Can Segmentation Models Understand the World? Proactive Affordance Reasoning via Vision-Grounded Chain-of-Thought

Yuchen Guo[1], Junli Gong[2], Hongmin Cai[3], Yiu-ming Cheung[4], Weifeng Su[5]
[1]Northwestern University [2]Northeastern University [3]South China University of Technology
[4]Hong Kong Baptist University [5]IRADS, Beijing Normal-Hong Kong Baptist University
yuchenguo2027@u.northwestern.edu, wfsu@bnbu.edu.cn

Abstract

Recent segmentation models couple large language models (LLMs) with mask decoders to ground complex language expressions into masks, yet their instructions remain target-referential: they describe, constrain, or imply the region to be segmented. However, in real-world embodied interaction, human instructions are often at the intent-level, which includes the desired outcome without naming the region that enables it. To bridge this gap, we introduce ***SegWorld***, where the model reasons about the scene through a multi-level vision-grounded chain-of-thought (CoT) before committing to a mask. Before receiving any instructions, it proactively observes the scene, describing visible objects and inferring plausible events they may support. Given an instruction, it continues the chain: from the object relevant to the intent, through the action that satisfies it, to the physical interaction site, the object part that affords the action. We formalize SegWorld as probabilistic inference, in which proactive observation supplies a linguistic scene context that improves mask prediction when instructions are given at the level of intent. We construct an ***Intent2Part*** benchmark for evaluating affordance-bearing part segmentation from high-level goals. Experiments show SegWorld matches instruction-driven baselines on target-referential instructions and improves substantially on intent-level ones. Code and dataset are available at: https://github.com/Eurekaarrow/SegWorld.

1 Introduction

Language-driven segmentation has expanded the role of segmentation from category recognition to instruction following. Recent models integrate LLMs or multimodal LLMs with mask decoders, enabling them to ground complex natural-language expressions, including spatial relations, object attributes, and implicit reasoning cues, into masks (Lai et al., 2024; Lu et al., 2025; Liu et al., 2026; Li et al., 2025; Gong et al., 2025). These instructions are commonly studied as explicit referring expressions (Kazemzadeh et al., 2014; Yu et al., 2016; Liu et al., 2023) or implicit reasoning instructions (Lai et al., 2024; Jang et al., 2025; Wan et al., 2025). Recent works further introduce vision-grounded chain-of-thought reasoning into segmentation, using textual rationales, point or region cues, or positional priors to connect reasoning with grounding (Kao et al., 2026; Yuan et al., 2025). Despite this progress, the dominant paradigm remains query-conditioned: the model is given an instruction and is expected to return the corresponding region. The visual scene becomes relevant only after a question is asked, and the instruction is treated as a specification of the target.

But this is not how human instructions work in the real world (Sperber and Wilson, 1986). Human-centric instructions are pragmatic and underspecified, requiring the listener or agent to infer the intended goal from context (Shridhar et al., 2020; Tellex et al., 2011; Misra et al., 2016). Models shouldn't view scenes through the narrow aperture of current instructions. Before being asked anything, it can already take in what is present, what could happen, and how things can be used. The instruction then serves not as a full description of the target, but as a way of selecting one actionable possibility from this prior understanding.

This distinction becomes critical when instructions are expressed at the level of intent. Consider a tabletop scene with a mug, a kettle, and a plate, and the instruction "I want to drink water." The desired mask is not named by the instruction at all. A query-conditioned model may ground it to the mug or the kettle, both semantically related to drinking, yet neither specifies the physical site through which the intended action should be executed. Reaching the right region requires several steps of everyday reasoning, and the region it ar-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 38650–38662
October 24-29, 2026 ©2026 Association for Computational Linguistics

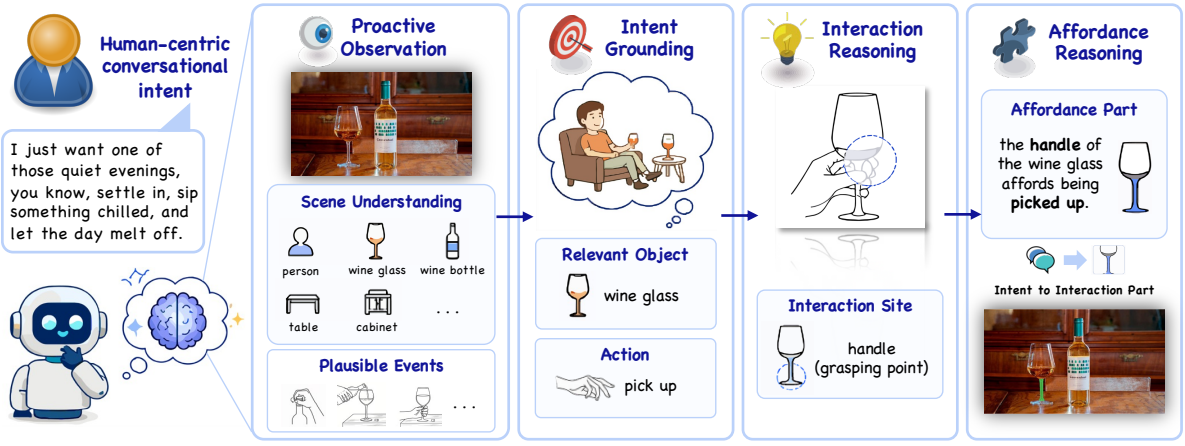

Figure 1: Given a human-centric conversational intent that does not name the target region, SegWorld first proactively observes the scene, then grounds the intent to a relevant object and action, reasons about the physical interaction site, and finally segments the affordance-bearing part that enables the intended action.

rives at is often not a whole object but a part. This is consistent with affordance and part-level grounding studies, where interaction is localized to functional regions rather than entire objects (Do et al., 2018; Nagarajan et al., 2019; Deng et al., 2021; Mo et al., 2021; Wan et al., 2025; Tang et al., 2025). We call this setting *intent-level segmentation*: recovering an affordance-bearing part that lies several reasoning steps below a high-level goal.

We introduce ***SegWorld***, which replaces query-conditioned mapping with a proactive, multi-level reasoning process. Rather than waiting for an instruction, SegWorld first reads the scene on its own: it describes the visible objects and the events they could plausibly support, producing a linguistic scene context before any instruction arrives. When the instruction is given, SegWorld does not search the image from scratch; it traverses this context through an eight-level vision-grounded chain-of-thought organized into three stages. The first stage performs proactive observation. The second identifies the object implicated by the intent and the action that would fulfill it. The third resolves the object part that affords this action and the affordance it carries. Each level is expressed in language and remains explicit and inspectable, turning a single opaque mapping from instruction to pixels into a traceable reasoning process.

This design is principled, not merely procedural. We formalize SegWorld as probabilistic inference: the proactive stage draws a scene context, and the instruction-conditioned stages estimate the mask conditioned on both the instruction and that context. The scene context acts as an informative prior. When an instruction is target-referential, this prior matters little, since the instruction already localizes the answer. But when the instruction is intent-level, the answer is underdetermined by the instruction alone, and the scene context is what makes the posterior tractable. Proactive observation, in other words, is most valuable exactly where query-conditioned models are weakest.

To study this setting, we construct ***Intent2Part***, a benchmark for intent-level segmentation in which instructions name a high-level goal and the target is an affordance-bearing object part. On *Intent2Part*, *SegWorld* matches strong query-conditioned baselines when instructions directly specify their target, and improves over them substantially when the target must be recovered from intent. Beyond accuracy, its vision-grounded chain-of-thought exposes where a prediction comes from, making both correct and incorrect masks interpretable.

Our contributions are summarized as follows:

- We identify ***intent segmentation***, a setting where instructions name a goal rather than a target, and where the answer is an affordance-bearing object part lying several reasoning steps below the instruction.
- We propose ***SegWorld***, a proactive segmentation framework that performs an eight-level vision-grounded chain-of-thought, and formalize it as probabilistic inference in which a proactive scene context serves as a prior for mask prediction.
- We construct ***Intent2Part***, a benchmark for evaluating intent-level segmentation, and show that *SegWorld* matches query-conditioned baselines on direct instructions

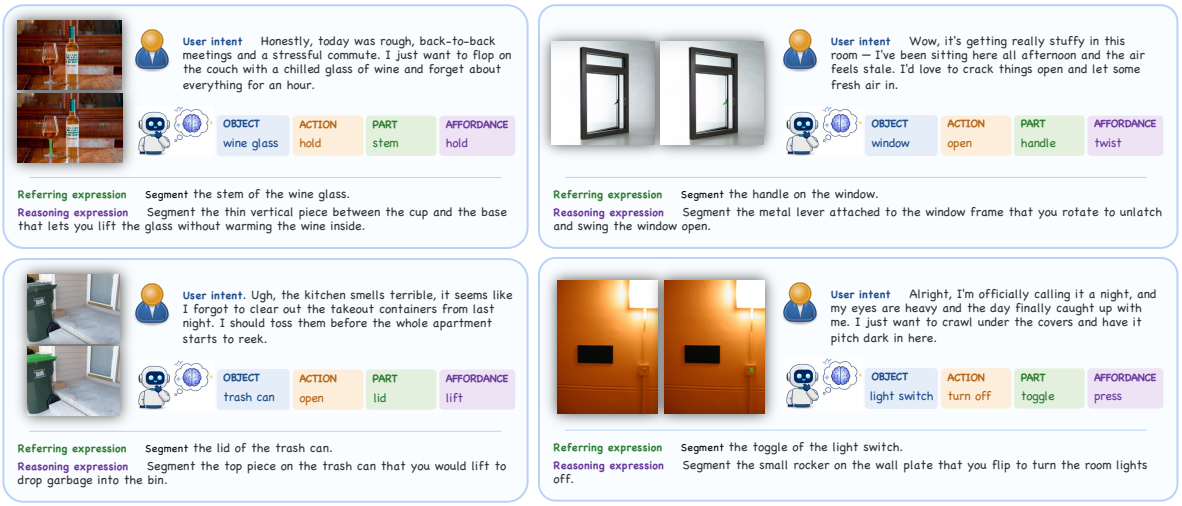

Figure 2: Examples from ***Intent2Part***. Each sample pairs a human-centric intent-level instruction with the latent reasoning chain needed to recover the target: object, action, part, and affordance. The benchmark also retains target-referential referring and reasoning expressions, allowing models to be evaluated under both direct target descriptions and high-level intents.

while improving substantially on intent-level ones.

2 Related Work

Language-driven segmentation. With the rapid development of LLMs, recent studies integrate LLMs into segmentation models (Lu et al., 2025; Liu et al., 2026; Lai et al., 2024), enabling them to segment objects from natural-language instructions rather than single-word category labels (Li et al., 2025; Gong et al., 2025). Such instructions are broadly categorized into explicit referring expressions (Kazemzadeh et al., 2014; Yu et al., 2016; Liu et al., 2023) and implicit reasoning instructions (Lai et al., 2024; Jang et al., 2025; Wan et al., 2025). To strengthen reasoning, some works introduce a vision-grounded chain-of-thought into the segmentation pipeline, using intermediate textual rationales, region or point annotations, or positional priors that couple reasoning with grounding (Kao et al., 2026; Yuan et al., 2025). However, both referring segmentation and reasoning segmentation rely on language expressions that explicitly or implicitly describe the target region to be segmented. In contrast, our intent-level segmentation is human-centric, aiming to better serve real-world embodied agents.

Affordance and part-level grounding. In real-world embodied interaction (Intelligence et al., 2025) and perception (Guo et al., 2026), the target of an action is often not a whole object but the part through which the action is executed (Wan et al., 2025). Prior work grounds affordances in images, videos, and 3D scenes (Tang et al., 2025), from dense affordance segmentation and graspable-region prediction to interaction hotspots and part-level affordance (Do et al., 2018; Nagarajan et al., 2019; Deng et al., 2021; Mo et al., 2021), with recent methods using vision-language models for open-vocabulary affordances (Qian et al., 2024; Chu et al., 2025). These methods take an action or interaction query as input and localize where it applies. Our setting asks the reverse: given only a high-level intent, the model infers the object, the action, and the affordance-bearing part in turn. Affordance is thus not the input query but the terminal semantics of intent-level segmentation.

3 Intent Segmentation

We address a segmentation task defined over an image I and a natural-language instruction q, where the goal is to predict a binary mask M. This setup is shared with referring and reasoning segmentation; the distinction lies in the nature of q and the semantics of M.

In *query-conditioned* segmentation, q is *target-referential*: it identifies the target region by naming it, describing its attributes, or specifying a relation that singles it out. The target is therefore determined by I and q together. In *intent-level segmentation*, q instead specifies a desired outcome, or *intent*, without referring to any image region. Because fulfilling an intent requires acting on the scene, and physical action is localized to a specific

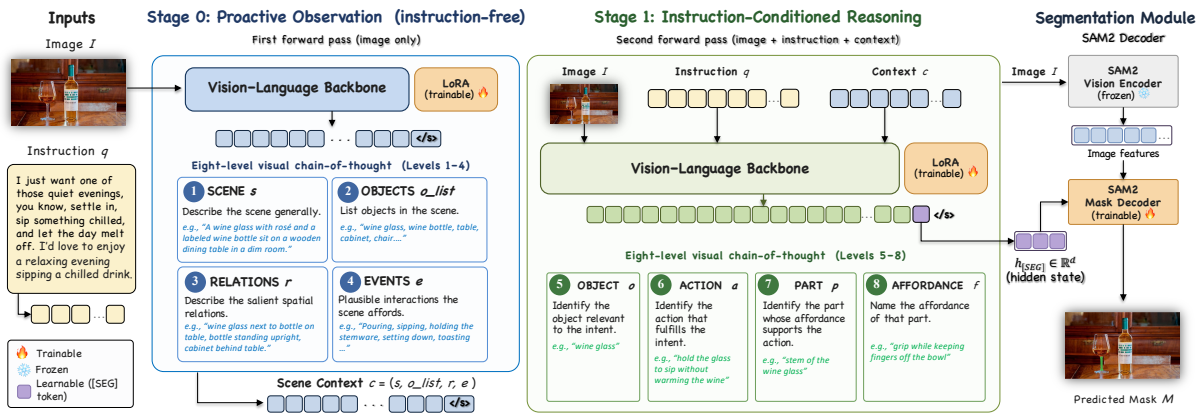

Figure 3: Overview of ***SegWorld***. Stage 0 performs instruction-free proactive observation, producing a scene context from the image alone. Stage 1 conditions on this context and the instruction to reason through object, action, part, and affordance. Shared VLM weights; two forward passes.

object part, the target M is an *affordance-bearing part*: the part whose affordance enables the action that satisfies the intent.

Satisfying q in scene I implicitly involves four latent quantities: an object o relevant to the intent, an action a on o that fulfills it, a part p of o that supports the action, and the corresponding affordance f. The ground-truth mask $M^\star$ is the segmentation of this part. Intent-level segmentation therefore requires recovering the latent chain $q \to o \to a \to (p, f) \to M^\star$, whose intermediate terms are not given at inference time.

This structure makes intent-level segmentation harder than the query-conditioned case. When q is target-referential, the target can often be identified by matching the instruction against image regions. When q is an intent, the model must infer which object can realize the goal, which action should be taken, and which object part affords that action before producing a mask. This motivates the proactive reasoning approach introduced in Section 4.

4 The SegWorld Method

4.1 Overview

SegWorld replaces the direct mapping $q \to M$ with a two-pass, three-stage reasoning process over a shared vision-language backbone. In the first pass, the model observes the image I without access to the instruction and emits a linguistic scene context c. In the second pass, the model takes the instruction q together with this context, resolves the intent into an object and action, grounds the action to an affordance-bearing part, and predicts the target mask M. Together, the stages form an eight-level visual chain-of-thought spanning proactive observation, intent resolution, and part-level grounding. The chain terminates in a special `[SEG]` token whose contextual hidden state drives a SAM2 mask decoder (Ravi et al., 2025).

Probabilistically, SegWorld defines a latent-context model in which the instruction-free scene context c is conditionally independent of the instruction q given the image I. Because Stage 0 is architecturally denied access to q, the model factorizes as

$$P_\theta(M \mid I, q) = \sum_{c \in \mathcal{C}} P_\theta(c \mid I)\, P_\theta(M \mid I, q, c), \tag{1}$$

where $\mathcal{C}$ denotes the space of linguistic scene contexts. This is an identity under SegWorld's model factorization, rather than a claim that the instruction-independent factorization exactly matches the true data-generating distribution.

Exact marginalization over all possible contexts is intractable. At inference, we therefore decode one instruction-free context,

$$\begin{aligned} \hat{c} &= \mathrm{Decode}(P_\theta(c \mid I)), \\ \widehat{P}_\theta(M \mid I, q) &= P_\theta(M \mid I, q, \hat{c}), \end{aligned} \tag{2}$$

and predict the mask conditioned on $\hat{c}$. Thus, the approximation in practice arises from single-context decoding, not from dropping q from $P_\theta(c \mid I)$: the latter is an architectural constraint that makes c an instruction-independent and cacheable scene prior.

4.2 Proactive Observation

The first stage models the instruction-free distribution $P_\theta(c \mid I)$. The instruction q is withheld entirely: the observation pass receives only the image and an observation prompt, ensuring that the

resulting context is a property of the scene rather than of the query.

We decompose the context into four levels, $c = (s, o_{\text{list}}, r, e)$. The SCENE level gives a one-sentence description s of the overall setting. The OBJECTS level enumerates salient visible objects o_{list}. The RELATIONS level captures spatial and functional relations r among them. The EVENTS level predicts a small set of plausible interactions e supported by the scene, independent of any particular instruction. These levels summarize what is present, how it is arranged, and what actions the scene could plausibly afford before any goal is specified.

4.3 Instruction-Conditioned Reasoning

Given the image I, instruction q, and scene context c, the second pass models $P_\theta(M \mid I, q, c)$ through two further stages. The intent-resolution stage identifies the object o relevant to the intent and the action a that would satisfy it. The affordance-grounding stage identifies the part p of o that supports the action and names its affordance f. These four levels recover the latent chain introduced in Section 3, $q \to o \to a \to (p, f) \to M^\star$, making the intermediate reasoning explicit rather than collapsing the instruction directly into a mask.

The final token of the chain is `[SEG]`, added to the backbone vocabulary. Its contextual hidden state $h_{\text{[SEG]}} \in \mathbb{R}^d$ is linearly projected into a prompt embedding for the SAM2 mask decoder. The decoder fuses this prompt with image features from a frozen SAM2 vision encoder to produce the binary mask M. Thus, the mask decoder receives a learned language prompt that encodes the full reasoning chain, rather than the raw instruction alone.

Why the scene context helps. Equation 1 clarifies when proactive observation is useful. For target-referential instructions, q already identifies the target, so additional scene context has limited effect. For intent-level instructions, however, $P_\theta(M \mid I, q)$ is diffuse: multiple objects and parts may be semantically related to the intent, but only some are actionable in the observed scene. Conditioning on c restricts prediction to candidates consistent with the scene structure and plausible events, sharpening the posterior. Whether this restricted, instruction-independent context family is useful is an empirical question. Section 6 evaluates this design through the proactive-context ablation and a one-pass joint alternative.

4.4 Training

Objective. SegWorld is jointly trained with mask and language-modeling objectives:

$$\begin{aligned}\mathcal{L} = \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}(M, M^\star) + \lambda_0\mathcal{L}_{\text{LM}}^{(0)}(c) \\ + \lambda_1\mathcal{L}_{\text{LM}}^{(1)}(o, a, p, f),\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{mask}}$ is binary cross-entropy plus dice loss, and $\mathcal{L}_{\text{LM}}^{(0)}$ and $\mathcal{L}_{\text{LM}}^{(1)}$ supervise the observation chain $c = (s, o_{\text{list}}, r, e)$ and the instruction-conditioned chain (o, a, p, f), respectively. We use $\lambda_{\text{mask}} = 1.0$, $\lambda_0 = 0.5$, and $\lambda_1 = 1.0$. The mask loss back-propagates through the `[SEG]` hidden state into the second pass, while the observation pass is supervised by its language-modeling loss.

Supervision. The instruction-conditioned levels (o, a, p, f) are taken from the ground-truth annotations of *Intent2Part* (Section 5). The observation levels $(s, o_{\text{list}}, r, e)$ are synthesized offline with a vision-language model using an image-only prompt, so they describe the scene without reference to any instruction. During training, the instruction q is sampled from a pool containing the target-referential instructions inherited from InstructPart and our intent-level instruction, exposing the model to both direct and intent-level query distributions.

Scheduled sampling. At inference, the second pass conditions on the context generated by the first pass. During early training, this context is noisy, so we first condition on the synthesized observation context and gradually replace it with self-generated contexts after a warmup of S_{warmup} steps. The replacement probability increases linearly up to $p_{\max} = 0.5$. Self-generated contexts are detached before entering the second pass, while the observation pass is always supervised by its language-modeling loss. This curriculum reduces the train-test gap while preserving early optimization stability.

Parameterization. The vision-language backbone is fine-tuned with LoRA adapters applied to attention and MLP projections. The SAM2 vision encoder is frozen, and the SAM2 mask decoder is trained. Full implementation details and hyperparameters are reported in Appendix A.

5 The Intent2Part Benchmark

To evaluate intent-level segmentation, we construct ***Intent2Part***, a benchmark in which instruc-

tions express high-level intents and the target is an affordance-bearing object part. Intent2Part builds on InstructPart (Wan et al., 2025), which provides images, object and part labels, part affordances, high-level actions, part masks, and target-referential instructions. We retain these annotations and add one intent-level instruction for each image-annotation pair.

5.1 Construction

Intent-level instructions. For each sample, we author an instruction that names only the desired outcome, without mentioning the target object, target part, action, or affordance. To reduce language and stylistic bias from any single generator, we generate candidate instructions independently with three frontier vision-language models: `Claude Opus 4.7`, `GPT-5.5`, and `Gemini 3.1 Pro`. Each model is prompted in a first-person voice with the image and a brief description of a plausible user scenario. We then filter candidates with an automatic validator that enforces first-person intent form, length constraints, and exclusion of target-revealing terms and their variants. Candidates that fail validation are rewritten manually. The full generation prompt, validator, and quality-control procedure are given in Appendix B.

Observation-level annotations. To supervise SegWorld's proactive observation stage, we additionally synthesize image-level scene contexts for training images. Given an image without any instruction, a vision-language model produces a scene description, salient objects, spatial and functional relations, and plausible events. These annotations supervise the observation pass during training and are never used for evaluation.

5.2 Statistics and Splits

Intent2Part contains 1,800 training and 600 test image-annotation pairs, following the splits of InstructPart, and adds intent-level instructions to both training and test samples. We evaluate on two test splits: **test_official**, the original 600-sample InstructPart test split, and **test_clean**, a 232-sample subset whose base images do not appear in training. Since **test_official** contains repeated base images with different part annotations, **test_clean** serves as our primary split for measuring image-level generalization. Each sample is evaluated under both the inherited target-referential instruction and the newly authored intent-level instruction.

6 Experiments

6.1 Setup

Datasets and metrics. All experiments are conducted on *Intent2Part* (Section 5). Models are trained on the 1,800-sample training split and evaluated under two instruction conditions: *target-referential* instructions inherited from InstructPart and our newly authored *intent-level* instructions. Intent-level evaluation is reported on *test_clean* (232 source-disjoint samples; primary split) and *test_official* (600 samples). We report mean Intersection-over-Union (mIoU), cumulative IoU (cIoU), and the `[SEG]` emission rate, defined as the fraction of samples for which the model produces a valid segmentation token.

Baselines. We compare with representative reasoning and affordance segmentation baselines: LISA (Lai et al., 2024), Sa2VA (Yuan et al., 2025), Affordance-R1 (Wang et al., 2026), and SAM3-I (Li et al., 2025). We also include a frozen **SAM3 Agent** cascade that uses Qwen-VL (Bai et al., 2025) to translate the instruction into a target description and SAM3 (Carion et al., 2025) to predict the mask. Trainable baselines are fine-tuned on *Intent2Part* with the same data and training budget as SegWorld. We additionally report off-the-shelf Sa2VA without task-specific fine-tuning.

SegWorld variants. SegWorld is built on Sa2VA-Qwen2.5-VL-7B. **SegWorld v1** is trained only with target-referential instructions and is evaluated zero-shot on intent-level instructions. **SegWorld v2** adds our intent-level instructions to the training pool with 20% sampling probability. Implementation details are provided in Appendix A.

6.2 Qualitative Examples

Figure 4 compares SegWorld with representative baselines on intent-level instructions. Existing methods often identify objects that are semantically related to the intent but fail to isolate the actionable part: bottles instead of pourable regions, whole benches or couches instead of seats, and entire appliances or luggage instead of handles. SegWorld more consistently grounds the intent to the affordance-bearing part, reflecting the benefit of explicitly reasoning through object, action, part, and affordance before emitting the segmentation token.

Figure 4: Qualitative comparison on intent-level instructions. Existing methods often ground the intent to a semantically related whole object or broad region, while SegWorld localizes the affordance-bearing part needed to realize the intent, such as a bottle body for pouring, a bench seat for sitting, an oven handle for opening, or a suitcase handle for carrying.

6.3 Main Results

Table 1 quantifies the trends shown in Figure 4, comparing SegWorld with existing segmentation systems across target-referential and intent-level instructions.

SegWorld preserves target-referential segmentation. On the standard referring and reasoning conditions, SegWorld v1 already improves over the strongest fine-tuned baseline, and v2 maintains this performance after adding intent-level supervision. This shows that proactive observation and multi-level reasoning do not trade off against conventional instruction-driven segmentation.

Proactive reasoning transfers to intent-level instructions. Without seeing intent-level instructions during training, SegWorld v1 achieves 0.514 mIoU on intent-level *test_clean*, outperforming the strongest baseline, Affordance-R1, by 3.4 points. Since v1 is trained only on target-referential instructions, this gain indicates that the two-pass reasoning structure learns a reusable intent-to-part inference pattern rather than merely memorizing intent templates.

Intent supervision further improves intent-to-part grounding. Adding intent-level instructions in SegWorld v2 raises mIoU on *test_clean* from 0.514 to 0.612, while leaving target-referential performance essentially unchanged. Compared with Affordance-R1, which applies MLLM-based affordance reasoning to this setting, SegWorld v2 improves intent-level mIoU on *test_clean* from 0.480 to 0.612. Although *test_clean* is source-disjoint from training, it yields higher scores than *test_official*; Section 6.4 shows that the overlapping-image subset of *test_official* contains conflicting part annotations and is empirically harder.

6.4 Analysis

Reasoning aligns with actionable parts. Figure 5 visualizes SegWorld's grounding behavior. The grounding heatmaps concentrate on the part that realizes the intent before the final mask is produced, such as a chair seat, sofa cushion, or bed surface. The intent-region similarity matrix shows a bright diagonal, indicating that matched intent-image pairs align in the learned joint space. These visualizations suggest that SegWorld's gains come from reasoning toward affordance-bearing parts rather than merely selecting semantically related objects.

Two-pass training induces mask commitment. On intent-level *test_official*, off-the-shelf Sa2VA

	Referring		Reasoning		Intent-level				Per-Action mIoU			
	InstructPart		*InstructPart*		*test_clean*		*test_official*		*Intent, test_official*			
Method	mIoU	cIoU	mIoU	cIoU	mIoU	cIoU	mIoU	cIoU	hold	cut	sit	**[SEG]%**
vanilla Sa2VA	0.510	0.367	0.498	0.359	0.151	0.099	0.141	0.093	0.178	0.073	0.075	57.0
SAM3 Agent	0.592	0.426	0.580	0.418	0.387	0.252	0.366	0.238	0.150	0.502	0.548	98.0
LISA-finetune	0.583	0.420	0.572	0.412	0.221	0.144	0.205	0.133	0.155	0.205	0.265	**100.0**
UniPixel-finetune	0.627	0.451	0.615	0.443	0.253	0.164	0.241	0.157	0.162	0.290	0.335	**100.0**
Affordance-R1	0.685	0.493	0.672	0.484	0.480	0.313	0.420	0.273	0.175	0.610	0.660	99.0
SAM3-I	0.701	0.505	0.689	0.496	0.342	0.222	0.318	0.207	0.165	0.445	0.498	**100.0**
Sa2VA-finetune	0.712	0.513	0.703	0.506	0.305	0.198	0.283	0.184	0.170	0.390	0.410	92.0
SegWorld v1	0.782	0.563	0.770	0.554	0.514	0.345	0.446	0.310	0.184	0.655	0.729	**100.0**
SegWorld v2	**0.783**	**0.564**	**0.772**	**0.556**	**0.612**	**0.411**	**0.495**	**0.345**	**0.253**	**0.734**	**0.775**	**100.0**

Table 1: Main results on *Intent2Part*. We report mIoU/cIoU across target-referential and intent-level instructions, with *test_clean* as the primary intent-level split. Per-action mIoU covers three representative actions, and **[SEG]%** denotes valid mask-token emission. Best results are in **bold**.

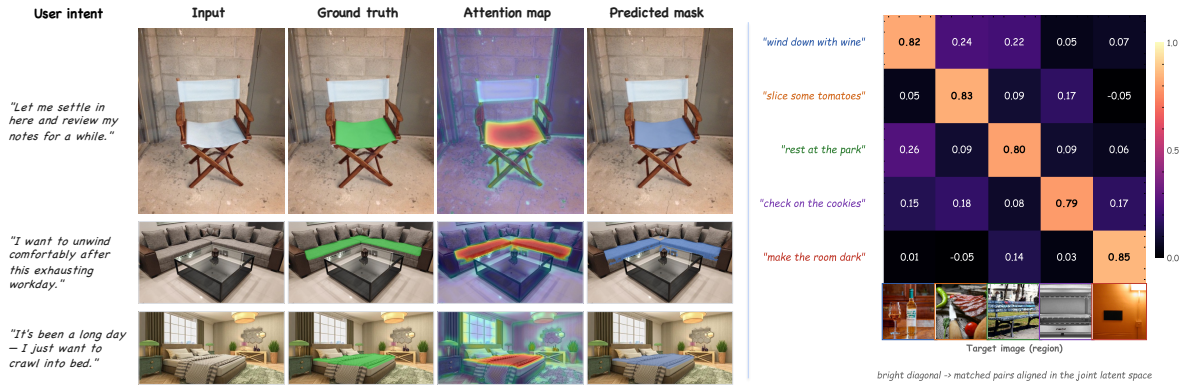

Figure 5: Reasoning visualization for SegWorld. Left: attention maps focus on the actionable part before mask prediction, and the predicted masks align with the affordance-bearing regions required by the intent. Right: intent-region similarity scores form a bright diagonal, showing that matched intent-region pairs align in the learned joint space.

fails to emit [SEG] on 43.0% of samples, often responding conversationally instead of producing a mask. Both SegWorld variants emit valid [SEG] tokens on all samples. This suggests that the two-pass training procedure changes the model's behavior from answering an intent in language to committing the intent to a segmentation output.

Image overlap introduces part conflict. Although *test_clean* is source-disjoint from training, it yields higher intent-level scores than *test_official*. The latter includes 368 overlapping-image samples whose base images appear in training with different object-part annotations. These samples are harder: their implied mIoU is 0.403 for SegWorld v1 and 0.422 for v2, both below the corresponding *test_clean* scores. The gap therefore reflects part-level conflict, not image memorization.

Robustness to human-written intents. Three annotators independently rewrote the intents for 60 *test_clean* samples without seeing generator outputs. Table 2 shows that the ranking is preserved: SegWorld v2 achieves 0.581 mIoU, compared with 0.452 for the strongest baseline. Its gain over v1 remains 0.093, close to the 0.098 gain on the original intents from the same samples, indicating that the improvement is not explained solely by the lexical distribution of the intent generators.

Model	Human-written	LLM-authored
Affordance-R1	0.452	0.474
Sa2VA-finetune	0.276	0.298
SegWorld v1	0.488	0.509
SegWorld v2	**0.581**	**0.607**

Table 2: Intent-level mIoU on 60 *test_clean* samples with independently human-written intents. LLM-authored results use the same samples.

Per-action behavior. Per-action results, reported in Appendix C, show that intent supervision is es-

Variant	*test_clean*		*test_official*	
	mIoU	cIoU	mIoU	cIoU
SegWorld v2	**0.612**	**0.411**	**0.495**	**0.345**
SegWorld-joint	0.549	0.352	0.457	0.317
w/o intent supervision (v1)	0.514	0.345	0.446	0.310
w/o event level	0.573	0.386	0.472	0.329
w/o proactive context	0.438	0.294	0.382	0.266
w/o Stage-1 CoT	0.402	0.270	0.358	0.249

Table 3: Ablation on intent-level segmentation. *SegWorld-joint* performs all eight reasoning levels in one instruction-conditioned pass. Separating instruction-free observation from intent grounding improves mIoU by 6.3 points on *test_clean*.

pecially helpful for small interaction parts such as handles, knobs, and straps, where the intended object is easy to identify but the actionable part is spatially small.

Ablation. Table 3 isolates the contribution of SegWorld's reasoning components. Removing the event level weakens performance, indicating that plausible scene interactions provide useful priors beyond static scene description. Removing the proactive context causes a larger drop, showing that instruction-free observation is a key source of intent-level grounding. The largest degradation comes from removing the Stage-1 chain, confirming that explicit object-action-part-affordance reasoning is not merely explanatory text but contributes directly to mask prediction. SegWorld-joint merges observation and intent grounding into one instruction-conditioned pass while retaining the same eight reasoning levels, data, and training budget. Its 0.549 mIoU is 6.3 points below SegWorld v2 on *test_clean*, showing that the gain is not due only to a longer reasoning trace. With the instruction visible from the start, observation is no longer instruction-independent and loses its role as a reusable scene prior.

7 Conclusion

We introduced ***SegWorld***, a proactive framework for intent-level segmentation. SegWorld first observes the scene without an instruction, then reasons from a high-level intent to an object, action, affordance-bearing part, and mask through a vision-grounded chain-of-thought. We also introduced ***Intent2Part***, a benchmark for evaluating part segmentation from high-level goals. Experiments show that SegWorld maintains strong target-referential performance while substantially improving intent-level grounding, suggesting a step toward segmentation systems that reason about what a scene affords rather than only following instructions.

Limitations

SegWorld studies intent-level segmentation in a controlled part-segmentation setting. First, Intent2Part focuses on static images and single-step intents. Many embodied scenarios involve temporal context, multi-step plans, or changing object states, which are outside the scope of this work. Second, our intent-level instructions are authored with the aid of frontier vision-language models and filtered by automatic validators. Although this reduces direct target leakage and single-model language bias, the resulting language distribution may still differ from naturally collected human instructions. Third, SegWorld makes its reasoning explicit through language, which improves inspectability but also makes the system sensitive to the quality of intermediate textual predictions: errors in scene context, object selection, or action inference can propagate to the final mask. Finally, the benchmark defines one ground-truth affordance-bearing part per intent, while real scenes may admit multiple valid ways to satisfy the same goal. Our evaluation therefore does not measure multi-solution intent grounding.

Ethical Considerations

SegWorld is designed for intent-level part segmentation and may be useful for assistive interfaces, embodied agents, and human-centered visual grounding. As with other vision-language systems, its predictions can be affected by biases in the underlying models and datasets, including cultural assumptions about common goals, household objects, and typical object use. Incorrect intent grounding may also lead to inappropriate or unsafe action suggestions if used directly in a robotic system. Our work only predicts segmentation masks and does not execute actions; any deployment in embodied settings should include human oversight, task-specific safety checks, and evaluation under the target environment. The intent-level instructions in Intent2Part are generated and filtered to avoid revealing target regions, but they may not fully reflect the diversity of naturally occurring human language. We will release the dataset construction protocol and validation rules to support reproducibility and auditing.

Acknowledgments

Our work was supported in part by the Guangdong Provincial Key Laboratory of IRADS (2022B1212010006), and in part by Guangdong Higher Education Upgrading Plan, and in part by Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme (2025A0505000004), and in part by the National Key Research, and Development Program of China (2024YFF1206600), and in part by the MSAI Conference Funding of MSAI program of Northwestern University.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, and 1 others. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.

Hengshuo Chu, Xiang Deng, Qi Lv, Xiaoyang Chen, Yinchuan Li, Jianye Hao, and Liqiang Nie. 2025. 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. *arXiv preprint arXiv:2502.20041*.

Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 2021. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1778–1787.

Thanh-Toan Do, Anh Nguyen, and Ian Reid. 2018. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE.

Sitong Gong, Lu Zhang, Yunzhi Zhuge, Xu Jia, Pingping Zhang, and Huchuan Lu. 2025. Reinforcing video reasoning segmentation to think before it segments. *arXiv preprint arXiv:2508.11538*.

Yuchen Guo, Junli Gong, Wenjun Dong, Yiuming Cheung, and Weifeng Su. 2026. Adding thermal awareness to visual systems in real-time via distilled diffusion models. *arXiv preprint arXiv:2605.06010*.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, and 1 others. 2025. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*.

Donggon Jang, Yucheol Cho, Suin Lee, Taehyeon Kim, and Dae-Shik Kim. 2025. Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. *International Conference on Learning Representations*.

Shiu-hong Kao, Chak Ho Huang, Huaiqian Liu, Yu-Wing Tai, and Chi-Keung Tang. 2026. Cot-seg: Rethinking segmentation with chain-of-thought reasoning and self-correction. *arXiv preprint arXiv:2601.17420*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589.

Jingjing Li, Yue Feng, Yuchen Guo, Jincai Huang, Wei Ji, Qi Bi, Yongri Piao, Miao Zhang, Xiaoqi Zhao, Qiang Chen, and 1 others. 2025. Sam3-i: Segment anything with instructions. *arXiv preprint arXiv:2512.04585*.

Chang Liu, Henghui Ding, and Xudong Jiang. 2023. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Chen. 2026. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. *Advances in Neural Information Processing Systems*, 38:126078–126108.

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Yan Feng, Bin Chen, Ke Chen, and Yaowei Wang. 2025. Coprs: Learning positional prior from chain-of-thought for reasoning segmentation. *arXiv preprint arXiv:2510.11173*.

Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. 2016. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300.

Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. 2021. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823.

Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. 2019. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. 2024. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, and 1 others. 2025. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pages 28085–28128.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Yihe Tang, Wenlong Huang, Yingke Wang, Chengshu Li, Roy Yuan, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. 2025. Uad: Unsupervised affordance distillation for generalization in robotic manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3822–3831. IEEE.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 25, pages 1507–1514.

Zifu Wan, Yaqi Xie, Ce Zhang, Zhiqiu Lin, Zihan Wang, Simon Stepputtis, Deva Ramanan, and Katia P Sycara. 2025. Instructpart: Task-oriented part segmentation with instruction reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24202–24227.

Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. 2026. Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9738–9746.

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

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and 1 others. 2025. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*.

A Training and Implementation Details

Model. SegWorld is built on Sa2VA-Qwen2.5-VL-7B. The vision-language backbone is fine-tuned with LoRA adapters applied to attention and MLP projections, with rank $r = 64$, $\alpha = 128$, and dropout 0.05. The SAM2 vision encoder is frozen throughout training, while the SAM2 mask decoder and the projection from the `[SEG]` hidden state to the decoder prompt embedding are trained.

Optimization. We train with AdamW using learning rate 2×10^{-5}, weight decay 0.05, bf16 precision, and DeepSpeed ZeRO-2. The effective batch size is 16, with batch size 1 per device, 8-step gradient accumulation, and 2 A100-80G GPUs. Each run is trained for 2,250 optimizer steps and takes approximately 4 hours. Gradient checkpointing is disabled to avoid double gradient reduction across the two-pass forward computation.

Scheduled sampling. During inference, Stage 1 conditions on the scene context generated by Stage 0. During early training, Stage 1 is conditioned on the synthesized Stage-0 context. After a warmup of S_{warmup} steps, we gradually replace the synthesized context with the model's self-generated context. The replacement probability is $p_{\text{self}}(t) = \min(t/S_{\text{warmup}}, 1) \cdot p_{\max}$, with $p_{\max} = 0.5$. Self-generated contexts are decoded under `no_grad` and detached before entering Stage 1. Stage 0 is always supervised with its language-modeling loss.

B Intent2Part Construction Details

Intent-level instruction generation. For each image-annotation pair, we generate candidate intent-level instructions with three frontier vision-language models: `Claude Opus 4.7`, `GPT-5.5`, and `Gemini 3.1 Pro`. Each model receives the image and a brief description of a plausible user scenario, and is prompted to write a first-person instruction that expresses a desired outcome without naming the target object, part, action, or affordance. Using multiple generators reduces language and stylistic bias from any single model.

Validation. Each candidate instruction is checked by an automatic validator. The validator requires a first-person intent form, length between

6 and 25 words, and absence of the ground-truth `object`, `part`, `action`, and `affordance` terms. It also filters plural forms, common morphological variants, and near-synonyms that would reveal the target region. If multiple candidates pass validation, we select the most natural one. If no candidate passes, the instruction is manually rewritten and re-validated. Every released intent-level instruction passes the validator.

Observation-level annotations. Stage-0 supervision is synthesized with a vision-language model using an image-only prompt. The model is asked to produce four fields: a one-sentence scene description, a list of salient visible objects, spatial and functional relations among them, and plausible events supported by the scene. The prompt does not include any target-referential or intent-level instruction. These annotations are used only for training the proactive observation stage and are never used during evaluation.

Split construction. Intent2Part follows the original InstructPart train/test split, with 1,800 training and 600 test image-annotation pairs. Because InstructPart may contain multiple annotations for the same base image, the official test split includes some images whose base image also appears in training with a different object-part annotation. We therefore define *test_clean* as the subset of test samples whose base image does not appear in training. This yields 232 source-disjoint test samples and 368 overlapping-image test samples. We report both *test_official* and *test_clean*, using *test_clean* as the primary split for image-level generalization.

Action	n	vanilla	v1	v2
hold	93	0.178	0.184	**0.253**
cut	72	0.073	0.655	**0.734**
sit	71	0.075	0.729	**0.775**
pick up	66	0.101	0.152	**0.482**
open	50	0.020	0.455	**0.565**
support	44	0.234	**0.459**	0.413
grasp	33	0.147	0.379	**0.405**
contain	32	0.405	0.402	**0.627**

Table 4: Per-action mIoU on intent-level *test_official*.

C Per-Action Results

Table 4 reports intent-level mIoU on *test_official* for the eight most frequent action categories. SegWorld v1 already improves substantially over off-the-shelf Sa2VA on actions requiring localized part grounding, while v2 further improves categories where the actionable part is small relative to the object, such as *pick up* and *open*.

D Additional Evaluation of Intent Reasoning

This section evaluates whether Intent2Part measures scene-conditioned intent-to-part reasoning beyond conventional part grounding, and whether SegWorld's explicit chain recovers the intended intermediate semantics. Unless otherwise specified, all results are reported on the intent-level *test_clean* split.

D.1 Text-Only Resolution and Modular Cascades

Hypothesis-only resolution. We first measure how much of the target can be inferred from the intent text alone. We prompt `GPT-5.5` with the intent string, without the image, to predict the target object and part labels. Correctness is normalized exact match; semantic equivalents are adjudicated by `GPT-4o`, and all judge-flagged cases are verified by the authors. Table 6 shows that text-only script knowledge resolves the object more often than the specific object-part pair. Thus, although intents encode useful commonsense priors, the part-level target remains substantially underdetermined without scene information.

Unconfounded modular cascade. We next test whether intent-level segmentation can be reduced to an external language resolver followed by a conventional referring grounder. We use Sa2VA-finetune as the grounder because it is trained on the same 1,800 samples, with the same budget as SegWorld, and achieves 0.712 mIoU on referring instructions. The resolver predicts an object-part pair, which is rendered as the template "the *part* of the *object*" and passed to Sa2VA-finetune. This construction separates intent resolution from part grounding without the training-data and part-granularity confounds of a frozen cascade.

Table 7 reports both a text-only resolver and an image-conditioned resolver. Even when the frontier resolver observes the image, routing its prediction through a discrete object-part interface remains below SegWorld. This gap suggests that the composition of scene-conditioned intent reasoning and pixel-level grounding is not recovered by a hard textual intermediate representation alone.

Model	Ref. ✓	Intent ✓	Ref. ✓ Intent ×	$P(\text{intent}✓ \mid \text{ref.}✓)$
Sa2VA-finetune	78.0	31.0	49.6	0.36
SAM3 Agent	63.4	40.1	26.3	0.59
Affordance-R1	74.1	50.0	29.3	0.60
SegWorld v2	**84.1**	**65.5**	**19.8**	**0.76**

Table 5: Per-sample contingency analysis (%) on *test_clean*. Higher is better except Ref. ✓, Intent ×.

Resolver	Object	Part	Object–Part
GPT-5.5 (text only)	61.2	47.4	38.8

Table 6: Text-only label resolution accuracy (%) on intent-level *test_clean*.

Method	mIoU	cIoU
Text-only GPT-5.5 → Sa2VA-ft	0.341	0.222
Image-conditioned GPT-5.5 → Sa2VA-ft	0.487	0.319
SegWorld v2	**0.612**	**0.411**

Table 7: Intent-level segmentation with modular resolver–grounder cascades. The text-only resolver receives only the intent; the image-conditioned resolver receives both the image and intent.

Reasoning level	v1	v2
Stage-0 OBJECTS recall	91.4	**92.2**
Stage-0 EVENTS hit	62.9	**69.0**
Stage-1 OBJECT	78.4	**84.5**
Stage-1 ACTION	74.1	**81.0**
Stage-2 PART	65.5	**72.4**
Stage-2 AFFORDANCE	67.2	**74.1**

Table 8: Intermediate reasoning accuracy (%) on intent-level *test_clean*.

D.2 Contingency with Referring Segmentation

Each *test_clean* sample has the same ground-truth mask paired with both a target-referential and an intent-level instruction. We therefore measure per-sample contingency between the two conditions. A prediction is considered correct when its IoU is at least 0.5. The *referring-correct, intent-wrong* cell identifies samples for which a model can localize the part when it is described explicitly, but cannot recover it from the corresponding high-level intent.

Table 5 shows that intent-level segmentation does not merely re-measure part grounding. For Sa2VA-finetune, 49.6% of samples are grounded correctly under a referring expression but fail under the corresponding intent. SegWorld v2 reduces this gap to 19.8% and raises $P(\text{intent}✓ \mid \text{ref.}✓)$ from 0.36 to 0.76, indicating improved recovery of actionable parts from intent.

D.3 Intermediate Chain Evaluation

We directly evaluate the semantically supervised intermediate levels of SegWorld. For Stage 0, OBJECTS recall measures whether the ground-truth object is listed in the generated scene context, and EVENTS hit measures whether the ground-truth action occurs among the predicted plausible events. For Stages 1–2, we evaluate normalized exact match for the generated object, action, part, and affordance labels. Semantic equivalents are adjudicated with GPT-4o and verified by the authors when flagged.

Intent supervision improves every intermediate level (Table 8), with the largest gains on the object, action, part, and affordance decisions that connect intent to pixels. We further measure error propagation for SegWorld v2. The mean mask IoU is 0.706 when the selected object is correct, compared with 0.098 when it is incorrect, and $P(\text{part}✓ \mid \text{object}✓) = 0.86$. Among intent-level samples with IoU below 0.5, 45% select the wrong object, 35% select the correct object but the wrong part, and 20% generate a correct chain but produce an imprecise mask. These results locate most remaining errors in scene-conditioned object and part selection rather than in mask decoding.

D.4 Oracle Intent Resolution

To quantify how much of the referring–intent gap is attributable to intent resolution, we inject the ground-truth object and action as the Stage-1 outputs and allow SegWorld v2 to generate the remaining part, affordance, and [SEG] token. This oracle raises intent-level mIoU from 0.612 to 0.734 on *test_clean*. Relative to the 0.791 mIoU obtained under the corresponding referring instruction, the oracle closes $(0.734-0.612)/(0.791-0.612) = 68\%$ of the referring–intent gap. Thus, most of the observed difficulty arises from scene-conditioned intent resolution, while the remaining gap reflects the challenge of grounding the selected part under intent-style language.