Seeing as Humans Do: Learning from Motion to Segment Anything Without Supervision

Weijian Jian[1,*], Xiaoyue Zhang[2,*], Bin Xiao[3], Chunyu Xie[1], Yixiao He[4], Yutao Liu[1], Dawei Leng[1,✉], and Yuhui Yin[1]

[1] 360 AI Research
lengdawei@360.cn
[2] Independent Researcher
[3] University of Ottawa
[4] Beijing University of Posts and Telecommunications

Abstract. The Segment Anything Model (SAM) relies heavily on massive manual annotations, creating a fundamental bottleneck for model scaling. While unsupervised methods attempt to learn object concepts from motion, they typically overfit to moving entities, lacking both multi-granularity understanding and the ability to generalize to static objects. To overcome this, we introduce **Mo**tion-Grounded **S**egment **A**nything (**MoSA**), a highly scalable unsupervised framework that learns a transferable objectness prior from unlabeled videos. MoSA operates in three progressive stages: (1) automatically generating multi-granularity motion pseudo-labels from large-scale video data; (2) training a **P**erceptual **G**rouping **M**odel (**PGM**) via contrastive learning to internalize a generalized, appearance-driven concept of objects; and (3) transferring this learned prior into a prompt-guided architecture for segment-anything-style inference on images. Extensive zero-shot evaluations across seven challenging benchmarks (e.g., COCO and ADE20K) demonstrate that MoSA significantly outperforms existing unsupervised methods. Notably, despite using zero manual annotations, MoSA achieves segmentation performance comparable to the fully supervised SAM. Our findings reveal that harnessing large-scale unlabeled motion is a feasible and highly scalable alternative to annotation-driven segment-anything pipelines.

Keywords: Unsupervised Learning · Segment Anything

1 Introduction

The Segment Anything Model (SAM) [15] has established a new standard for zero-shot generalization in computer vision. However, its success heavily relies on the massive, manually annotated SA-1B dataset. This dependence on costly supervision creates a fundamental bottleneck for further model scaling. Consequently, annotation-free unsupervised learning has emerged as a crucial direction for scalable, general-purpose segmentation.

*Equal contribution, ✉Corresponding author

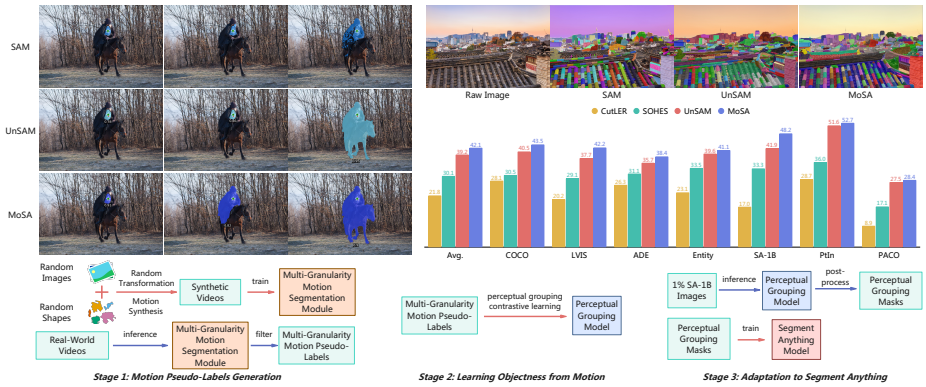

Fig. 1: Overview of our proposed **MoSA**. The **top-left** and **top-right** panels respectively show qualitative comparisons of MoSA against the fully supervised SAM [15] and the unsupervised UnSAM [37] on point-based promptable segmentation and whole-image segmentation. The middle-right panel shows that MoSA achieves SOTA performance compared to other unsupervised segmentation methods. The **bottom** panel illustrates the pipeline of MoSA's three-stage unsupervised framework. Note that UnSAM is not strictly unsupervised, as it utilizes the supervised CascadePSP [7] refiner.

Existing unsupervised segmentation methods generally fall into two categories: image-based and video-based approaches. Image-based methods, such as CutLER [36] and UnSAM [37], leverage self-supervised representations (e.g., DINO [4]) to perform feature clustering or pseudo-label generation, achieving competitive unsupervised segmentation performance. Nevertheless, these approaches infer objectness primarily from static appearance cues, such as texture, color, and shape, overlooking the cognitive priors humans develop through continuous observation of the dynamic world.

Research in developmental psychology, particularly Gestalt theory [16, 17], suggests that the human ability to group objects is largely driven by motion cues. Video-based methods naturally exploit these motion cues. Early approaches typically focus on detecting moving foreground regions, often producing binary masks without explicitly modeling multiple object instances or detailed structures [20, 40, 42]. Recent advances extend unsupervised video object segmentation (U-VOS) toward instance- and scene-level understanding. For example, OCLR [39] proposes an object-centric model trained on synthetic data to discover and track multiple moving objects. Other methods rely on videos during training but can perform inference on static images without requiring motion information. Methods such as RCF [22] and DyStaB [41] exploit motion-induced grouping to generate object masks, often combining motion and appearance refinement within bootstrapping frameworks. Furthermore, approaches like DIOD [14] integrate motion-guided discovery with self-distillation to improve pseudo-label quality. Although these methods prove that motion is a powerful supervisory signal for object concepts, they face two major limitations. First, they easily overfit

to moving objects, struggling to generalize to static or weakly moving entities (e.g., the roof tiles in Fig. 1). Second, they are largely confined to instance-level segmentation, lacking multi-granularity understanding. To truly emulate human perception, a scalable model must generalize from moving entities to *anything* in static scenes, expanding its object vocabulary from growing unlabeled videos.

Motivated by this insight, we introduce **Mo**tion-Grounded **S**egment **A**nything (**MoSA**), a three-stage unsupervised framework designed to learn a transferable, general-purpose segmentation capability from unlabeled videos (see Fig. 1). In the first stage, we develop a **M**ulti-**G**ranularity **M**otion **S**egmentation (**MGMS**) pipeline. Pre-trained on synthetic data and subsequently applied to real-world unlabeled videos, MGMS generates high-quality motion-based pseudo-labels at scale without manual annotation. In the second stage, we train a **P**erceptual **G**rouping **M**odel (**PGM**) using these motion pseudo-labels. This stage encourages the model to internalize a generalized, appearance-driven notion of objectness that no longer depends on motion cues. Finally, an efficient adaptation stage transfers the learned objectness prior to a prompt-guided, high-resolution segmentation architecture, enabling strong zero-shot segment-anything performance. We conduct extensive evaluations on seven challenging segmentation benchmarks, including COCO [23] and ADE20K [45]. Experimental results demonstrate that MoSA consistently outperforms existing unsupervised segmentation methods across all benchmarks. Notably, despite being trained entirely without manual annotations, MoSA achieves zero-shot segmentation performance comparable to SAM, which relies on large-scale supervised data. These findings suggest that learning general object perception from large-scale unlabeled motion is both feasible and scalable, offering a promising path toward foundational general-purpose segmentation. Our main contributions are as follows:

- We propose **MoSA**, a motion-grounded segmentation framework that learns from unlabeled videos to acquire a general-purpose, multi-granularity segmentation capability without manual annotation.
- We introduce a **P**erceptual **G**rouping **M**odel (**PGM**) with a contrastive learning objective that allows the model to generalize beyond motion-specific cues, facilitating the discovery and segmentation of unseen static objects.
- MoSA achieves new state-of-the-art performance in unsupervised segmentation across extensive experiments, proving that learning general objectness from motion is a scalable and practical route to developing foundational vision models.

2 Related Work

Unsupervised Image Segmentation Unsupervised image segmentation aims to automatically identify and segment objects from static images without manual annotations [13, 34]. Methods such as LOST [31] and TokenCut [38] leverage patch-level features from pretrained vision transformers (e.g., DINO family [4, 27, 32]) to localize salient objects via feature affinity and graph-based partitioning. CutLER [36] introduces a cut-and-learn pipeline that discovers pseudo-

masks and trains segmentation models in a self-supervised manner, demonstrating strong multi-instance detection and segmentation capabilities. UnSAM [37] adopts a divide-and-conquer strategy that combines hierarchical grouping and mask refinement to generate multi-granularity pseudo-labels, enabling both automatic and promptable segmentation. A common thread in these image-based approaches is their reliance on DINO features and the subsequent use of hierarchical clustering to generate pseudo-labels. However, they may lack the robust, world-grounded prior knowledge that humans develop through extensive, long-term observation.

Motion-to-Objectness Bootstrapping A growing body of research leverages motion as a natural supervisory signal for self-supervised object discovery. The standard paradigm extracts motion segments from videos to train static image models, allowing the concept of "objects" to generalize beyond initially moving regions. For instance, DyStaB [41] partitions the motion field by minimizing the mutual information between segments, then employs a dynamic-static bootstrapping strategy that iteratively alternates between motion-based segmentation and static model learning. RCF [22] combines relaxed common-fate grouping with appearance refinement to generate object masks. Furthermore, Bao *et al*. [2] utilize a slot-attention framework to discover independently moving objects, while DIOD [14] integrates motion guidance with self-distillation to enhance pseudo-label quality. However, existing methods remain confined to narrow, specialized scenarios. Their training objectives focus heavily on extracting precise instance masks rather than learning a universal, transferable "object prior" necessary for foundational "segment-anything" capabilities. Furthermore, they are strictly restricted to instance-level predictions and cannot comprehend objects across multiple granularities (e.g., distinguishing a part from a whole). Consequently, their learned representations suffer from a severe *motion bias*—they overfit to motion-salient regions and fail on static or rarely moving objects. Ultimately, building a scalable, general-purpose segmentation model solely from motion cues remains an open challenge.

3 Motion Pseudo-Label Generation

To learn object concepts from motion without manual supervision, MoSA begins with a **M**ulti-**G**ranularity **M**otion **S**egmentation (**MGMS**) pipeline that generates high-quality pseudo-labels from unlabeled videos. Within the MGMS pipeline, motion segmentation modules are first trained on synthetic video data. These modules are subsequently applied to large-scale real-world videos to produce a large set of candidate masks (see Fig. 2). Finally, a dedicated quality assessment metric is introduced to filter the generated candidates, resulting in a high-quality, multi-granularity motion pseudo-label dataset for subsequent-stage training.

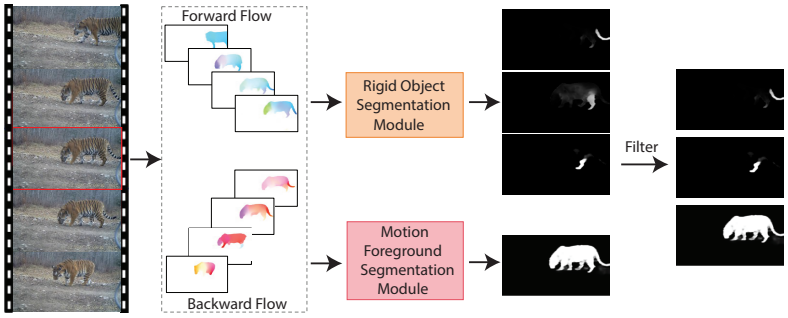

Fig. 2: An overview of our Multi-Granularity Motion Segmentation (MGMS) pipeline for generating high-quality pseudo-labels from unlabeled videos. The pipeline first extracts bidirectional optical flow from videos and processes it through two parallel modules: the Rigid Object Segmentation Module and the Motion Foreground Segmentation Module. These generate masks for rigid parts and the moving foreground of the central frame (highlighted in red). Finally, masks are filtered and processed with NMS to yield the final multi-granularity pseudo-labels.

3.1 Synthetic Video Generation

Segmenting moving objects from real-world videos is inherently challenging due to confounding factors such as complex camera motion, occlusions, and diverse object dynamics (e.g., rigid versus non-rigid motion). To mitigate these challenges and provide clean supervisory signals, we introduce a video synthesis pipeline to train our motion segmentation modules. Inspired by [19], we generate synthetic videos with accurately constructed ground-truth masks. We extend the original framework to support multiple objects and complex occlusions. Specifically, for each synthetic video, a background image is sampled as the background texture, and random affine transformations are applied to simulate camera motion. We then generate multiple foreground objects, represented as randomly shaped and textured patches. Each object moves independently along a smooth B-spline trajectory and undergoes random perspective transformations, with inter-frame smoothness constraints to mimic natural motion and deformation. Although synthetic videos lack semantic realism, they provide unambiguous motion cues and pixel-accurate masks, making them well-suited for training general-purpose motion segmentation models.

3.2 Multi-Granularity Motion Segmentation

To capture motion at varying levels of detail, we design a pipeline with two complementary modules, both trained on our synthetic data. The first module focuses on instance-level segmentation of rigidly moving objects, while the second identifies the entire moving foreground. This multi-granularity strategy ensures a comprehensive capture of diverse motion patterns. Further training details for each module are provided in the **supplementary material**.

Rigid Object Segmentation Module (ROSM) This module is designed to segment individual objects that undergo coherent motion. In our context, "rigid motion" refers to any motion pattern that can be well-approximated by a single perspective transformation. This assumption holds not only for truly rigid objects (e.g., a car), but also for instance parts (e.g., a person's arm) or even static elements exhibiting parallax due to camera motion. Our synthetic data, where each object follows a distinct transformation, provides an ideal training ground for this task. We formulate this as an instance segmentation problem, adapting the SOLOv2 framework [35] to operate on bidirectional optical flow. For a given flow field, the module outputs a set of instance masks, each corresponding to a distinct moving entity, along with a confidence score.

Motion Foreground Segmentation Module (MFSM) While the rigid object module excels at segmenting coherent motion, it struggles with complex, non-rigid deformations, such as a walking person whose limbs move semi-independently. To address this, we introduce a module that performs a simpler, yet highly effective task: separating the entire moving foreground from the background. This yields high-quality masks for complex non-rigid objects, complementing the instance-level output of the first module. To this end, we employ a U-Net architecture [30] that takes bidirectional optical flow as input and outputs a single binary mask delineating all moving regions within the central frame.

3.3 Mask Quality Assessment and Filtering

ROSM and MFSM produce a large volume of candidate masks, many of which are low-quality and can degrade the performance of downstream models. To prune these, we introduce a **Mask Quality Score** (S_{quality}), which combines two components to retain high-fidelity predictions:

Maskness ($S_{maskness}$) This score measures the model's confidence in the pixels it predicts as belonging to the mask. Given a soft mask prediction $p \in [0,1]^{H \times W}$, we define the set of positive pixels as $\mathcal{P}_{\text{pos}} = \{(x,y) \mid p(x,y) > \tau_c\}$, where τ_c is a confidence threshold. The score is then the mean prediction value for this set of pixels:

$$S_{\text{maskness}} = \frac{1}{|\mathcal{P}_{\text{pos}}|} \sum_{(x,y) \in \mathcal{P}_{\text{pos}}} p(x,y). \tag{1}$$

Boundary Sharpness ($S_{sharpness}$) This score rewards masks with well-defined, sharp edges. Let M_b be the mask binarized from p at threshold τ_c. Let B be a d-pixel wide band around the contour of M_b, and ∇p be the gradient magnitude of p. The sharpness is the proportion of boundary pixels with a gradient magnitude above a threshold γ_g:

$$S_{\text{sharpness}} = \frac{1}{|B|} \sum_{(x,y) \in B} \mathbb{I}(\|\nabla p(x,y)\| > \gamma_g), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The final quality score is a weighted sum of these two components: $S_{\text{quality}} = \beta \cdot S_{\text{maskness}} + (1-\beta) \cdot S_{\text{sharpness}}$, where β is a weighting factor. We discard masks with a score below a predefined threshold. Finally, we apply Non-Maximum Suppression (NMS) [26] to the filtered masks to eliminate duplicates, yielding the final set of pseudo-labels.

3.4 Pseudo-Label Dataset Construction

We applied our MGMS pipeline to a diverse collection of large-scale public video datasets[5]. The videos span a wide range of scenes, including human activities, animal behaviors, and driving footage. The complementary nature of our two segmentation modules allows us to generate a rich, multi-granularity dataset. The rigid instance module identifies distinct objects and parts, while the motion foreground module captures complete non-rigid entities. From approximately 10,000 hours of raw video, we generated around **21 million** high-quality motion pseudo-labels. This large-scale, automatically curated dataset forms the foundation for training our Perceptual Grouping Model, as detailed in the next section.

4 Learning from Motion to Segment Anything

While the pseudo-labels from Section 3 are derived from motion, our ultimate goal is to train a universal segmentation model that operates on static images, without any reliance on motion cues at inference time. However, a fundamental challenge arises from the nature of our pseudo-supervision: the masks generated from motion are high-quality but inherently sparse, covering only a fraction of the objects present in an image, e.g., one of the tiger's paws in Fig. 2. To bridge this gap, we introduce the **P**erceptual **G**rouping **M**odel (**PGM**), a framework designed to learn a generalizable concept of "objectness" from this limited supervision. Central to our approach is a novel contrastive training strategy, which enables the model to discover unlabeled objects beyond the provided masks.

4.1 Perceptual Grouping Model Architecture

PGM is built upon a standard Vision Transformer (ViT) [8] backbone. To capture objects at various scales, PGM incorporates K parallel linear projection heads on top of the ViT features. During training, each pseudo-mask is assigned to a specific head based on its bounding box size. The loss for that mask is computed exclusively using the features from its assigned head. This strategy encourages different heads to specialize in segmenting objects of different scales. Furthermore, to enhance robustness against occlusions and complex scenes, we employ the Copy-Paste data augmentation technique [9].

[5] We emphasize that no manual annotations (e.g., labels, boxes, masks) from these datasets were used, and we ensured no overlap with our downstream evaluation benchmarks.

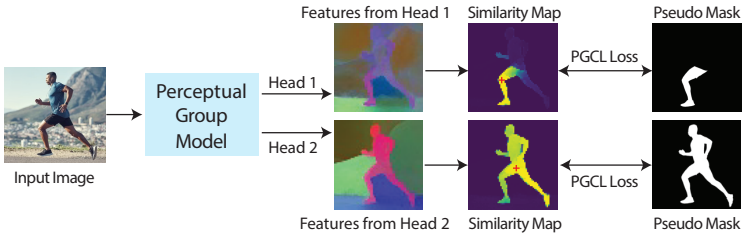

Fig. 3: Illustration of Perceptual Grouping Contrastive Learning (PGCL). Our PGM features multiple heads, each trained on pseudo-masks of a specific scale range. For each assigned mask, we compute a similarity map by comparing an in-mask anchor patch (+) with all patches in the corresponding head's feature space. The PGCL loss supervises this map using the mask as the ground truth, training the model to group coherent patches into objects.

4.2 Perceptual Grouping Contrastive Learning (PGCL)

Standard segmentation objectives, such as Binary Cross-Entropy (BCE) [30] or Dice loss [33], are suboptimal for our training data. Since motion-based pseudo-labels are inherently sparse, these losses erroneously penalize unannotated object instances as background. While Slot Attention [14] addresses sparsity, it struggles in open-world settings: its fixed slot capacity clashes with the varying number of objects in open-world scenes, and its rigid pixel-to-slot assignment hampers multi-granularity segmentation. Crucially, slot-centric representations tend to overfit to the specific object categories seen during training, limiting *zero-shot generalization* to unseen domains.

To overcome these significant limitations and cultivate a more robust and generalizable notion of "objectness" independent of explicit object count or categories, we introduce **Perceptual Grouping Contrastive Learning (PGCL)**, as illustrated in Fig. 3. The central idea is to impose intra-object feature compactness while promoting inter-object feature separability in the patch-level feature space. Specifically, for an input image tokenized into N patches, each of the K heads produces a set of patch embeddings $F^k \in \mathbb{R}^{N\times D}$. Consider a pseudo-mask $\mathcal{M}_i$ assigned to head H_k. We sample an anchor patch embedding $\mathbf{f}_t^k$ from within the mask region. We then compute its cosine similarity $s_{t,j}^k$ with all patch embeddings $\{\mathbf{f}_j^k\}_{j=1}^N$ from the same head. The probability $p_{t,j}^k$ that patch j belongs to the same object as anchor t is modeled via a sigmoid function with a learnable temperature τ_k:

$$p_{t,j}^k = \sigma(\tau_k \cdot s_{t,j}^k), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function. We define a binary label $m_{t,j}$, which is 1 if both the anchor patch t and patch j are within the mask $\mathcal{M}_i$, and 0 otherwise. The PGCL loss is a BCE loss, averaged across all heads and sampled anchors:

$$\mathcal{L}_{\text{PGCL}} = -\frac{1}{KN_aN}\sum_{k=1}^{K}\sum_{t=1}^{N_a}\sum_{j=1}^{N}\Big[m_{t,j}\log p_{t,j}^k + (1-m_{t,j})\log(1-p_{t,j}^k)\Big], \tag{4}$$

where N_a is the number of sampled anchors. By explicitly contrasting positive pairs (patches within the same mask) against negative ones, PGCL enables the model to capture object coherence based on feature similarity. This allows the network to generalize beyond the sparse pseudo-labels, effectively handling the "segment anything" task for unseen objects.

4.3 Adaptation to Segment Anything

While PGM learns a general notion of objectness, it is trained on relatively low-resolution video frames and lacks prompt-based interaction capabilities. To bridge this gap and enable high-performance, high-resolution segmentation with prompt-based interaction, we leverage PGM to generate high-quality masks on high-resolution static images. We then utilize these perceptual grouping masks to train two specialized models for distinct segment anything tasks.

PGM Inference Our perceptual grouping masks generation begins by feeding a static image into the pre-trained PGM. PGM processes the image to generate an initial pool of candidate soft masks, derived from patch embeddings through pairwise cosine similarity. These candidates are then filtered based on a predicted mask quality score (Sec. 3.3) and refined using a dense Conditional Random Field (CRF) [18] to sharpen object boundaries. For high-resolution imagery, we employ a multi-scale, sliding-window approach: PGM inference is applied independently to overlapping tiles at various resolutions. Masks from all tiles are subsequently projected back, merged, and filtered via Non-Maximum Suppression (NMS) to remove duplicates and produce the final comprehensive and fine-grained results. Further implementation details can be found in the **supplementary material**.

Training the Segment Anything Models We apply the aforementioned inference pipeline to generate perceptual grouping masks on a 1% subset of the high-resolution SA-1B dataset [15]. We then train two Segment Anything models, following the approach of UnSAM [37]:

- **Whole-Image Segmentation:** For universal object segmentation, we train a Mask2Former model [6]. This model accepts a full high-resolution image as input and automatically outputs a set of masks, where each mask corresponds to a distinct instance discovered in the scene.
- **Promptable Segmentation:** To build a prompt-driven model, we adopt an architecture based on Semantic-SAM [21], which excels at predicting masks at multiple granularity levels from a single click. To enable interactivity, we simulate user clicks during training by randomly sampling a point within the foreground of each mask to serve as a positive point prompt.

Table 1: Evaluation results of PGM on seven datasets. †: The pseudo-labels used by UnSAM [37] are not strictly unsupervised as they are refined by CascadePSP [7].

Methods	PtIn	LVIS	Entity	PACO	ADE	COCO	SA-1B	**Average**
UnSAM [37] (CRF [18])	23.2	16.5	22.1	10.8	15.1	21.9	15.3	17.8
UnSAM [37] (CascadePSP† [7])	27.3	19.9	24.3	12.8	17.8	24.8	24.7	21.7
PGM (**ours**) (w/o refinement)	34.7	27.9	28.2	16.1	25.5	32.6	36.6	28.8
PGM (**ours**) (CRF [18])	**36.1**	**29.3**	**30.1**	**16.9**	**26.7**	**33.8**	**37.8**	**30.1**

5 Experiments

5.1 Training Setups

MGMS Pipeline Training Settings Our MGMS Pipeline comprises two key modules: ROSM employs a SOLO-v2 architecture [35] with a ResNet-18 backbone [12], while MFSM utilizes a Res-UNet architecture [44], also with a ResNet-18 backbone. Both modules take a sequence of 7 frames as input and were trained on a synthetic dataset for 50,000 iterations with an input resolution of 512×512 and a batch size of 32.

PGM Training Settings The training data for PGM comprises 21 million pseudo-labels distributed across 10 million static frames. This dataset was generated from a 10,000-hour video corpus constructed from three major sources: Kinetics-700 [5], which contains approximately 650,000 10-second clips covering a wide range of human actions; BDD100K [43], a large-scale driving video dataset designed for diverse road scenarios; and a 5% random sample from YouTube-8M [1], a broad-coverage video collection sourced from YouTube. PGM utilizes a ViT-base architecture [8] with an input size of 512×512 pixels. The architecture is topped by four parallel heads, each implemented as a linear layer that produces 128-dimensional feature vectors. We employ the AdamW [25] optimizer with a learning rate of 1×10^{-4} and a batch size of 32.

Segment Anything Model Training Settings For whole-image segmentation, we construct a model by integrating a Mask2former [6] decoder with a ResNet-50 [12] backbone. Training is conducted for 8 epochs, configured with a 5×10^{-5} learning rate, a batch size of 16, and 0.05 weight decay. For promptable segmentation, we leverage the Semantic-SAM [21] architecture with a Swin-Transformer [24] Tiny model as the backbone to improve performance. Both models are trained with only a 1% subset of unlabeled images from SA-1B [15].

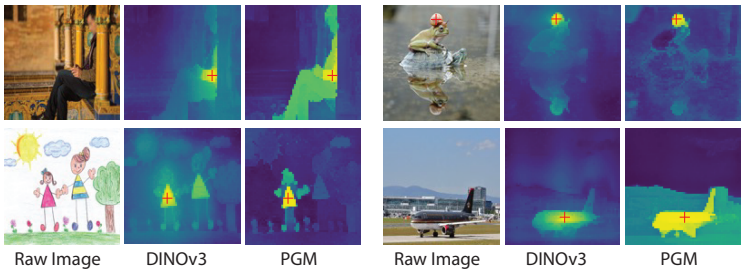

Fig. 4: Qualitative comparison of PGM (ours) with DINOv3 family [32], where the red + denotes anchor patches.

5.2 Evaluation Datasets and Metrics

We evaluate MoSA under two segmentation settings: *Whole-Image Segmentation* and *Point-Based Promptable Segmentation*. For Whole-Image Segmentation, we evaluate MoSA on seven benchmark datasets, including COCO [23], LVIS [10], ADE20K [45], EntitySeg [28], SA-1B [15], PartImageNet [11], and PACO [29]. Importantly, each dataset only labels a subset of objects, whereas our model generates masks across heterogeneous levels and open-world categories. As a result, the standard COCO-style Average Precision (AP) metric may not fully capture the model's ability to segment diverse, open-world entities. Therefore, following prior unsupervised segmentation works [3, 36, 37], we adopt Average Recall (AR) as the primary metric for whole-image segmentation comparisons. For Point-Based Promptable Segmentation, we evaluate on COCO Val2017 [23]. Following prior promptable segmentation methods [15,21], we report two metrics: MaxIoU and OracleIoU. MaxIoU measures the Intersection over Union (IoU) between the ground truth mask and the highest-confidence predicted mask, while OracleIoU reports the maximum IoU achieved among all predicted masks.

5.3 Evaluation Results

Perceptual Grouping Masks We quantitatively evaluate PGM on seven benchmarks. As shown in Table 1, PGM generates high-quality masks. Notably, even without refinement, they significantly outperform the pseudo-labels used by UnSAM [37], which are post-processed with CascadePSP [7] (a method relying on manual annotations). Qualitatively, Fig. 4 compares feature similarity maps of PGM and the strong self-supervised vision model DINOv3 [32]. The results highlight PGM's superior ability to learn a holistic concept of "objectness". For instance, in the top-right image, DINOv3 only activates on part of the snail, whereas PGM correctly segments the entire snail. Moreover, PGM shows remarkable zero-shot generalization. When applied to an out-of-distribution hand-drawn image (bottom-left), it successfully identifies the child figure, whereas the baseline fails to grasp this abstract concept. This suggests that our PGCL effectively teaches the model a generalizable notion of object coherence from sparse motion-based supervision.

Table 2: Results on the whole-image segmentation task. †: UnSAM [37] utilizes CascadePSP [7], a method that relies on manual annotations, and is therefore not strictly unsupervised. MoSA achieves SOTA performance among unsupervised methods only using a ResNet-50 backbone and delivers results comparable to SAM [15].

Methods	Backbone #params	# images	**Avg.**	with Whole Entities					with Parts	
				COCO	LVIS	ADE	Entity	SA-1B	PtIn	PACO
SAM [15]	VIT-B(85M)	11M	42.1	49.6	46.1	45.8	45.9	60.8	28.3	18.1
FreeSOLO [34]	RN-101 (45M)	1.3M	7.3	11.6	5.9	7.3	8.0	2.2	13.8	2.4
CutLER [36]	RN-50 (23M)	1.3M	21.8	28.1	20.2	26.3	23.1	17.0	28.7	8.9
SOHES [3]	VIT-B(85M)	0.2M	30.1	30.5	29.1	31.1	33.5	33.3	36.0	17.1
UnSAM† [37]	RN-50(23M)	0.1M	39.2	40.5	37.7	35.7	39.6	41.9	51.6	27.5
MoSA (**ours**)	RN-50(23M)	0.1M	**42.1**	**43.5**	**42.2**	**38.4**	**41.1**	**48.2**	**52.7**	**28.4**

Whole-Image Segmentation Table 2 presents the whole-image segmentation performance, measured by AR, of our unsupervised MoSA against baselines. Notably, MoSA, trained with only 1% of the SA-1B training data and utilizing a ResNet-50 backbone (only 23M parameters), achieves an average AR of 42.1%. This performance significantly outperforms previous unsupervised methods and the improvement is consistent across all seven datasets. Furthermore, on PartImageNet [11] and PACO [29], MoSA surpasses the supervised SAM (ViT-B, 11M images) by 24.4% and 10.3% in AR, respectively. This suggests that pre-training on video motion cues enables the discovery of fine-grained objects and details often overlooked by human annotators. Qualitative results in Fig. 1 further illustrate this; notably, MoSA effectively segments even completely static objects, such as tiles on a roof. This capability is a direct consequence of PGCL, which extrapolates from sparse, motion-derived pseudo-labels to learn a universal concept of "objectness". MoSA's results align more closely with human perception of object instances and exhibit less segmentation noise.

Point-Based Promptable Segmentation As shown in Table 3, MoSA achieves 41.6% MaxIoU and 63.4% OracleIoU on COCO [23]. This notably outperforms the previous SOTA, UnSAM [37], by +1.3% and +3.9% respectively. Crucially, UnSAM is not strictly unsupervised, as it relies on CascadePSP [7], a refiner trained with manual annotations. While fully supervised SAM [15] sets a higher benchmark (52.1% MaxIoU, 68.2% OracleIoU), MoSA's performance is compelling given its unsupervised nature and high efficiency, requiring 3.4× fewer parameters and 100× less SA-1B data. Qualitatively (Fig. 1), MoSA provides stable, perceptually-aligned, multi-granularity labels. For instance, even in challenging scenarios such as segmenting a person riding a horse (where their motion is often congruent), MoSA leverages video-learned semantics to accurately distinguish them. Additional qualitative results are in the **supplementary materials**.

Table 3: Quantitative comparison on the point-based promptable image segmentation task. MoSA achieves superior performance over the previous SOTA, UnSAM [37], which is not strictly unsupervised as it employs CascadePSP [7].

Methods	Backbone (# params)	% of SA-1B	Point (Max)	Point (Oracle)
SAM [15] (supervised)	ViT-B/8 (85M)	100%	52.1	68.2
UnSAM [37]	Swin-Tiny (25M)	1%	40.3	59.5
MoSA (**ours**)	Swin-Tiny (25M)	1%	**41.6**	**63.4**

Table 4: Ablation study of PGM training.

Method	AR_{1000}	AR_s	AR_m	AR_l
BCE+Dice	3.8	1.2	2.8	6.6
Slot Attention [14]	8.2	5.2	8.6	8.7
PGCL (ours)	**28.3**	**14.7**	**26.6**	**36.1**

5.4 Ablation Study

PGM Training To verify the effectiveness of our proposed PGCL, we conduct a comparative analysis against strong baselines using identical backbone architectures and training data (1,000 hours of video). Specifically, we compare PGCL with: (1) standard **BCE+Dice** loss, and (2) **Slot Attention** [14] with 256 slots. We evaluate the zero-shot transfer performance on 1,000 images from the SA-1B dataset, reporting Average Recall (AR) overall (AR_{1000}) and across varying scales (AR_s, AR_m, AR_l). As presented in Table 4, the conventional BCE+Dice baseline fails to segment most objects because sparse motion pseudo-labels wrongly suppress unlabeled static regions as background. Similarly, Slot Attention exhibits suboptimal performance (8.2 AR). We attribute this to the limitation of fixed slot capacity, which struggles to accommodate the uncertain number of objects and the multi-granular hierarchies inherent in open-world scenes. In contrast, our method achieves a substantial improvement, reaching 28.3 AR. This result confirms the superiority of PGCL in handling the complexity of the "segment anything" task.

MGMS Pipeline We ablate the MGMS Pipeline's key components: the Rigid Object (ROSM) and Motion Foreground (MFSM) Segmentation Modules. Their efficacy is evaluated based on PGM's Average Recall (AR) on 1,000 SA-1B images when trained on their generated pseudo-labels. As shown in Table 5, ROSM alone excels on small objects (20.6% AR_s) but struggles with larger ones. Conversely, MFSM is effective for large objects (43.1% AR_l) but performs poorly on small ones (13.7% AR_s). The full pipeline integrates both modules to achieve the best overall score (36.6% AR_{1000}). This confirms their complementary nature:

Table 5: Ablation study of modules in the MGMS pipeline.

ROSM	MFSM	AR_{1000}	AR_s	AR_m	AR_l
✓		28.8	20.6	31.1	28.7
	✓	34.7	13.7	34.9	43.1
✓	✓	**36.6**	**20.6**	**35.4**	**45.1**

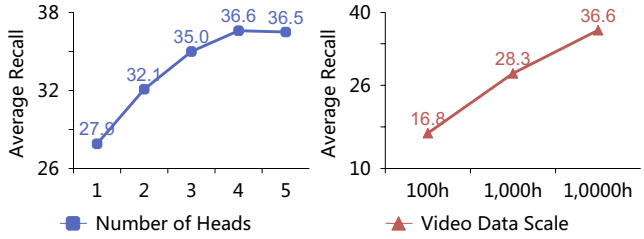

Fig. 5: Impact of the number of prediction heads (left) and pre-training video data scale (right) in PGM.

combining ROSM's precision for small objects with MFSM's broader motion capture is essential for generating robust, multi-scale motion cues.

Number of Prediction Heads We ablate the number of prediction heads in PGM for multi-granularity segmentation. As shown in Fig. 5 (left), performance improves substantially from 27.9% AR with a single head to 36.6% AR with four heads. The most significant gain occurs when adding the second head (+4.2% AR), confirming the fundamental benefit of specialized heads. However, performance saturates beyond four heads, dropping marginally to 36.5% AR with five. This indicates that four heads provide the optimal trade-off, effectively capturing the full range of object scales from motion cues without introducing redundancy.

Impact of Pre-training Video Data Scale We investigate how the amount of unlabeled video data impacts PGM's ability to learn motion priors. As illustrated on the right side of Fig. 5, we observe a strong positive correlation between data scale and performance. Specifically, increasing the training data from 100h to 1,000h, and further to 10,000h, substantially boosts AR from 16.8% to 28.3% and finally to 36.6%. This clear, non-saturating trend underscores that our PGM effectively leverages large-scale video corpora and strongly suggests that its performance can be further enhanced by scaling to even larger datasets.

6 Conclusion

Perceptual grouping is fundamental to human visual intelligence, enabling us to transform raw visual data into semantic understanding through simple, yet

powerful rules. Inspired by this biological process, we introduce MoSA, a novel three-stage unsupervised framework for learning generalized segmentation from motion. Extensive experiments demonstrate that by observing visual motion alone, MoSA learns robust segmentation abilities applicable to diverse objects and scenes. This approach not only outperforms existing unsupervised methods, but also offers a more interpretable and scalable path toward realizing "segment-anything." Our work highlights the immense potential of learning object perception from motion as a fundamental building block for artificial visual intelligence, bridging the gap towards human-like perceptual understanding.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. CoRR **abs/1609.08675** (2016)
2. Bao, Z., Tokmakov, P., Jabri, A., Wang, Y.X., Gaidon, A., Hebert, M.: Discovering objects that can move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11789–11798 (2022)
3. Cao, S., Gu, J., Kuen, J., Tan, H., Zhang, R., Zhao, H., Nenkova, A., Gui, L., Sun, T., Wang, Y.: SOHES: self-supervised open-world hierarchical entity segmentation. In: ICLR. OpenReview.net (2024)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9630–9640. IEEE (2021)
5. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. CoRR **abs/1907.06987** (2019)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1280–1289. IEEE (2022)
7. Cheng, H.K., Chung, J., Tai, Y., Tang, C.: Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In: CVPR. pp. 8887–8896. Computer Vision Foundation / IEEE (2020)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. OpenReview.net (2021)
9. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR. pp. 2918–2928. Computer Vision Foundation / IEEE (2021)
10. Gupta, A., Dollár, P., Girshick, R.B.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR. pp. 5356–5364. Computer Vision Foundation / IEEE (2019)
11. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J., Liu, S., Yang, C., Yu, Q., Yuille, A.L.: Partimagenet: A large, high-quality dataset of parts. In: ECCV (8). Lecture Notes in Computer Science, vol. 13668, pp. 128–145. Springer (2022)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778. IEEE Computer Society (2016)
13. Hu, Y., Huang, J., Schwing, A.G.: Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In: ECCV (1). Lecture Notes in Computer Science, vol. 11205, pp. 813–830. Springer (2018)
14. Kara, S., Ammar, H., Denize, J., Chabot, F., Pham, Q.C.: Diod: Self-distillation meets object discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3975–3985 (2024)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: ICCV. pp. 3992–4003. IEEE (2023)
16. Koffka, K.: Principles of Gestalt psychology. routledge (2013)
17. Köhler, W.: Gestalt psychology. Psychologische forschung **31**(1), XVIII–XXX (1967)

18. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. pp. 282–289. Morgan Kaufmann (2001)
19. Lamdouar, H., Xie, W., Zisserman, A.: Segmenting invisible moving objects. In: BMVC. p. 231. BMVA Press (2021)
20. Lee, M., Cho, S., Lee, S., Park, C., Lee, S.: Unsupervised video object segmentation via prototype memory network. In: WACV. pp. 5913–5923. IEEE (2023)
21. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L., Gao, J.: Segment and recognize anything at any granularity. In: ECCV (48). Lecture Notes in Computer Science, vol. 15106, pp. 467–484. Springer (2024)
22. Lian, L., Wu, Z., Yu, S.X.: Bootstrapping objectness from videos by relaxed common fate and visual grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14582–14591 (2023)
23. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV (5). Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014)
24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 9992–10002. IEEE (2021)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (Poster). OpenReview.net (2019)
26. Neubeck, A., Gool, L.V.: Efficient non-maximum suppression. In: ICPR (3). pp. 850–855. IEEE Computer Society (2006)
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
28. Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Torr, P.H.S., Lin, Z., Jia, J.: Open world entity segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(7), 8743–8756 (2023)
29. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., Mousavi, A., Song, Y., Dubey, A., Mahajan, D.: PACO: parts and attributes of common objects. In: CVPR. pp. 7141–7151. IEEE (2023)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (3). Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015)
31. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing objects with self-supervised transformers and no labels. In: BMVC. p. 310. BMVA Press (2021)
32. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S.E., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3. CoRR **abs/2508.10104** (2025)
33. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: DLMIA/ML-CDS@MICCAI. Lecture Notes in Computer Science, vol. 10553, pp. 240–248. Springer (2017)
34. Wang, X., Yu, Z., Mello, S.D., Kautz, J., Anandkumar, A., Shen, C., Álvarez, J.M.: Freesolo: Learning to segment objects without annotations. In: CVPR. pp. 14156–14166. IEEE (2022)

35. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. In: NeurIPS (2020)
36. Wang, X., Girdhar, R., Yu, S.X., Misra, I.: Cut and learn for unsupervised object detection and instance segmentation. In: CVPR. pp. 3124–3134. IEEE (2023)
37. Wang, X., Yang, J., Darrell, T.: Segment anything without supervision. In: NeurIPS (2024)
38. Wang, Y., Shen, X., Yuan, Y., Du, Y., Li, M., Hu, S.X., Crowley, J.L., Vaufreydaz, D.: Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. IEEE Trans. Pattern Anal. Mach. Intell. **45**(12), 15790–15801 (2023)
39. Xie, J., Xie, W., Zisserman, A.: Segmenting moving objects via an object-centric layered representation. In: NeurIPS (2022)
40. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: ICCV. pp. 7157–7168. IEEE (2021)
41. Yang, Y., Lai, B., Soatto, S.: Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2826–2836 (2021)
42. Yang, Y., Loquercio, A., Scaramuzza, D., Soatto, S.: Unsupervised moving object detection via contextual information separation. In: CVPR. pp. 879–888. Computer Vision Foundation / IEEE (2019)
43. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: CVPR. pp. 2633–2642. Computer Vision Foundation / IEEE (2020)
44. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geosci. Remote. Sens. Lett. **15**(5), 749–753 (2018)
45. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. Int. J. Comput. Vis. **127**(3), 302–321 (2019)