UnSAMv2: Self-Supervised Learning Enables Segment Anything at Any Granularity

Anonymous authors
Paper under double-blind review

Abstract

The Segment Anything Model (SAM) family has become a widely adopted vision foundation model, but its ability to control segmentation granularity remains limited. Users often need to refine results manually — by adding more prompts or selecting from pre-generated masks — to achieve the desired level of detail. This process can be ambiguous, as the same prompt may correspond to several plausible masks, and collecting dense annotations across all granularities is prohibitively expensive, making supervised solutions infeasible. To address this limitation, we introduce UnSAMv2 which enables segment anything at any granularity without human annotations. UnSAMv2 extends the divide-and-conquer strategy of UnSAM by discovering abundant mask-granularity pairs and introducing a novel granularity control embedding that enables precise, continuous control over segmentation scale. Remarkably, with only $6K$ unlabeled images and 0.02% additional parameters, UnSAMv2 substantially enhances SAM-2, achieving segment anything at any granularity across interactive, whole-image, and video segmentation tasks. Furthermore, we introduce a language-guided segmentation pipeline that treats granularity as a predictable variable by grounding it in text semantics via an external VLM, improving previous SOTA in 1-IoU (40.4 $\rightarrow$ 47.3). Across 11+ benchmarks, UnSAMv2 achieves consistent gains over SAM/SAM-2 baselines, improving NoC_{90} (5.69 $\rightarrow$ 4.75), 1-IoU (58.0 $\rightarrow$ 73.1), and AR_{1000} (49.6 $\rightarrow$ 68.3) by a large margin. Our work demonstrates that a small amount of unlabeled data can deliver large, practical improvements to strong supervised baselines!

1 Introduction

What is an object? This question has long been debated in both vision and cognition. Should an object be defined by a model's learned semantics or by an annotator's subjective judgment? We propose a third perspective: granting users full flexibility to define objectness through a single point prompt and a continuous granularity score.

Imagine clicking on a single point in an image and smoothly adjusting a slider that controls segmentation granularity. At low granularity values, the model reveals fine-grained parts with precise boundaries; as the granularity increases, it gradually merges regions into larger and more semantically coherent entities. In essence, this transforms SAM's (Kirillov et al., 2023) three discrete mask hypotheses into a continuous granularity axis capable of producing multiple masks per point, each corresponding to a user-defined granularity. Such controllable segmentation enables users to flexibly define what constitutes an "object" for their usage, whether it involves part-level analysis, instance grouping, or large-scale region editing.

The emergence of the Segment Anything family (Kirillov et al., 2023; Ravi et al., 2024) has positioned SAM as a vision foundation model, significantly advancing diverse tasks such as video object tracking (Yang et al., 2023; Ye et al., 2025), multi-scale perception (Huang et al., 2025; Wang et al., 2025), and compositional understanding (Doersch et al., 2024; Wang et al., 2023a; Chen et al., 2025). However, SAM and its successors rely heavily on *supervised learning* from SA-1B dataset (Kirillov et al., 2023). This pipeline directly ties the model's notion of an "object" to human annotation bias and limits its output to three discrete mask hypotheses per prompt. As a result, segmentation in SAM is *defined by supervision* rather than *discovered*

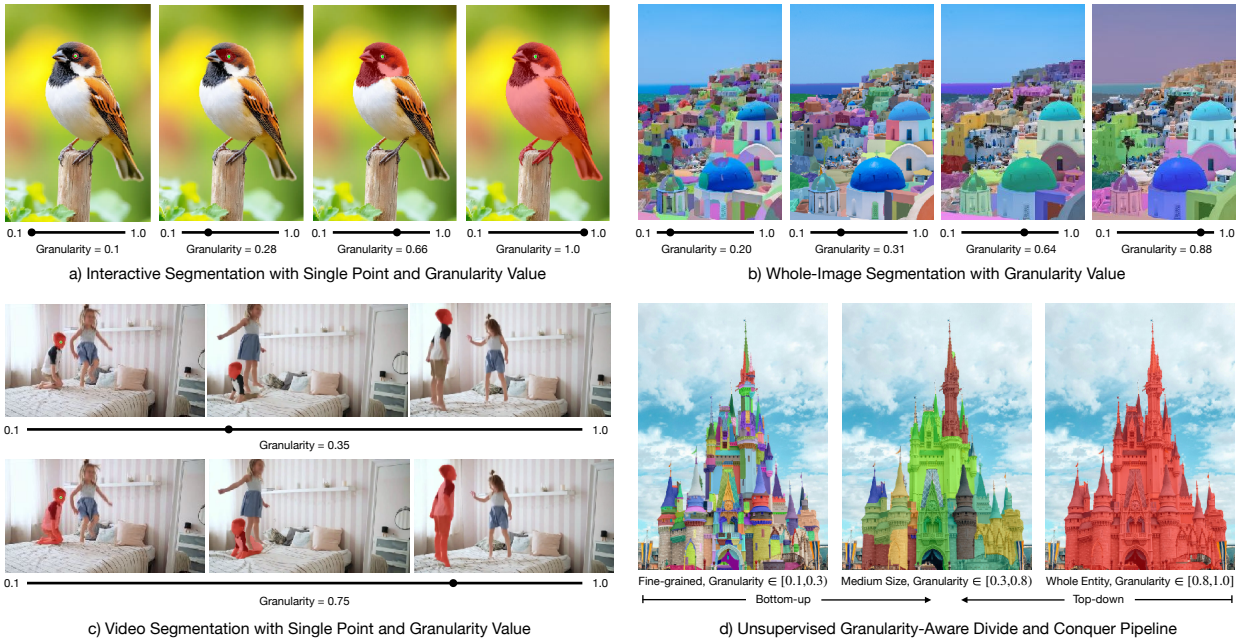
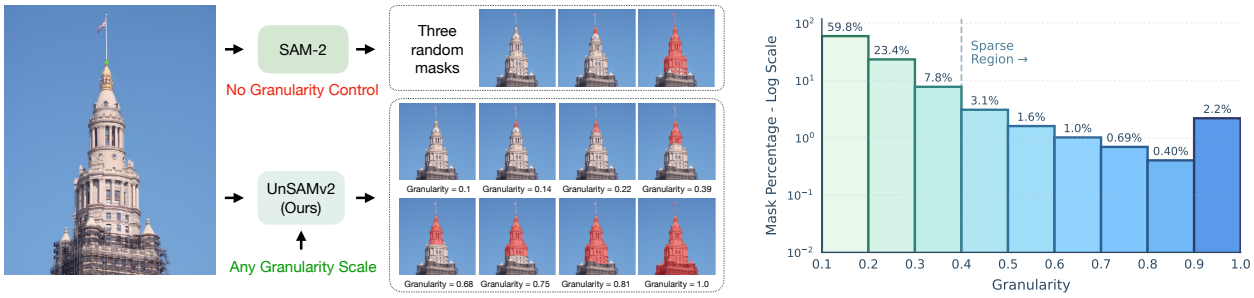


Figure 1: What is an object? The notion has long been debated: should it follow a model's learned semantics or an annotator's subjective judgment? UnSAMv2 takes a third path, granting users full flexibility to define objectness through promptable segmentation with a single point and a continuous, controllable granularity score. Built on SAM-2 (Ravi et al., 2024), UnSAMv2 introduces a granularity-aware, self-supervised training pipeline based on divide-and-conquer pseudo-labels (Wang et al., 2024c). Trained on just 6000 unlabeled images, it segments anything from fine-grained parts to holistic objects, achieving state-of-the-art performance across interactive, whole-image, and video segmentation tasks.



(a) From ambiguity to control **(b)** Granularity distribution

Figure 2: From ambiguity to control, and learning across scales. (a) Without granularity input, SAM-2 may output multiple plausible masks per point; UnSAMv2 introduces a continuous granularity variable to retrieve the intended object at any scale with a single prompt. **(b)** Our unsupervised divide-and-conquer pipeline yields a rich, left-tailed granularity distribution dominated by fine-grained pseudo-masks, yet UnSAMv2 learns stable semantics across all scales.

from data. This design assumes a shallow hierarchy of objectness, while real-world scenes exhibit complex, nested part–whole structures that cannot be captured through fixed human labels. Although SA-1B covers a broad range of object sizes, it lacks explicit correspondences between instance- and part-level masks, making it difficult for supervised models to learn how granularity should vary continuously.

We argue that segmentation granularity should be learned through unsupervised learning, which allows models to infer object hierarchies directly from image statistics instead of depending on predefined labels. To this end, we introduce **UnSAMv2**, a self-supervised framework that enables segmentation at any granularity

without human supervised labels. UnSAMv2 learns a continuous representation of granularity that bridges the gap between parts and wholes, giving users full control over segmentation masks.

UnSAMv2 builds upon UnSAM (Wang et al., 2024c), which introduced an unsupervised divide-and-conquer strategy for discovering hierarchical masks. While UnSAM focused on unsupervised hierarchy construction, **UnSAMv2** extends this idea to *granularity-controllable segmentation*. In the *divide* stage, we use the normalized-cut method MaskCut (Wang et al., 2023b) to extract instance-level masks. In the *conquer* stage, we recursively merge similar pixels within each instance to discover finer parts, forming hierarchical pseudo-labels that encode relative scale. From these hierarchies, we compute a continuous granularity scalar for each mask, representing its position along the part–whole continuum.

We then augment SAM-2 (Ravi et al., 2024) with a lightweight granularity encoder and a granularity-aware mask token. Given a point prompt and a granularity scalar g, UnSAMv2 predicts the mask corresponding to the desired granularity, turning segmentation into a controllable function of scale. Training only the lightweight SAM-2 decoder for four hours with 2 A100 GPUs on 6,000 unlabeled images (0.02% additional parameters) enables smooth interpolation between coarse and fine structures and reveals the latent hierarchy within SAM's feature space.

Across interactive, whole-image, and video segmentation benchmarks, UnSAMv2 consistently surpasses SAM-2 (Ravi et al., 2024) and previous state-of-the-art promptable segmentation methods. Evaluated on more than 11 widely used datasets, including SA-1B, COCO, and PartImageNet, UnSAMv2 improves NoC_{90} from 5.69 to 4.75, 1-IoU from 58.0 to 73.1, and AR_{1000} from 49.6 to 68.3, all achieved using only unlabeled data. The resulting model empowers users to define their own notion of objectness and to explore segmentation as a continuous, controllable process rather than a static prediction. We further demonstrate the practical utility of granularity control when combined with semantic prompts. Inspired by previous language-guided segmentation methods such as LISA (Lai et al., 2024), SA2VA (Yuan et al., 2025), and SAM-3 (Carion et al., 2025), we pair our granularity control with an external VLM that predicts an appropriate granularity value from the user's text prompt. This improves the strongest baseline (Yuan et al., 2025) from 40.41 to 47.29 in 1-IoU, demonstrating that granularity can be effectively grounded in text semantics.

2 Related Work

Multi-Granularity Segmentation. SAM (Kirillov et al., 2023; Ravi et al., 2024; Carion et al., 2025) has greatly advanced segmentation performance by leveraging large-scale human-annotated data and extensive compute. Semantic-SAM (Li et al., 2024) improves fine-grained predictions through a multiple-choice design (Guzman-Rivera et al., 2012; Li et al., 2018). However, these approaches constrain point-prompt predictions to a fixed number of candidate masks, forcing users to manually select from limited outputs or give additional prompts. This restriction highlights the need for explicit control over mask granularity. Recent work (Wang et al., 2024a; Lin et al., 2025; Li et al., 2023a) has begun to tackle such ambiguity. GARField (Kim et al., 2024b) and SAMPart3D (Yang et al., 2024c; 2025) address scale ambiguity in 3D scene decomposition via absolute scale conditioning, while GraCo (Zhao et al., 2024) achieves granularity-controllable interactive segmentation by extending SimpleClick (Liu et al., 2023) with discrete granularity inputs. In contrast, UnSAMv2 tackles mask ambiguity in a fully self-supervised manner by treating granularity as a continuous, relative concept within a hierarchical structure.

Self-Supervised Learning and Unsupervised Segmentation. SSL methods such as MAE (He et al., 2022b), JEPA (Assran et al., 2023), and DINO (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) demonstrate that large-scale pretraining can endow vision transformers with strong representations, benefiting a wide range of downstream tasks (Jiang et al., 2025; Gao et al., 2024; Zhang et al., 2025; Yang et al., 2024a;b; Wu et al., 2025; Karmann & Urfalioglu, 2025). In parallel, unsupervised segmentation has gained lots of attention (Wang et al., 2023c; Li et al., 2023b; Wang et al., 2022; Siméoni et al., 2021; Kim et al., 2024a; Hamilton et al., 2022; Arica et al., 2024; Niu et al., 2024; Zhou et al., 2020; Hahn et al., 2025; Karmann & Urfalioglu, 2025; Sick et al., 2025). CutLER (Wang et al., 2023b; 2024b) greatly advanced unsupervised instance segmentation by using MaskCut (Shi & Malik, 2000) to iteratively extract objects from images. SOHES (Cao et al., 2024) adopts a bottom-up merging scheme, grouping pixels based on cosine similarity to progressively discover objects. Recently, UnSAM (Wang et al., 2024c) introduced a

divide-and-conquer paradigm to generate hierarchical pseudo labels. Building on these efforts, UnSAMv2 leverages the hierarchical mask discovery perspective and extends it to assign an explicit granularity scale for each pseudo mask. This enables unsupervised segmentation at any granularity level of detail explicitly.

Language-Guided Segmentation. Language-guided segmentation (Xia et al., 2024; Ren et al., 2024; Wei et al., 2025; Chen et al., 2024; Bai et al., 2024; Wang et al., 2024d) grounds text semantics into pixel masks by combining VLMs with segmenters. LISA (Lai et al., 2024) decodes a segmentation token from an MLLM into masks, while Sa2VA (Yuan et al., 2025) and SAM-3 (Carion et al., 2025) further integrate SAM-style models with language for image/video prompting. We are the first to treat *granularity* as a predictable variable by introducing a VLM-based pipeline that infers granularity from text semantics and combines it with granularity-controllable segmentation.

3 Method

We present **UnSAMv2**, a self-supervised framework that enables segmentation at arbitrary levels of granularity without human annotations. Unlike the supervised SAM (Kirillov et al., 2023) pipeline, which depends on human-labeled object masks, UnSAMv2 learns granularity directly from image statistics through a hierarchy-aware divide-and-conquer process. This enables segmentation granularity to be continuously controlled by a single scalar input, rather than restricted to a fixed number of discrete mask tokens.

We first review prior unsupervised segmentation methods (Section 3.1) and the limitations of supervised training paradigms (Section 3.2). We then present our granularity-aware divide-and-conquer algorithm that automatically constructs mask–granularity pairs from unlabeled data (Section 3.3). Next, we describe the architectural design that empowers any promptable segmentation model to interpret and control segmentation granularity (Section 3.4). We then briefly discuss the difference with prior supervised learning works (Section 3.5). Then, we present UnSAMv2+, a lightly supervised variant that integrates SA-1B annotations to further refine the granularity learning process (Section 3.6). Finally, we associate granularity with semantics using vision-language models (Section 3.7).

3.1 Preliminaries

UnSAM, MaskCut, and SOHES. UnSAM (Wang et al., 2024c) introduced a divide-and-conquer strategy to generate pseudo masks without supervision. In the *divide* stage, a cut-based segmentation method MaskCut (Wang et al., 2023b) is applied to obtain instance-level masks. Then, inspired by bottom-up hierarchical segmentation, *e.g.*, SOHES (Cao et al., 2024), the *conquer* stage iteratively merges similar pixels under a sequence of thresholds, constructing part–whole hierarchies. Formally, for a local image region I_{local}, patch-level features $K = \text{DINO}(I_{\text{local}})$ are extracted using DINO (Siméoni et al., 2025). Neighboring patches are merged according to the cosine similarity of their DINO features against thresholds $\theta_1, \ldots, \theta_l$, yielding part masks nested inside instances. This process produces a rich hierarchy of mask granularity.

Segment Anything family. Segment Anything models (Kirillov et al., 2023; Ravi et al., 2024; Carion et al., 2025) have advanced promptable segmentation with a scalable encoder–decoder design. **(1) Image encoder:** a ViT that maps an input image to multi-scale dense embeddings while preserving spatial structure. **(2) Prompt encoder:** embeddings for user inputs such as points, boxes, or masks that condition the segmentation process and guide attention to regions of interest. **(3) Mask decoder:** a lightweight transformer that fuses image and prompt features to predict segmentation masks.

Despite these strengths, the training pipeline is fully *supervised* on SA-1B (Kirillov et al., 2023), which ties the notion of objectness to human-labeled masks. Moreover, the decoder employs three fixed mask tokens (small, medium, large), producing at most three hypotheses per prompt. This discretization limits controllable granularity and discourages hierarchical reasoning about parts and wholes, motivating an unsupervised formulation that learns granularity from data rather than from fixed labels.

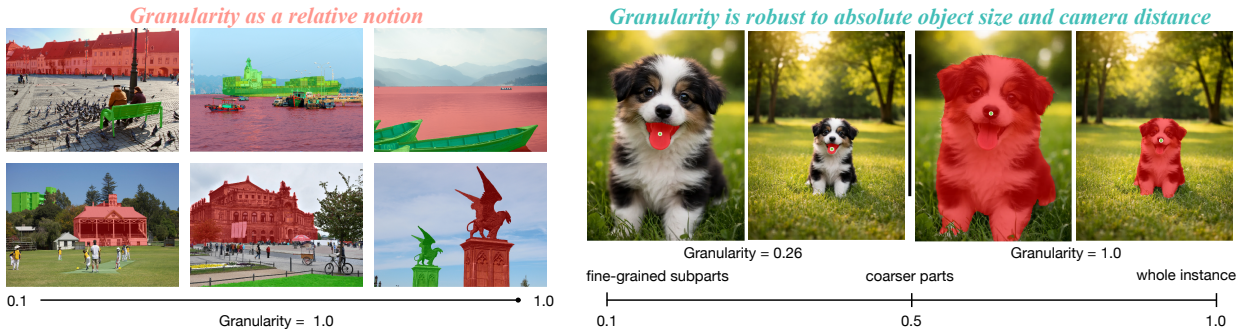


Figure 3: Granularity is a relative notion and robust to camera distance. At a fixed granularity value, mask sizes vary widely across scenes, showing that UnSAMv2 learns granularity relationally, consistent with human perception of parts and wholes rather than simply associating it with absolute size.

3.2 Limitations of SAM's Supervised Paradigm

Lack of granularity control. When a single point corresponds to multiple plausible objects (*e.g.*, a part versus the whole), SAM generates up to three discrete masks and requires manual selection by the user. Without an explicit granularity variable, the model cannot traverse scales continuously—fine details and coarse structures remain disconnected. This limitation not only reduces efficiency in interactive segmentation but also prevents smooth, interpretable control over the level of detail.

Lack of hierarchical reasoning. Supervised training on human-labeled masks encourages SAM to learn a flat object representation, where parts and instance-level segments are treated as isolated entities rather than components within a hierarchy. As a consequence, SAM lacks structural awareness and fails to capture relationships across scales. It struggles to segment scenes at intermediate levels of detail or to uncover the nested hierarchical structure of visual scenes (Fig. 2a). This limitation underscores the necessity of *unsupervised learning*, which can recover hierarchical dependencies directly from image statistics rather than relying on human annotations.

3.3 Granularity-Aware Divide-and-Conquer

We extend UnSAM's divide-and-conquer framework to automatically construct dense mask–granularity pairs from unlabeled images. The process consists of four stages.

Stage 1: Instance discovery via N-Cuts. We employ the normalized-cut-based method CutLER (Wang et al., 2023b) to generate segmentation masks $\mathcal{M} = \{m_1, \ldots, m_n\}$ at varying levels of granularity. We filter out noisy CutLER mask outputs with a confidence threshold τ_{conf},

$$\mathcal{M}_{\text{high}} = \{m_i \mid \text{conf}(m_i) \geq \tau_{\text{conf}}\}, \forall i \in [n]. \tag{1}$$

Stage 2: Instance–part relationship discovery. We identify instance-level masks $\mathcal{M}_{\text{inst}} \subseteq \mathcal{M}_{\text{high}}$ based on two criteria: (i) their area-to-image-area ratio exceeds the threshold τ_{area}, and (ii) they dominate overlapping masks, which means for any $m_i \in \mathcal{M}_{\text{inst}}$ and $m_j \in \mathcal{M}_{\text{high}}$, we have

$$\text{IoU}(m_i, m_j) \geq \tau_{\text{overlap}} \Rightarrow \text{Area}(m_i) \geq \text{Area}(m_j). \tag{2}$$

The remaining masks form $\mathcal{M}_{\text{rest}} = \mathcal{M}_{\text{high}} \setminus \mathcal{M}_{\text{inst}}$. Each $m_r \in \mathcal{M}_{\text{rest}}$ is assigned as a part of m_i if $\text{IoU}(m_r, m_i) > \tau_{\text{overlap}}$; otherwise, it is discarded. Therefore, for every instance mask m_i, we have a set of part-level masks $\mathcal{M}_{i,\text{part}}$.

Stage 3: Fine-grained mask discovery. To enrich granularity, we further apply patch merging to each $m_i \in \mathcal{M}_{\text{inst}}$, decomposing it into finer structures. The resulting fine-grained masks $\mathcal{M}_{i,\text{conquer}}$ are merged with $\mathcal{M}_{i,\text{part}}$ through Non-Maximum Suppression (NMS), denoted as $\mathcal{M}_{i,\text{final}}$. This process expands granularity richness in a hierarchical perspective while maintaining mask quality.

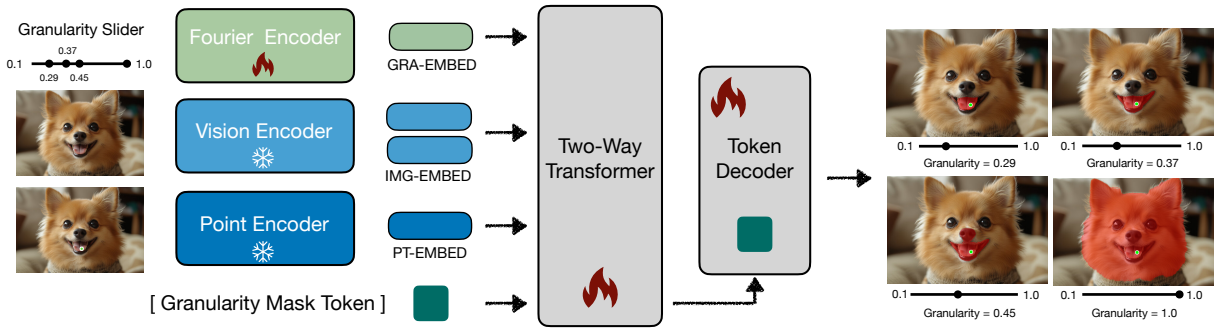


Figure 5: Architecture of UnSAMv2. Built on SAM-2, UnSAMv2 introduces a Fourier-based granularity encoder and a mask token to enable segmentation at any granularity. A scalar granularity input $g \in [0.1, 1]$ is mapped to a high-dimensional embedding via Fourier transformation and an MLP, then injected into the transformer alongside the sparse point embedding and dense image embedding. The granularity-aware mask token attends to image, point, and granularity embeddings, and is finally decoded by a token decoder into a mask at the requested granularity.

Stage 4: Continuous granularity assignment. After obtaining the instance and part hierarchies, we assign each mask $m_i \in \mathcal{M}_{i,\text{final}}$ a continuous granularity score $g_i \in [0.1, 1.0]$ according to its relative area within the hierarchy:

$$g_i = \left(\frac{\sqrt{A_i} - \sqrt{A_{\min}}}{\sqrt{A_{\max}} - \sqrt{A_{\min}}} \right) \cdot 0.9 + 0.1, \tag{3}$$

where A_i demonstrates the area of the mask m_i in pixel space, and $A_{\min}$ and $A_{\max}$ correspond to the smallest and largest masks in the current mask hierarchy.

As shown in Fig. 2(b), our divide-and-conquer pipeline generates a dense hierarchy of masks spanning fine to coarse scales. We find that most masks have $g_i < 0.4$, indicating an abundance of part-level structures; yet UnSAMv2 learns stable semantics across all scales, and our granularity is robust to absolute object sizes (Fig. 3). We quantify the relationship between granularity and absolute mask size in Fig. 4 with our unsupervised masks and granularity data in 6,000 images. As granularity increases, the mean mask area grows, showing effective granularity controllability; the variance widens because instance-level masks can differ greatly in area across instances.

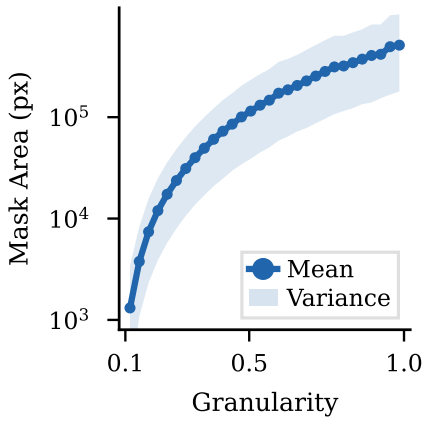


Figure 4: Quantitative analysis on granularity and absolute mask areas.

3.4 UnSAMv2 Architecture

Granularity encoding. As shown in Fig. 5, we introduce a lightweight granularity encoding module that converts a scalar g into a high-dimensional Fourier embedding (Tancik et al., 2020), denoted as $\phi(g) \in \mathbb{R}^{d_{\text{Fourier}}}$. Then, we project $\phi(g)$ into the decoder feature space using a three-layer MLP: $E_g = \text{MLP}(\phi(g)) \in \mathbb{R}^{d_{\text{decoder}}}$. This granularity embedding is concatenated with the point prompt features E_p: $E_{\text{prompt}} = \text{Concat}(E_p, E_g)$. This design allows the mask decoder to jointly interpret spatial prompts and granularity cues.

Granularity-aware mask decoding. We replace SAM's three fixed-size mask tokens with a single learnable *granularity-aware mask token*. This token attends to both prompt embeddings and image features through self- and cross-attention, producing a mask with the target granularity. The design is model-agnostic and can be integrated with any promptable segmentation framework.

Parameter efficiency. The encoding and token modules introduce less than 0.02% extra parameters while providing continuous mask granularity control.

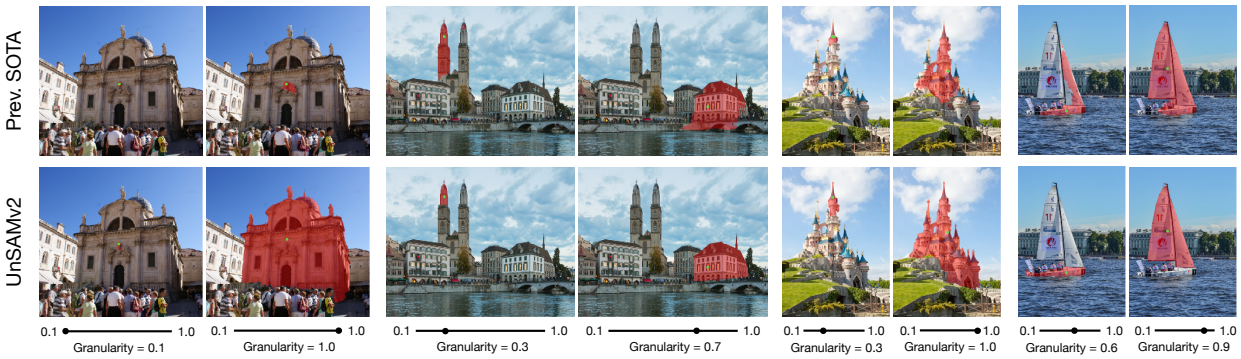


Figure 6: Qualitative comparison with the previous state-of-the-art method (Zhao et al., 2024). Each scene shows results at different target granularity values. Prior methods often break one object into parts at high granularity or include extra regions at low granularity. In contrast, UnSAMv2 produces clear and consistent masks with smooth transitions across scales.

3.5 Distinctions with prior methods.

Prior methods such as GraCo (Zhao et al., 2024) and GARField (Kim et al., 2024b) treat granularity as a discrete variable or directly associate it with absolute mask size. However, discretizing granularity imposes artificial boundaries between adjacent levels, preventing smooth transitions across scales, while absolute-size definitions fail to account for contextual differences, *e.g.*, a "small" mask could correspond to an entire object in one image but only a part in another. In contrast, UnSAMv2 formulates granularity as a ***continuous, relative*** measure within the instance–part hierarchy, aligning more closely with human perception, where objects and parts are interpreted in relation to one another rather than by their absolute scales. This formulation enables UnSAMv2 to capture fine-grained hierarchical structure and traverse segmentation levels seamlessly.

3.6 UnSAMv2+: Light-Supervised Variant

Although the unsupervised pipeline generates abundant mask–granularity pairs, pseudo labels can be noisy. To further stabilize learning, we introduce **UnSAMv2+**, a lightly supervised variant that incorporates SA-1B ground-truth masks into the divide stage: $\mathcal{M}_{\text{UnSAMv2+}} = \mathcal{M}_{\text{CutLER}} \cup \mathcal{M}_{\text{SA-1B}}$. The combined masks are then processed through the same conquer and granularity-assignment stages as in UnSAMv2. This hybrid supervision effectively balances human-curated and self-discovered hierarchies, producing cleaner masks while preserving the scalability and flexibility of unsupervised granularity learning.

3.7 Language-Guided Segmentation with Granularity Control

Inspired by previous work in language-guided segmentation, *e.g.* LISA (Lai et al., 2024), Sa2VA (Yuan et al., 2025), and SAM-3 (Carion et al., 2025), we integrate an external vision-language model (VLM) to our granularity-controllable segmentation pipeline. Intuitively, we combine UnSAMv2 with language since the text semantics also reflect object granularity. The VLM acts as a semantic bridge that predicts the appropriate granularity from a natural point–text prompt. As shown in Fig. 7, the user provides an image I, a click p, and a text description t of the target, *e.g.* "the cat's left eye." We query a VLM-based predictor ϕ to infer a scalar granularity g, and then prompt UnSAMv2 with (p, g) to produce the final mask:

$$g = \phi(I, p, t), \qquad \hat{m} = f_\theta(I, p, g), \tag{4}$$

where f_θ denotes our granularity-controllable segmentation model.

To train ϕ, we leverage the pretrained VLM and our unsupervised mask, granularity pairs $\{(m_i, g_i)\}$ from our divide-and-conquer pipeline in Sec 3.3. For each mask m_i, we sample a point $p_i \in m_i$ and associate it with a short phrase t_i describing the region, yielding training tuples (I, p_i, t_i, g_i). We extend the VLM vocabulary with a special `<gra>` token. During training, the model is prompted to produce a response ending

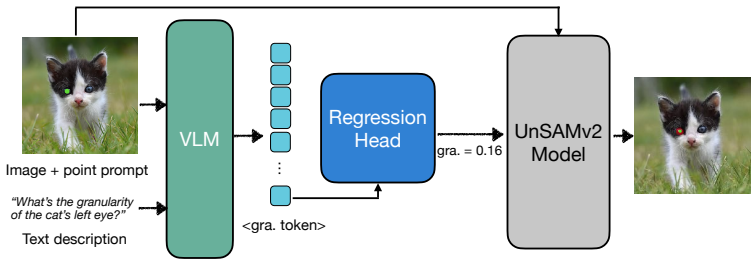


Figure 7: Language-guided segmentation. Conditioned on the image and the text description, the VLM infers a granularity value, which is then provided to UnSAMv2 for segmentation.

with <gra>, and we attach a lightweight regression head h_θ, a two-layer MLP with a sigmoid output, to the hidden state $\mathbf{z}$ at the <gra> token position. The predicted granularity is $\hat{g} = h_\theta(\mathbf{z})$, and we supervise it with the hierarchy-defined granularity label:

$$\mathcal{L}\text{gra} = \left\| h_\theta(\mathbf{z}\texttt{<gra>}) - g_i \right\|_1 , \tag{5}$$

where $\mathbf{z}$<gra> is the VLM's last-layer hidden representation at the <gra> position, which encodes the full image–point–text context through causal attention. We train the LoRA adapters and h_θ, while the VLM backbone remains frozen. At inference time, the model performs single-shot point-and-text segmentation by autoregressively generating up to the <gra> token, extracting the predicted granularity $\hat{g}$ from the regression head, and then feed the predicted granularity to UnSAMv2 to derive desired segmentation result.

4 Experiments

4.1 Model Training Settings

Pseudo mask-granularity data generation. For the divide stage of our unsupervised data pipeline, we set τ_{conf} to 0.3 and τ_{overlap} to 0.8. In the conquer stage, we leverage DINOv3 (Siméoni et al., 2025) vision transformer to extract feature embedding and set the iterative merging thresholds as $\theta = [0.9, 0.8, 0.7, 0.6, 0.5]$. We run the unsupervised pipeline to generate mask-granularity pairs on 6,000 images from SA-1B. **UnSAMv2 training.** We adopt SAM-2-small as the base model, and with only 8 A-100 GPU hours, we finetune the base model for 5 epochs with 6,000 unlabeled images. During training, we freeze the vision encoder and train the granularity encoding module, granularity mask token, and the two-way transformer block in the mask decoder with LoRA (Hu et al., 2022). **VLM training.** We adopt Qwen2.5-VL-7B as our base model and use image-text labels from PascalPart and PartImageNet as our training data. We train the base model for 2 epochs with LoRA. We provide more experimental details in the appendix.

4.2 Evaluation Datasets and Metrics

Interactive Image Segmentation. Following SimpleClick (Liu et al., 2023) and GraCo (Zhao et al., 2024), we evaluate UnSAMv2's capability to segment objects at various granularity levels with the number of clicks (NoC) required to reach a certain Intersection over Union (IoU) threshold to assess the model's efficiency in segmenting the target object. Specifically, we choose NoC_{80} and NoC_{90} which record the average number of clicks needed to achieve 80% and 90% IoU between predicted mask and ground-truth mask. In addition, we measure the IoU between the predicted mask and the ground-truth mask with just one click, denoted as 1-IoU in the table. For object levels, we conduct evaluations on 5 commonly used datasets (Rother et al., 2004; McGuinness & O'connor, 2010; Perazzi et al., 2016; Hariharan et al., 2011; Kirillov et al., 2023), and for part level, we evaluate models on (Chen et al., 2014; Deng et al., 2009). Note that for the fairness of comparison, we only compare datasets across models that are not in-distribution with their training data, e.g., SAM-2 (Ravi et al., 2024) is trained on SA-1B, SimpleClick (Liu et al., 2023) and GraCo (Zhao et al., 2024) are trained with SBD and PascalPart. We provide descriptions of evaluation datasets in the appendix.

Whole-Image Segmentation. We evaluate our model's performance in identifying all possible masks for given images in a zero-shot setting across 5 datasets that span a broad range of granularity (Lin et al., 2014; Gupta et al., 2019; Zhou et al., 2019; Qi et al., 2022; Kirillov et al., 2023). Importantly, each benchmark annotates only specific semantic classes and typically emphasizes certain levels of the part–instance hierarchy,

Datasets →	Averaged			GrabCut			Berkeley			SBD			PascalPart			PartImageNet		
Method	$NoC_{80}\downarrow$	$NoC_{90}\downarrow$	1-IoU	NoC_{80}	NoC_{90}	1-IoU	NoC_{80}	NoC_{90}	1-IoU	NoC_{80}	NoC_{90}	1-IoU	NoC_{80}	NoC_{85}	1-IoU	NoC_{80}	NoC_{85}	1-IoU
SAM-2 (Ravi et al., 2024)	4.02	5.69	58.0	1.48	1.54	76.8	1.29	1.72	80.0	2.85	6.73	64.6	9.37	11.80	29.0	5.09	6.64	39.5
UnSAMv2*	**3.41**	**4.75**	**73.1**	**1.16**	**1.30**	**91.4**	**1.10**	**1.60**	**90.8**	**2.75**	**5.97**	**74.8**	**7.84**	**9.59**	**51.5**	**4.21**	**5.29**	**57.1**
vs. sup. SAM-2	**-0.61**	**-0.94**	**+15.1**	**-0.32**	**-0.24**	**+14.6**	**-0.19**	**-0.12**	**+10.8**	**-0.10**	**-0.76**	**+10.2**	**-1.53**	**-2.21**	**+22.5**	**-0.88**	**-1.35**	**+17.6**
UnSAMv2+*	3.30	4.56	74.7	1.22	1.26	92.6	1.12	1.37	91.1	2.55	5.72	77.3	7.87	9.67	52.0	3.75	4.77	60.7

Table 1: Comparison with SAM-2 on interactive segmentation benchmarks. Trained on 6,000 images with purely unsupervised pseudo-labels, UnSAMv2 significantly outperforms SAM-2. We report promptable segmentation performance in terms of NoC and 1-IoU across five benchmarks. * Following (Zhao et al., 2024), we select the optimal granularity from 0.1 to 1.0 in steps of 0.1 and report averaged results.

Datasets →	Averaged			GrabCut			Berkeley			DAVIS			SA-1B			PartImageNet		
Method	$NoC_{80}\downarrow$	$NoC_{90}\downarrow$	1-IoU	NoC_{80}	NoC_{90}	1-IoU	NoC_{80}	NoC_{90}	1-IoU	NoC_{80}	NoC_{90}	1-IoU	NoC_{80}	NoC_{90}	1-IoU	NoC_{80}	NoC_{85}	1-IoU
SimpleClick (Liu et al., 2023)	3.32	4.87	60.2	1.32	1.48	89.6	1.26	2.44	84.6	2.88	5.38	75.6	4.80	7.27	21.9	6.34	7.76	29.2
GraCo (Zhao et al., 2024)*	2.35	3.42	74.4	1.24	1.32	91.1	1.17	1.63	88.7	2.84	5.09	74.0	2.41	4.01	66.2	4.11	5.03	51.7
SAM-2 (Ravi et al., 2024)	2.44	3.63	69.0	1.48	1.54	76.8	1.29	1.72	80.0	2.51	4.50	71.7	**1.82**	3.76	77.1	5.09	6.64	39.5
UnSAMv2*	2.28	3.40	79.3	**1.16**	1.30	91.4	**1.10**	1.60	90.8	2.75	4.66	84.7	2.18	4.16	72.4	4.21	5.29	57.1
UnSAMv2+*	**2.07**	**3.10**	**81.7**	1.22	**1.26**	**92.6**	1.12	**1.37**	**91.1**	**2.36**	**4.40**	**85.7**	1.87	**3.69**	**78.4**	**3.75**	**4.77**	**60.7**
vs. prev. SOTA	**-0.28**	**-0.32**	**+7.30**	**-0.02**	**-0.06**	**+1.50**	**-0.05**	**-0.26**	**+2.40**	**-0.48**	**-0.69**	**+11.70**	**-0.54**	**-0.32**	**+12.20**	**-0.36**	**-0.26**	**+9.00**

Table 2: State-of-the-art performance on interactive segmentation across benchmarks. Trained with only 6,000 images using a combination of supervised and unsupervised labels, UnSAMv2+ demonstrates superior segmentation quality and indicates how effectively unsupervised methods can complement supervised data. Results of (Liu et al., 2023; Zhao et al., 2024) are reproduced with official code and checkpoints. *: Following (Zhao et al., 2024), we select optimal granularity from 0.1 to 1.0 with a step of 0.1 and report average results.

whereas our method predicts masks at all levels and for any class. Consequently, COCO Average Precision (AP) does not faithfully capture open-world performance. Following prior work (Wang et al., 2023b; Cao et al., 2024; Wang et al., 2024c), we report Average Recall (AR_{1000}) for comparison across methods.

Language-Guided Segmentation. To evaluate our text-guided segmentation pipeline in Section 3.7, we test UnSAMv2 augmented with an external VLM on PartImageNet (He et al., 2022a). For each test image, we take the center point of the ground-truth mask and its text label and query the VLM to predict a granularity value $g \in [0.1, 1.0]$. We then prompt UnSAMv2 with the same center point and the predicted granularity to produce the segmentation mask and report 1-IoU. For fair comparison, we provide the same center-point and text-label inputs to the baselines LISA (Lai et al., 2024) and Sa2VA (Yuan et al., 2025).

4.3 Experimental Results

Interactive Segmentation. Remarkably, UnSAMv2 surpasses SAM-2 across all datasets in zero-shot evaluation settings as summarized in Table 1. Finetuned with only 6,000 images with unsupervised pseudo-labels, UnSAMv2 demonstrates superior performance in segmenting objects at various granularity levels. It indicates that our model architecture and unsupervised data pipeline effectively guide the base model to understand the meaning of granularity scaler. On average, UnSAMv2 surpasses SAM-2 by 15.2% in NoC_{80}, 16.5% in NoC_{90}, and 26.0% in 1-IoU. With UnSAMv2, user could segment their desired objects accurately at any granularity level with fewer prompt points, which greatly enhances the flexibility of object segmentation and prepares for downstream tasks. In addition, as shown in Tab. 2, UnSAMv2+, trained with a combination of supervised and unsupervised labels on 6,000 images achieves state-of-the-art performance on all metrics across multiple evaluation datasets.

Qualitative comparisons with previous SOTA (Zhao et al., 2024) are shown in Fig. 6. For fair comparison with the three mask tokens design in SAM-2, we additionally report results using three granualrity scales 0.2, 0.5, and 1.0 in Table 3.

Whole-Image Segmentation. Apart from SOTA results in interactive segmentation, UnSAMv2 achieves superior performance on whole-image segmentation across datasets with instances of abundant granularity levels, outperforming SAM (Kirillov et al., 2023) by 37.7% and UnSAM (Wang et al., 2024c) by 29.8% in AR_{1000}. The results are shown in Table 4. With a granularity scalar as input, UnSAMv2 can surface instances over a wide range of detail. As shown in Fig. 8, users simply set the desired granu-

Method	$NoC_{80}\downarrow$	1-IoU
SAM-2 (3 tok.)	4.02	58.0
UnSAMv2 (3 gra.)	3.76	66.7
UnSAMv2 (10 gra.)	**3.41**	**73.1**

Table 3: Evaluation with three granularity scales.

Figure 8: Visualizations for whole-image segmentation. Low granularity reveals fine parts, while higher values recover whole objects. UnSAMv2 offers controllable whole-image segmentation, even for scenes with many densely packed entities.

Methods	COCO	LVIS	ADE	Entity	SA-1B
SAM (Kirillov et al., 2023)	49.6	46.1	45.8	45.9	60.8
UnSAM (Wang et al., 2024c)	40.5	37.7	35.7	39.6	41.9
UnSAM+ (Wang et al., 2024c)	52.2	50.8	45.3	49.8	64.8
UnSAMv2	74.5	63.8	68.4	64.3	70.6
UnSAMv2+	**79.1**	**70.3**	**72.7**	**70.3**	**77.9**
vs. prev. SOTA	**+26.9**	**+19.5**	**+27.4**	**+20.5**	**+13.1**

Table 4: State-of-the-art performance on whole image segmentation. The evaluation metric is AR_{1000}. For UnSAMv2, we aggregate masks at granularity from 0.1 to 1.0 in step of 0.1 and filter out low-confidence masks.

Methods	# Images	$NoC_{80\downarrow}$	1-IoU
SAM-2	–	5.09	39.5
UnSAMv2	1K	4.48	55.5
UnSAMv2	3K	4.27	55.7
UnSAMv2	6K	4.21	57.1

Table 5: Ablations for training data size. UnSAMv2 starts to demonstrate a decent understanding of granularity when trained with only 1,000 images with unsupervised pseudo-labels. We report metrics on PartImageNet.

larity to obtain all candidate masks in images at that level, creating new possibilities on how to integrate segmentation models into vision task pipelines.

Video Segmentation Results. Despite being trained solely on images, UnSAMv2 demonstrates promising capabilities on interactive video segmentation. Although we keep SAM-2's memory module frozen during training, UnSAMv2 still achieves competitive results on video data, demonstrating that the granularity embeddings and mask token propagate across frames effectively. This further shows that our self-supervised pipeline enables granularity to be assimilated into pretrained model's reasoning flow. We show qualitative video results in Fig. 9.

Language-Guided Segmentation Results. By integrating an external VLM into our granularity-controllable segmentation pipeline, we leverage the strong correlation between granularity and text semantics to further unlease the potential of UnSAMv2. The VLM can reason on what a reasonable granularity scale is given a short text description of the desired object or part and works as a bridge between users' text prompts and granularity scales. By combining UnSAMv2 with an external VLM, we achieve promising results compared with widely used language segmentation models (Lai et al., 2024; Yuan et al., 2025) as demonstrated in Table 6.

5 Ablations

Efficiency of UnSAMv2 training. In Table 5, we ablate the training data size used for UnSAMv2. We find that granularity training is highly sample-efficient: UnSAMv2 already shows a solid grasp of granularity with only 1,000 images. We attribute this efficiency to two factors. First, we derive the granularity scale hierarchically, mirroring how humans perceive scale—by relative size within a hierarchy rather than absolute size. Second, during training we update only 0.1% of the parameters and keep the rest frozen, which preserves pretrained model's segmentation ability. Overall, our procedure effectively teaches pretrained model the concept of mask granularity: with just a few example images, UnSAMv2 learns a latent representation of granularity rather than merely memorizing instances.

Granularity-aware mask token. We observe that training UnSAMv2 directly with the original mask tokens and token decoding module in SAM-2 leads to unsatisfying performance (Fig. 10). This phenomenon has been noticed in HQ-SAM (Ke et al., 2023). It indicates SAM-2's original mask tokens have already shown strong-prior knowledge on what constitutes as an object. Teaching these pretrained tokens to understand the meaning of granularity is difficult. Thus, we introduce a new mask token and its MLP to do mask decoding

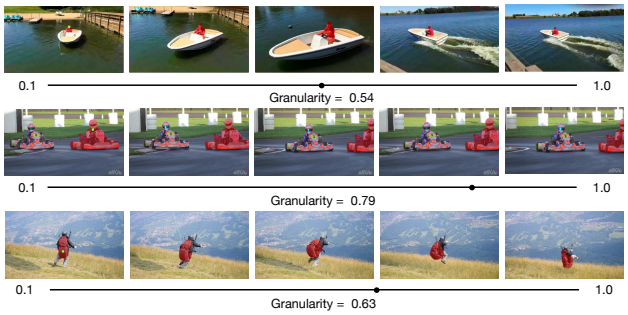


Figure 9: Granularity generalizes to video. We prompt UnSAMv2 with a point and granularity value on the first frame, then propagate the mask to later frames. Even though trained only on images, UnSAMv2 maintains consistent masks over time, showing strong temporal coherence and transferability.

Method	Base VLM	1-IoU
LISA	LLaVA 1.5	26.31
Sa2VA	Qwen2.5-VL-7B	40.41
UnSAMv2	LLaVA 1.5	43.73
UnSAMv2	Qwen2.5-VL-7B	45.07
UnSAMv2 tuned	Qwen2.5-VL-7B	**47.29**

Table 6: Evaluation results on language-guided segmentation. We use the PartImageNet (He et al., 2022a) test set, querying LISA (Lai et al., 2024), Sa2VA (Yuan et al., 2025), and UnSAMv2 with the ground-truth text label and object center point. Without any finetuning, the external VLMs interpret the granularity scale effectively, suggesting a strong alignment between our granularity and semantics.

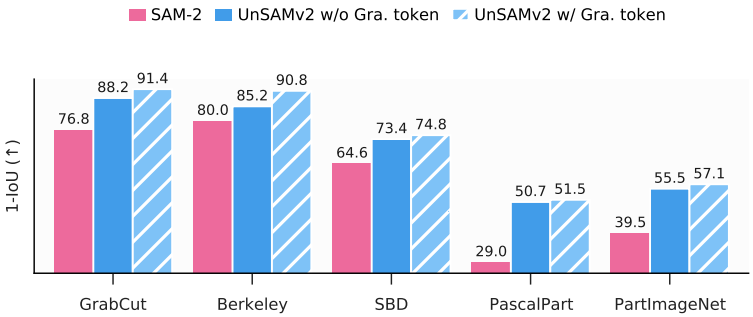


Figure 10: Ablation of granularity-aware mask token. Directly finetuning original SAM-2 mask tokens leads to limited gains, suggesting they already encode strong priors. Adding our granularity-aware token enables efficient learning of granularity.

Methods	Sup. Data	Unsup. Data	1-IoU PascalPart	1-IoU PtIn
SAM-2	–	–	29.0	39.5
UnSAMv2	✓	✗	42.5	54.3
UnSAMv2	✗	✓	51.5	57.1
UnSAMv2+	✓	✓	52.0	60.7

Table 7: Ablation for supervised granularity training. Training with only SA-1B supervised labels underperforms using data from our unsupervised pipeline, highlighting the importance of pseudo-labels for learning granularity.

that is only trained with mask data accompanied with granularity scales. By conducting self-attention with the granularity embedding encoded by Fourier module and cross-attention with image embeddings, the newly introduced token learn the representation of masks and their corresponding granularity simultaneously. The newly introduced token is a crucial component of UnSAMv2 and a key component to achieve the goal – segment anything at any granularity.

Unsupervised pseudo-labels are pivotal for granularity training, as shown in Table 7. We observe that when trained solely on human-labeled SA-1B annotations, UnSAMv2 performs unsatisfactorily compared with training that also incorporates pseudo-labels from our divide-and-conquer pipeline. This points to an inherent issue with supervised datasets: they are heavily biased by human labelers' notions of objects. By contrast, our unsupervised approach focuses on intrinsic relationships among patches, enabling coverage of both instances and parts hierarchically.

6 Conclusion

We presented UnSAMv2, a self-supervised framework that equips pretrained segmentation model to segment anything at any granularity. By deriving continuous granularity scales from unlabeled data, UnSAMv2 learns to traverse part–whole hierarchies and control segmentation with a single scalar. Trained on only 6K unlabeled images, it achieves state-of-the-art results across interactive, whole-image, and video segmentation. We further combine our granularity pipeline in language-guided segmentation, demonstrating the wide usage of our granularity pipeline. Our results highlight that self-supervised learning can unlock latent hierarchical structure in vision foundation models, transforming segmentation from discrete prediction into continuous, controllable reasoning.

A1 Evaluation Datasets

Interactive Segmentation. Specifically, we adopt the following 7 datasets as benchmarks.

- **GrabCut** (Rother et al., 2004) consists of 50 images, each containing a single instance.
- **Berkeley** (McGuinness & O'connor, 2010) includes 96 images with 100 instances, some of which are more challenging for segmentation.
- **DAVIS** (Perazzi et al., 2016) contains 50 high-quality videos and we use 345 frames for evaluation.
- **SA-1B** (Kirillov et al., 2023) contains 11 million high-resolution images (average size 3300×4950 pixels) and 1.1 billion high-quality segmentation masks. For evaluation, we randomly choose 1,000 images that are not included in UnSAMv2 training data as our evaluation set.
- **SBD** (Hariharan et al., 2011) includes 8,498 training images (with 20,172 instances) and 2,857 validation images (with 6,671 instances).
- **PascalPart** (Chen et al., 2014) provides part-level annotations for 20 classes from PascalVOC, totaling 193 part categories. We evaluate models on the whole validation set.
- **PartImageNet** (Deng et al., 2009) organizes 158 ImageNet classes into 11 super-categories and defines 40 distinct part categories. We utilize its validation set to evaluate model's performance on segmenting part-level objects, which consists of 1,206 images and 4,553 annotated parts.

Whole-Image Segmentation. We adopt the following 5 datasets to evaluate UnSAMv2's capability on discovering all instances in images.

- **COCO** (Common Objects in Context) (Lin et al., 2014) is a widely utilized object detection and segmentation dataset. It consists of 115,000 labeled training images and 5,000 labeled validation images. We evaluate our model on COCO `Val2017` with 5000 validation images in a zero-shot manner. We use averaged recall (AR_{1000}) as the metrics for the whole-image segmentation task.
- **SA-1B** (Kirillov et al., 2023) consists of 11 million high-resolution images and 1.1 billion segmentation masks. Following interactive segmentation, we randomly choose 1,000 images that are not included in UnSAMv2 training procedure as evaluation set.
- **LVIS** (Large Vocabulary Instance Segmentation) (Gupta et al., 2019) has 164,000 images with over 1,200 categories and 2 million high-quality instance-level segmentation masks. It covers a large number of object categories. We evaluate UnSAMv2 using its 5000 validation images.
- **EntitySeg** (Qi et al., 2022) is an open-world, class-agnostic dataset with 33,277 images, averaging 18.1 annotated entities per image. We conduct zero-shot evaluation on 1,314 low-resolution images in the validation set.
- **ADE20K** (Zhou et al., 2019) contains 25,574 training and 2,000 testing images covering 365 scenes, emphasizing semantic-level segmentation. It provides labels for 150 semantic categories and 707,868 objects drawn from 3,688 categories. We evaluate zero-shot whole-image segmentation performance on the 2,000-image test split.

Language-Guided Segmentation. We use the test of PartImageNet (He et al., 2022a) as our evaluation data. For each sample, the VLM receives the image and a text description of the target part and predicts a granularity score. A center click point is simulated on the ground-truth mask, and the predicted granularity along with the point prompt are passed to UnSAMv2. We report mean IoU at 1-click.

A2 Training Details.

Pseudo mask-granularity data generation. For the divide stage of our unsupervised data pipeline, we set τ_{conf} to 0.3 and τ_{overlap} to 0.8. In the conquer stage, we leverage DINOv3 (Siméoni et al., 2025) ViT-B/16 backbone to extract feature embedding from the last layer of vision transformer and merge adjacent patches together based on predefined cosine similarity thresholds $\theta = [0.9, 0.8, 0.7, 0.6, 0.5]$. We run the pipeline to generate mask-granularity pairs on 6,000 images from SA-1B in a fully unsupervised manner. On average, we have 112 pseudo-labels on each image. Note that unlike UnSAM (Wang et al., 2024c), UnSAMv2 is designed to learn instance–part relationships rather than only instance representations, so our granularity-aware divide-and-conquer pipeline intentionally produces fewer pseudo-labels than UnSAM, which produces 448 pseudo-labels per image.

UnSAMv2 Training. We finetune SAM-2-small model for 5 epochs with pseudo-labels on 6,000 images. During finetuning, we freeze the heavy-weight Hiera image encoder and only train the granularity encoding module, granularity mask token, and the two-way transformer block in the mask decoder. For the granularity encoder, we first adopt Fourier transformation to granularity scalar with dimension $d_{\text{Fourier}} = 128$ and followed by a 3-layer MLP. Following SAM-2 (Ravi et al., 2024), UnSAMv2 adopts a combination of focal loss and dice loss with a ratio of 20:1. The batch size is set to 4, with the learning rate initialized at 1e-4. We apply LoRA technology to all projection layers of the transformer, setting the LoRA rank to 8. All experiments are conducted on either 2 A-100 or 4 RTX 3090 GPUs.

VLM training. We construct the granularity training dataset from PascalPart and PartImageNet. Each sample consists of an image, a natural language description of the target part (e.g., "head of the person"), and a ground-truth granularity score in [0.1, 1.0] calculated with our divide-and-conquer pipeline. We adopt Qwen2.5-VL-7B-Instruct as the VLM backbone. The model is trained with LoRA with rank set to 16. A special `<gra>` token is appended to each response; its last-layer hidden state is fed through a lightweight regression head to produce the granularity prediction. The model is optimized with AdamW using a cosine learning rate schedule. All training is conducted on 2 A-100 GPUs.

A3 More Visualizations

We present more UnSAMv2's qualitative results on interactive image segmentation in Fig. A1, whole image segmentation in Fig. A2, and video segmentation in Fig. A3.

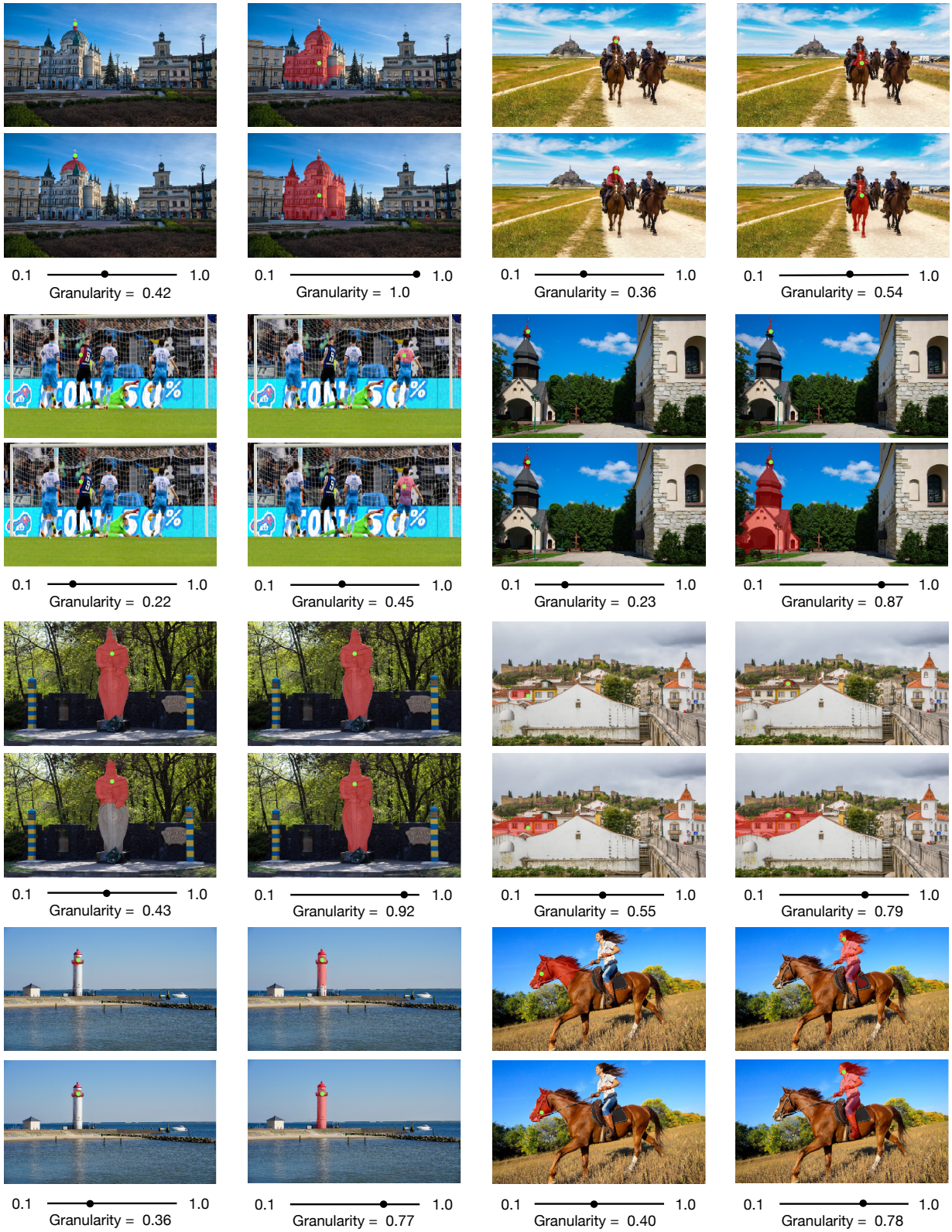


Figure A1: Interactive segmentation results on SA-1B. The top row is previous SOTA GraCo (Zhao et al., 2024) and the bottom row is UnSAMv2.

Figure A2: Whole image segmentation on SA-1B. From top to bottom are raw images, segmentation by SAM-2 and UnSAMv2.

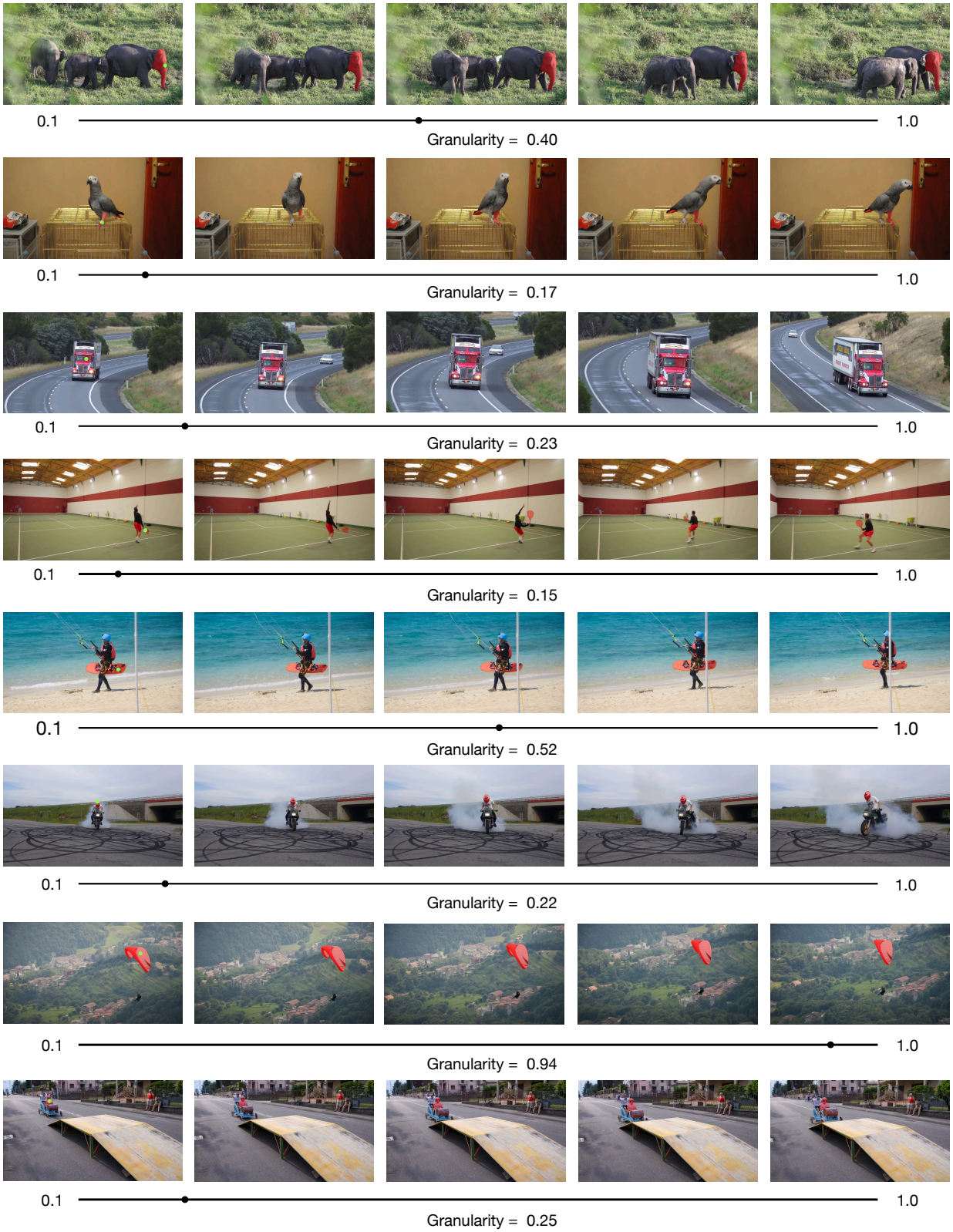


Figure A3: UnSAMv2's interactive video segmentation results on YoutubeVIS dataset at various granularity levels.

References

Shahaf Arica, Or Rubin, Sapir Gershov, and Shlomi Laufer. Cuvler: Enhanced unsupervised object discoveries through exhaustive self-supervised transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23105–23114, 2024.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024.

Shengcao Cao, Jiuxiang Gu, Jason Kuen, Hao Tan, Ruiyi Zhang, Handong Zhao, Ani Nenkova, Liang-Yan Gui, Tong Sun, and Yu-Xiong Wang. Sohes: Self-supervised open-world hierarchical entity segmentation. *arXiv preprint arXiv:2404.12386*, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Danhui Chen, Ziquan Liu, Chuxi Yang, Dan Wang, Yan Yan, Yi Xu, and Xiangyang Ji. Conformalsam: Unlocking the potential of foundational segmentation models in semi-supervised semantic segmentation with conformal prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24045–24055, 2025.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1978, 2014.

Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, Joao Carreira, et al. Bootstap: Bootstrapped training for tracking-any-point. In *Proceedings of the Asian Conference on Computer Vision*, pp. 3257–3274, 2024.

Zhirui Gao, Renjiao Yi, Yuhang Huang, Wei Chen, Chenyang Zhu, and Kai Xu. Partgs: Learning part-aware 3d representations by fusing 2d gaussians and superquadrics. *arXiv preprint arXiv:2408.10789*, 2024.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in neural information processing systems*, 25, 2012.

Oliver Hahn, Christoph Reich, Nikita Araslanov, Daniel Cremers, Christian Rupprecht, and Stefan Roth. Scene-centric unsupervised panoptic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24485–24495, 2025.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pp. 991–998. IEEE, 2011.

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pp. 128–145. Springer, 2022a.

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*, pp. 16000–16009, 2022b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, and Qianqian Wang. Segment any motion in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3406–3416, 2025.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025.

Markus Karmann and Onay Urfalioglu. Repurposing stable diffusion attention for training-free unsupervised interactive segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24518–24528, 2025.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023.

Chanyoung Kim, Woojung Han, Dayun Ju, and Seong Jae Hwang. Eagle: Eigen aggregation learning for object-centric unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3523–3533, 2024a.

Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21530–21539, 2024b.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

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*, pp. 9579–9589, 2024.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, pp. 467–484. Springer, 2024.

Kehan Li, Yian Zhao, Zhennan Wang, Zesen Cheng, Peng Jin, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Multi-granularity interaction simulation for unsupervised interactive segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 666–676, 2023a.

Xiang Li, Chung-Ching Lin, Yinpeng Chen, Zicheng Liu, Jinglu Wang, Rita Singh, and Bhiksha Raj. Paintseg: Painting pixels for training-free segmentation. *Advances in Neural Information Processing Systems*, 36:35–56, 2023b.

Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Zheng Lin, Nan Zhou, Chen-Xi Du, Deng-Ping Fan, and Shi-Min Hu. Refcut: Interactive segmentation with reference guidance. *arXiv preprint arXiv:2503.17820*, 2025.

Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22290–22300, 2023.

Kevin McGuinness and Noel E O'connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010.

Dantong Niu, Xudong Wang, Xinyang Han, Long Lian, Roei Herzig, and Trevor Darrell. Unsupervised universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22744–22754, June 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Zhongwei Ren, Zhicheng 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*, pp. 26374–26383, 2024.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Leon Sick, Dominik Engel, Sebastian Hartwig, Pedro Hermosilla, and Timo Ropinski. Cuts3d: Cutting semantics in 3d for 2d unsupervised instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21265–21275, 2025.

Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Bin Wang, Anwesa Choudhuri, Meng Zheng, Zhongpai Gao, Benjamin Planche, Andong Deng, Qin Liu, Terrence Chen, Ulas Bagci, and Ziyan Wu. Order-aware interactive segmentation. *arXiv preprint arXiv:2410.12214*, 2024a.

Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14176–14186, 2022.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023a.

Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3124–3134, 2023b.

Xudong Wang, Ishan Misra, Ziyun Zeng, Rohit Girdhar, and Trevor Darrell. Videocutler: Surprisingly simple unsupervised video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22755–22764, 2024b.

XuDong Wang, Jingfeng Yang, and Trevor Darrell. Segment anything without supervision. *Advances in Neural Information Processing Systems*, 37:138731–138755, 2024c.

XuDong Wang, Shaolun Zhang, Shufan Li, Konstantinos Kallidromitis, Kehan Li, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Segllm: Multi-round reasoning segmentation. *arXiv preprint arXiv:2410.18923*, 2024d.

Xuehao Wang, Zhan Zhuang, Feiyang Ye, and Yu Zhang. Mtsam: Multi-task fine-tuning for segment anything model. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15790–15801, 2023c.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20193–20203, 2025.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22193–22204, 2025.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3858–3869, 2024.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024b.

Yunhan Yang, Yukun Huang, Yuan-Chen Guo, Liangjun Lu, Xiaoyang Wu, Edmund Y Lam, Yan-Pei Cao, and Xihui Liu. Sampart3d: Segment any part in 3d objects. *arXiv preprint arXiv:2411.07184*, 2024c.

Yunhan Yang, Yufan Zhou, Yuan-Chen Guo, Zi-Xin Zou, Yukun Huang, Ying-Tian Liu, Hao Xu, Ding Liang, Yan-Pei Cao, and Xihui Liu. Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. *arXiv preprint arXiv:2507.06165*, 2025.

Mingqiao Ye, Seoung Wug Oh, Lei Ke, and Joon-Young Lee. Entitysam: Segment everything in video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24234–24243, 2025.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. *arXiv preprint arXiv:2510.23607*, 2025.

Yian Zhao, Kehan Li, Zesen Cheng, Pengchong Qiao, Xiawu Zheng, Rongrong Ji, Chang Liu, Li Yuan, and Jie Chen. Graco: Granularity-controllable interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3501–3510, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE transactions on image processing*, 29:8326–8338, 2020.