multimodal generation interface. Supplementary Appendix C provides further examples for each task family and for language-defined task variants beyond the standard benchmarks.

Table 5: Quantitative comparison with recent generalist visual models under the metrics reported in their original papers. For Vision Banana, † denotes a Gemini-assisted ReasonSeg result, and ‡ marks depth scores reported under an absolute-depth protocol, whereas VisionUMM uses affine-invariant depth evaluation; these entries are included only as reference comparisons.

(a) Comparison with Youtu-VL

Method	Detection mAP ↑	Sem. Seg. mIoU ↑	Ref. Seg. cIoU ↑	Depth δ_1 ↑
	COCO	Cityscapes	RefCOCO / + / g	NYUv2
Youtu-VL	47.1	70.4	80.7 / **76.2** / 76.5	90.4
VisionUMM	**53.7**	**71.2**	**81.3** / 76.0 / **80.3**	98.1

(b) Comparison with Vision Banana

Method	Sem. Seg. mIoU ↑	Ref. Seg. cIoU ↑	Rea. Seg. gIoU ↑	Depth δ_1 ↑				Normal mean err. ↓		
	Cityscapes	RefCOCOg	ReasonSeg	KITTI	NYUv2	DIODE	ETH3D	NYUv2	ScanNet	DIODE
Vision Banana	69.9	73.8	79.3†	91.5‡	94.8‡	91.7‡	93.5‡	17.8	15.1	13.8
VisionUMM	71.2	80.3	63.2	95.9	98.1	76.4	97.4	14.4	12.8	15.3

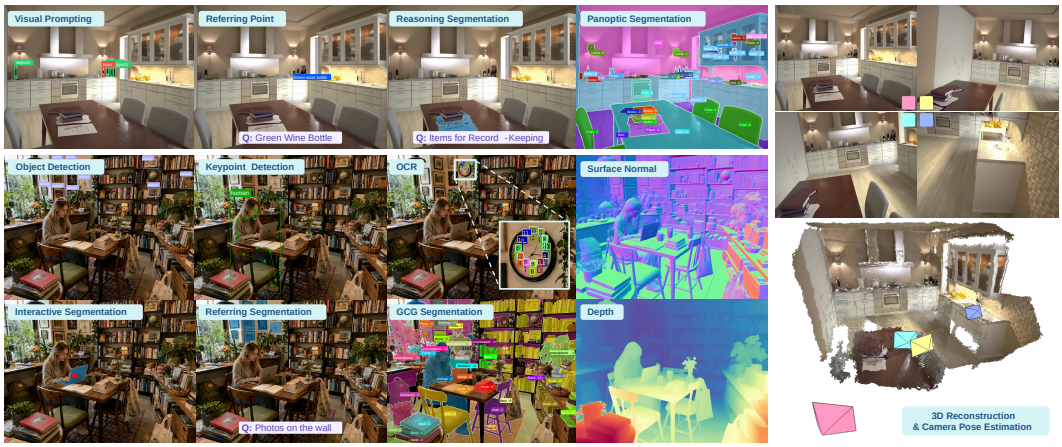


Figure 3: Qualitative overview of VisionUMM across four task families. Task instructions specify structured predictions, segmentation masks, dense geometric maps, and multi-view reconstruction and camera poses within a single multimodal generation interface.

5 Conclusion

In this work, we present unified multimodal generation as a formulation for computer vision, analogous to the role of GPT-style generative modeling in NLP. We present VisionCorpus by casting heterogeneous computer vision annotations into text, image, and mixed text-and-image generation targets, and train VisionUMM within the same formulation. This conversion enables large-scale training of a single UMM across structured visual understanding, dense geometric prediction, segmentation, and multi-view visual geometry without task-specific heads. The VisionUMM matches leading task-specialized systems and supports language-defined task variants, suggesting a scalable route for absorbing computer vision supervision into general-purpose foundation models. The generative modeling of diverse 2D and 3D perception tasks may further cultivate implicit spatial understanding, connecting this paradigm naturally to the rapidly emerging frontiers of physical intelligence. Future work could extend task specification through in-context examples, incorporate video supervision, and scale the corpus and model to acquire richer visual and spatial knowledge.

AI USE STATEMENT

We used generative AI in constructing training data. In the grounding and referring data engines, Qwen2.5-VL generated image descriptions, noun phrases, and referring expressions. These outputs were paired with spatial annotations and subjected to the matching and filtering procedures described in the data construction protocols. This use constitutes synthetic training-data generation, including synthetic annotations for existing images.

The data pipelines also use pretrained vision models and OCR tools for annotation, verification, and data curation: Rex-Omni for spatial annotation and expression verification, SAM for box-to-point conversion, PaddleOCR and RapidOCR for OCR re-annotation, MoGe-2 for dense geometric pseudo-labels, and LingBot-Depth for depth completion. We report these model-assisted steps separately from generative language-model assistance.

We additionally used generative AI tools during manuscript preparation to create and edit experiment code and to assist with manuscript drafting and editing, literature and submission-policy search, paper organization, LaTeX figure and table formatting, reference formatting, and document inspection. The authors reviewed and verified the final manuscript and take responsibility for all AI-assisted code, data, analyses, and claims.

REPRODUCIBILITY STATEMENT

The paper organizes the information needed to reproduce the work in the main text and in the accompanying Supplementary Material. Section 3 describes the formulation of the task and the composition of the corpus, Section 4.1 describes the training setup and Section 4 specifies the evaluation protocols and benchmarks. Supplementary Appendix A provides the source datasets, prompt and output formats, data-engine details, filtering rules, and the release summary in Supplementary Table S14. The code, pretrained weights, and released VisionCorpus are the accompanying reproducibility artifacts for these materials. Supplementary Appendix B provides the detailed optimization and multi-view training settings.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9535–9545, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7879–7890, 2026.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. doi: 10.48550/arXiv.2405.09818.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2Seq: A language modeling framework for object detection. In *International Conference on Learning Representations*, 2022a.

Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David Fleet, and Geoffrey E. Hinton. A unified sequence interface for vision tasks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31333–31346. Curran Associates, Inc., 2022b.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. arXiv preprint arXiv:1702.04405.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, et al. SenseNova-U1: Unifying multimodal understanding and generation with NEO-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, Liefeng Bo, Hailin Shi, Rui Zhu, Aashish Kumar, Aijin Li, Almaz Zinollayev, Anuar Askergaliyev, Arne Schumann, Binjie Mao, Byeongwon Lee, Chang Liu, Changrui Chen, Chunhong Pan, Chunlei Huo, Da Yu, DeChun Cong, Dening Zeng, Dheeraj Reddy Pailla, Di Li, Dong Wang, Donghyeon Cho, Dongyu Zhang, Furui Bai, George Jose, Guangyu Gao, Guizhong Liu, Haitao Xiong, Hao Qi, Haoran Wang, Heqian Qiu, HongLiang Li, Huchuan Lu, Ildoo Kim, Jaekyum Kim, Jane Shen, Jihoon Lee, Jing Ge, Jingjing Xu, Jingkai Zhou, Jonas Meier, Jun Won Choi, Junhao Hu, Junyi Zhang, Junying Huang, Kaiqi Huang, Keyang Wang, Lars Sommer, Lei Jin, and Lei Zhang. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

Juntong Fang, Zequn Chen, Weiqi Zhang, Donglin Di, Xuancheng Zhang, Chengmin Yang, and Yu-Shen Liu. MoRe: Motion-aware feed-forward 4d reconstruction transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28914–28924, June 2026.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Yilin Gao, Shuguang Dou, Junzhou Li, Zhiheng Yu, Yin Li, Dongsheng Jiang, and Shugong Xu. Visual bridge: Universal visual perception representations generating. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21234–21242, 2026. doi: 10.1609/aaai.v40i25.39268.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Jing He, Haodong Li, Mingzhi Sheng, and Ying-Cong Chen. Lotus-2: Advancing geometric dense prediction with powerful image generative model. *arXiv preprint arXiv:2512.01030*, 2025a.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, 2025b.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2VLM: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9535–9546, June 2026.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25472–25483, June 2026.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1156–1160, 2015. doi: 10.1109/ICDAR.2015.7333942.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*. IEEE, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In *European Conference on Computer Vision Workshops*, pp. 331–348, 2018.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4Seg: Reimagining image segmentation as text generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, and Jifeng Dai. Uni-Perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2691–2700, 2023.

Yi Li, Hongze Shen, Lexiang Tang, Xin Li, Xinpeng Ding, Yinsong Liu, Deqiang Jiang, Xing Sun, and Xiaomeng Li. DenseMLLM: Standard multimodal LLMs are intrinsic dense predictors, 2026. arXiv preprint arXiv:2602.14134.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Doll'ar. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Icdar 2023 competition on hierarchical text detection and recognition. *arXiv preprint arXiv:2305.09750*, 2023.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations*, 2023.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26439–26455, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018, June 2024.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction, 2021. arXiv preprint arXiv:2109.00512.

Zhongwei Ren, Zhizhong Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26374–26383, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Aadarsh Sahoo and Georgia Gkioxari. Conversational image segmentation: Grounding abstract concepts with scalable supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39476–39485, June 2026.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3260–3269, 2017.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2930–2937, 2013.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, pp. 746–760, 2012.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419. Springer, 2020.

Igor Vasiljevic, Nicholas Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falong Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A dense indoor and outdoor DEpth dataset, 2019. arXiv preprint arXiv:1908.00463.

Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-SAM: From segment anything to any segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26187–26196, 2026a. doi: 10.1609/aaai.v40i31.39822.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, Xiangxiang Chu, and Yao Zhao. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338*, 2025b.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23318–23340. PMLR, 17–23 Jul 2022.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 35928–35959. Curran Associates, Inc., 2025c.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. LocateAnything: Fast and high-quality vision-language grounding with parallel box decoding, 2026b. arXiv preprint arXiv:2605.27365.

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 (CVPR)*, pp. 20697–20709, 2024.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6830–6839, June 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140, October 2023b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977, 2025.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Ping Luo, Yu Qiao, and Jifeng Dai. VisionLLM v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 69925–69975. Curran Associates, Inc., 2024a. doi: 10.52202/079017-2235.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024b.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0688.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. UniTAB: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pp. 521–539. Springer, 2022.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Youtu-VL Team. Youtu-VL: Unleashing visual potential via unified vision-language supervision, 2026. arXiv preprint arXiv:2601.19798.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85, 2016.

Jingbo Zhang, Xiaoyu Li, Hongliang Zhong, Qi Zhang, Yanpei Cao, Ying Shan, and Jing Liao. Humanref-gs: Image-to-3d human generation with reference-guided diffusion and 3d gaussian splatting. *IEEE Transactions on Circuits and Systems for Video Technology*, 35:6867–6880, 2025. doi: 10.1109/TCSVT.2025.3540969.

Zheng Zhang, YeYao Ma, Enming Zhang, and Xiang Bai. PSALM: Pixelwise segmentation with large multi-modal model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 74–91, 2024. doi: 10.1007/978-3-031-72754-2_5.

Canyu Zhao, Yanlong Sun, Mingyu Liu, Huanyi Zheng, Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, and Chunhua Shen. DICEPTION: A generalist diffusion model for visual perceptual tasks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 40108–40155. Curran Associates, Inc., 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):65:1–65:12, 2018.

Lianghui Zhu, Bin Ouyang, Yuxuan Zhang, Tianheng Cheng, Rui Hu, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Li Yu, Wenyu Liu, and Xinggang Wang. LENS: Learning to segment anything with unified reinforced reasoning. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 13952–13960, 2026. doi: 10.1609/aaai.v40i16.38405.

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-Perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16804–16815, 2022.