

Conversational Image Segmentation: Grounding Abstract Concepts with Scalable Supervision

Aadarsh Sahoo Georgia Gkioxari
 California Institute of Technology

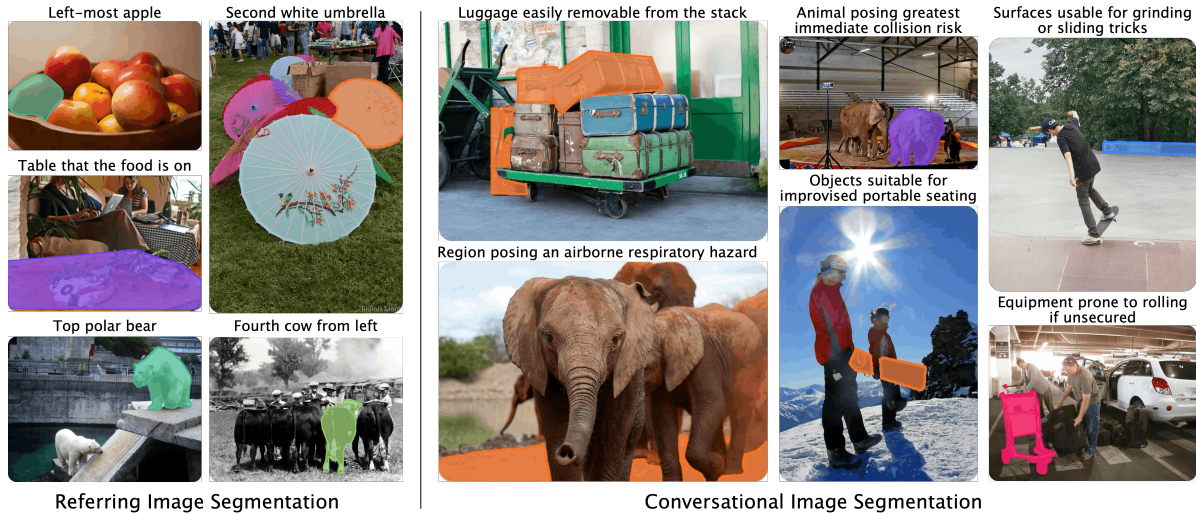


Figure 1. **Conversational Image Segmentation requires reasoning beyond object categories.** *Left:* Referring Image Segmentation (RIS) grounds descriptive phrases about object identity and spatial relations (e.g., “left-most apple”). *Right:* Conversational Image Segmentation (CIS) grounds abstract, intent-oriented concepts that require relational reasoning, physical understanding, and implicit constraints.

Abstract

Conversational image segmentation grounds abstract, intent-driven concepts into pixel-accurate masks. Prior work on referring image grounding focuses on categorical and spatial queries (e.g., “left-most apple”) and overlooks functional and physical reasoning (e.g., “where can I safely store the knife?”). We address this gap and introduce Conversational Image Segmentation (CIS) and CONVERSEG, a benchmark spanning entities, spatial relations, intent, affordances, functions, safety, and physical reasoning. We also present CONVERSEG-NET, which fuses strong segmentation priors with language understanding, and an AI-powered data engine that generates prompt-mask pairs without human supervision. We show that current language-guided segmentation models are inadequate for CIS, while CONVERSEG-NET trained on our data engine achieves significant gains on CONVERSEG and maintains strong performance on existing language-guided segmentation benchmarks. Project webpage: <https://glab-caltech.github.io/converseg>.

1. Introduction

“Which suitcases can I take without disturbing the stack?” For humans, the answer is immediate: we identify which pieces are load-bearing versus accessible, anticipate redistribution of weight, and filter candidates by the constraint of “easy removal.” Yet a segmentation model, trained to predict *suitcase* and *cart*, lacks any representation of support relations, occlusion ordering, or physical stability. Selecting easily removable luggage requires reasoning jointly over geometry, physics, and user intent – not merely recognizing object categories. This type of conversational, intent-driven language instruction reflects how humans naturally interact with their environments, yet remains beyond the reach of current perception systems.

In computer vision, grounding images using natural language expressions was first explored through the task of *Referring Image Segmentation* (RIS) [37]. However, existing benchmarks for this task, RefCOCO variants [37], primarily emphasize categorical and spatial references, such as “the white umbrella” or “the left-most apple”, shown on the left of Fig. 1. By contrast, functional or physical reasoning about objects and environments – such as “what object is prone to

rolling if unsecured” or “where can I safely store the knife?” – is largely underrepresented.

We address this gap by introducing *Conversational Image Segmentation (CIS)*, a task that grounds high-level conversational concepts into pixel-accurate masks in natural images. Examples are shown in Fig. 1. We call these concepts conversational because they mirror how humans naturally talk about objects and their surroundings. They span five families, inspired by human vision science [8, 24] and intuitive physics [2]: (1) **Entities** with open-vocabulary descriptions (“*weathered wooden furniture*”); (2) **Spatial & Layout** capturing complex geometric relations (“*items blocking the walkway*”); (3) **Relations & Events** describing interactions (“*the player about to catch the ball*”); (4) **Affordances & Functions** requiring use-case reasoning (“*surfaces suitable for hot cookware*”); and (5) **Physics & Safety** involving stability or hazard assessment (“*objects likely to tip over*”).

To measure progress in CIS, we introduce the CONVERSEG benchmark, featuring 1,687 human-verified image-mask pairs. Unlike previous benchmarks that focus on categorical entities and simple spatial relations, CONVERSEG offers coverage across all five concept families and a broader representation of conversational reasoning. We hope that advancing CIS with CONVERSEG will advance perception systems in assistive robotics, human-robot interaction and augmented reality – domains that require grounding abstract concepts.

We further introduce CONVERSEG-NET, a conversational segmentation model that maps an image and prompt to a grounding mask. Training it demands large-scale supervision over diverse prompts in natural images – a costly and cognitively intensive effort for human annotators who must produce pixel-accurate masks and natural, reasoning-rich prompts. To bypass this bottleneck, we build an automated, VLM-driven data engine that synthesizes high-quality prompt-mask pairs *without human supervision* via an iterative generate-and-verify loop, yielding 61K image-mask pairs across all five concept families. Trained on this data, CONVERSEG-NET achieves strong results on CONVERSEG and remains competitive on standard referring expression benchmarks, demonstrating both data quality and scalability.

In summary, our contributions are:

- We introduce *Conversational Image Segmentation (CIS)* and CONVERSEG, a benchmark of human-verified image-mask pairs targeting grounding of affordances, physics, and functional reasoning.
- We build an AI-powered data engine that synthesizes diverse, high-quality conversational prompt-mask pairs without human supervision.
- We design a baseline model, CONVERSEG-NET; trained on our engine’s data, it excels on CONVERSEG and remains strong on RIS benchmarks.

2. Related Work

Referring Expression Segmentation. Referring expression segmentation (RIS) localizes regions described by language. RefCOCO/+g [37] are standard benchmarks but are dominated by object-centric, low-level spatial phrases (e.g., “person on the left,” “red cup”). Early methods used multi-stage language-vision pipelines [11, 18]; recent work adopts Transformer-based vision-language encoders [13, 35]. Despite strong results on entities and simple spatial relations, these benchmarks seldom test affordances, stability, or user intent. Our CIS task and CONVERSEG explicitly target these gaps via five conversational concept families.

Reasoning and Implicit Segmentation. ReasonSeg [15] pairs images with implicit, reasoning-heavy instructions and masks, but queries still target entities or spatial relations, with limited coverage of affordances, safety, or physical constraints. Multi-modal LLM systems (LISA [15], GLaMM [27], PixelLM [30]) can perform multi-step reasoning and produce masks, yet rely on heavy backbones and multi-stage inference (chain-of-thought, tool calls), making deployment costly. In contrast, we pursue single-pass architectures that directly ground conversational concepts.

Promptable Segmentation Models. The Segment Anything Model (SAM) [14] enables promptable, class-agnostic segmentation from points or boxes; SAM2 [28] extends this to streaming video. These offer strong priors but lack native text conditioning. Bridging this, some [22, 29] pair SAM with text-conditioned detectors, while others integrate SAM-like decoders into VLMs [27, 39]. We leverage SAM’s learnt priors and combine them with lightweight vision-language adapters to enable end-to-end conversational grounding without sacrificing segmentation quality.

Vision-Language Models for Dense Prediction. Recent VLMs add heads for dense prediction: LISA [15] augments LLaVA [19] with a mask decoder, GLaMM [27] supports multi-turn grounded dialogue, and GroundHog [39], Kosmos-2 [26], and xGen-MM/BLIP-3 [33] push pixel-level grounding in ever-larger models. These systems excel at complex reasoning but demand substantial compute and often multiple passes per query. We take a complementary approach: rather than scaling model capacity, we scale *training data diversity* through automated data synthesis. A lightweight 3B VLM + SAM2 decoder, trained on 61K auto-generated prompt-mask pairs across five reasoning concepts, achieves competitive CIS and RIS performance.

Automated Data Synthesis. Synthetic data is a strong alternative to manual annotation, with synthetic pre-training approaching real-image performance for representation learning [31]. Pipelines like ELEVATER [16], synthetic region captioning [6, 36], and grounding verification [39] scale supervision, but mostly target literal descriptions or category la-



Figure 2. **CONVERSESEG Qualitative Examples.** Representative samples across five concept categories. Prompts require reasoning about attributes, spatial relations, interactions, functional properties, and physical constraints beyond standard object reference.

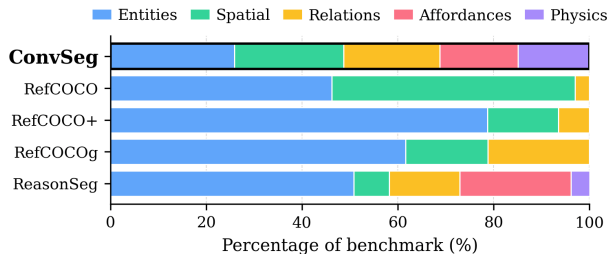


Figure 3. **Concept Coverage in Benchmarks.** Distribution of concepts across five existing benchmarks versus CONVERSESEG. Prior datasets primarily focus on entities/spatial relations, whereas CONVERSESEG offers near-uniform coverage across all five concepts.

bels. Our engine instead synthesizes conversational prompts aimed at affordances, layout constraints, and physical safety, then filters them via multi-stage visual verification.

3. Conversational Image Segmentation

We introduce the task of *Conversational Image Segmentation* (CIS) and the family of concepts that define it.

3.1. Task Definition

Given an image I and a natural language prompt p , the task is to predict a binary mask M_p identifying the pixels in I that satisfy the query p . Unlike referring image segmentation (RIS), prompts in CIS may require functional or physical reasoning (e.g., “surfaces stable enough to stack books”), target non-visible properties such as affordances or safety (e.g., “objects that might be hot”), and have context-dependent groundings (e.g., “comfortable seating areas”).

3.2. The Five Conversational Concepts

We organize prompts into five concept families according to the type of reasoning they require. These families reflect

how humans naturally query their surroundings and reveal reasoning abilities that extend beyond object-centric reference. We draw inspiration from human vision science [8, 24] and intuitive physics [2], which demonstrate that people infer functional properties and physical constraints directly from visual input – capabilities largely missing from current vision tasks and benchmarks.

1. **Entities.** Prompts that identify entities by category or attributes (e.g., “the bicycle with a basket”). While overlapping with RIS, we include open-vocabulary categories and complex attribute compositions.
2. **Spatial & Layout.** Prompts about spatial relations, ordering, and occupancy (e.g., “the rightmost orange in the bowl”, “the lamp behind the sofa”).
3. **Relations & Events.** Prompts targeting interactions or transient states (e.g., “the player serving the ball”, “the door being opened”).
4. **Affordances & Functions.** Prompts requiring functional reasoning about object use (e.g., “surfaces you could cut on”, “items that could serve as a shovel”).
5. **Physics & Safety.** Prompts invoking stability, support, and hazard assessment (e.g., “objects likely to tip”, “sharp objects posing a hazard”).

4. The CONVERSESEG Benchmark

In this section, we introduce the CONVERSESEG benchmark for conversational image segmentation and describe our data collection and annotation process.

Fig. 3 compares concept distributions across benchmarks. Existing datasets are heavily skewed toward entities and spatial relations (> 50%), with minimal coverage of affordances, events, or physical reasoning. In contrast, CONVERSESEG provides balanced representation across all five concepts.

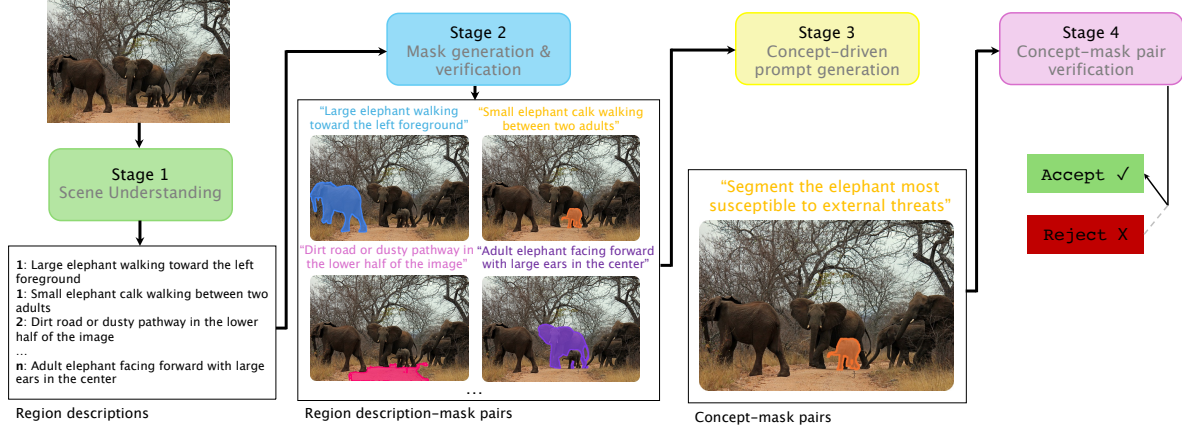


Figure 4. **Conversational Data Engine.** Five-stage pipeline for automated concept-mask synthesis: (1) VLM generates region descriptions; (2) Object detector and SAM2 produce masks, VLM filters and refines them. (3) Concept-driven meta-prompts generate conversational queries assigned to region-mask pairs; (4) Alignment verification ensures prompt-mask correctness.

4.1. Benchmark Construction and Verification

CONVERSESEG is constructed through a two-stage pipeline: automated candidate generation followed by human verification. We source images from the COCO [17] validation set and generate candidate tuples (I, p, M_p, c) where $c \in \mathcal{C}$ denotes one of five concept types.

Segmentation masks M_p in CONVERSESEG are either machine or human drawn, which lead to two evaluation sets:

Human-annotated Set. We initialize masks with human-drawn annotations from COCO, sourced from instance (objects) and panoptic annotations (objects and stuff). This split has high-quality mask annotations but is restricted to annotations within COCO.

SAM-seeded Set. Masks are extracted with SAM2 [28] prompted with bounding boxes generated by an object detector. This mode eliminates reliance on closed-vocabulary annotations from COCO and allows us to scale our evaluation data with minimal human supervision.

Conversational prompts p are collected with the help of our data engine, described in detail in Sec. 5, which proposes prompts p with corresponding masks M_p across all five concept families via a generate-then-verify process.

Human Verification. Each sample undergoes human verification for (i) prompt quality and correct concept assignment and (ii) mask accuracy. Verifiers accept/reject with a single click. The engine supplies image-tailored, diverse prompts across all five concept families; human checks ensure prompt and mask quality.

4.2. CONVERSESEG Statistics.

CONVERSESEG comprises 1,687 total samples across two splits: 1,194 SAM-seeded and 493 human-annotated. Prompts average 7.6 words (std: 1.2). Fig. 2 shows representative examples from each concept category. Additional

statistics, annotation protocols, and qualitative examples are provided in the Appendix.

5. The Conversational Data Engine

We introduce a fully automatic data engine for conversational image segmentation. Collecting pixel-accurate masks and, more critically, realistic prompts is prohibitively expensive at scale with human annotators. Our data engine synthesizes high-quality prompt-mask pairs *without human supervision*. Instead, it leverages high-performing VLMs that iteratively generate outputs and verify their quality.

5.1. Data Engine Architecture

Fig. 4 shows our pipeline, featuring modules for scene recognition, mask generation, prompt creation and verification.

Stage 1: Scene Understanding. Given an image I , a VLM generates 5–7 region descriptions in natural language, $\{d_1, \dots, d_n\}$. Each d_i specifies category, attributes, location, and relations in ≤ 15 words, e.g., “*large elephant walking toward left foreground*”. These descriptions serve as targets for mask generation.

Stage 2: Mask Generation. For each description d_i , we localize its segment in image I : Moondream3 [25] predicts a box b_i from (I, d_i) ; SAM2, prompted with b_i , returns the mask m_i . We choose Moondream3 for open-vocabulary detection and SAM2 for box-conditioned segmentation.

Stage 3: Mask Quality Verification. Pipeline systems suffer from error propagation, so we enforce two checks for mask fidelity:

Mask-text consistency check. A VLM verifies that (b_i, m_i) matches d_i in identity, attributes, and spatial location. It returns accept/reject; only accepts proceed to the next stage.

Mask refinement and selection. Passing masks may still have under/over-coverage or holes, often from noisy boxes. We

sample SAM2 with a dense point grid to form candidates and pick m'_i with highest IoU to m_i . A VLM then chooses the better of the two based on coverage, boundary precision, artifacts. The selected ($\hat{m}_i$) is the final mask.

Now, each image I has verified descriptions, d_i , paired with high-quality masks, $\hat{m}_i$.

Stage 4: Concept-Driven Prompt Generation. This stage derives image-tailored conversational prompts for I and its discovered regions.

Prompts must span a range of reasoning types within each concept c . For example, for affordances they should assess functional properties and context-dependent use ("surfaces safe for hot items"), canonical functions ("sources of water"), or counterfactual uses ("items that could prop a door"); for spatial relations, containment ("items inside containers") or ordinality ("the three leftmost cups"). We enforce this via concept-specific meta-prompts π_c (found in the Appendix).

For each c , a VLM receives: (1) indexed descriptions $\{d_i\}$, (ii) set-of-marks numbered overlay [34], and (iii) meta-prompt π_c . It generates up to three prompts p per concept and selects the corresponding regions. Trivial prompt-mask pairs (e.g., an image with one car and "segment the car") or prompts applicable to undiscovered regions are pruned.

Stage 5: Prompt-Mask Alignment Verification. For each tuple (I, p, M_p, c) , a VLM verifies that M_p : (1) matches p 's target, (2) excludes irrelevant content, and (3) is reasonably described by p . It returns accept/reject; only accepts proceed.

This five-stage pipeline combines VLM-driven generation with multi-stage verification to produce high-quality prompt-mask pairs without human annotation. All VLM components use Gemini-2.5-Flash [4].

5.2. Dual-Purpose Design

Our data engine is crafted to generate diverse conversational prompt-mask pairs in images. We use it for two purposes: **(1) Benchmark curation** (Sec. 4.1). We run the pipeline on COCO val images, seeding mask candidates from COCO annotations or SAM2. For benchmark-grade quality, generated prompt-mask pairs receive a final human check.

(2) Training data synthesis (Sec. 6.2). We run the data engine at scale on COCO train images, producing training pairs without human supervision. Ablations on pipeline components, failure mode analysis, and additional implementation details are in the Appendix.

6. Model

In this section, we present a single-pass model for conversational image segmentation that *grounds conversational concepts into pixels*. Our design philosophy is to avoid multi-component workflows like iterative tool use or multi-turn refinement to build a strong baseline for CIS. To this

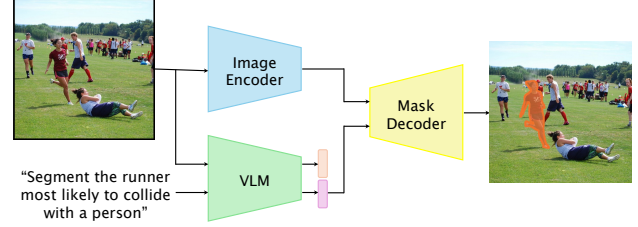


Figure 5. **Model Architecture.** An image encoder processes the image, while a VLM jointly encodes image and text. Lightweight adapters map the text-token embeddings from the VLM to a mask decoder that predicts the target segment.

end, we combine advances in promptable image segmentation, but lack text conditioning, with VLMs that integrate vision and language but do not perform segmentation. We describe our model architecture (Sec. 6.1), training strategy and implementation details (Sec. 6.2).

6.1. Model Architecture

Fig. 5 illustrates our architecture. It fuses SAM2 [28] (image-only components) with a compact vision-language backbone (Qwen-2.5-VL-3B [1]), connected via lightweight prompt adapters.

Image Encoder. We adopt the SAM2 image encoder which is an MAE [9] pre-trained Vision Transformer (ViT) [7] adapted for high-resolution inputs and keep it frozen. The encoder processes each image once, independent of the prompt, producing a spatial image embedding $\mathbf{z}_{\text{img}} \in \mathbb{R}^{H' \times W' \times D_{\text{img}}}$.

Prompt Encoder. We use Qwen2.5-VL-3B [1] as the prompt encoder, which jointly processes the image I and text prompt p to produce per-token hidden states at the final layer. We extract hidden states corresponding to text tokens only (they have attended to the image tokens through Qwen’s backbone), yielding a sequence $\{\mathbf{h}_1, \dots, \mathbf{h}_T, \mathbf{h}_{\text{EOS}}\} \in \mathbb{R}^{D_t}$, where T is the text length and D_t is the hidden dimension.

Following SAM’s design, we represent prompts as sparse and dense embeddings. The text token sequence $\{\mathbf{h}_1, \dots, \mathbf{h}_T\}$ serves as sparse embeddings, capturing fine-grained text information. The hidden state at the EOS position serves as the dense embedding, capturing global image-text context. Two lightweight adapters project these representations to the decoder’s input space: $\mathbf{e}_{\text{sparse}} = \text{Linear}_{D_t \rightarrow D_{\text{dec}}}(\{\mathbf{h}_1, \dots, \mathbf{h}_T\})$ and $\mathbf{e}_{\text{dense}} = \text{MLP}_{D_t \rightarrow D_{\text{dec}}}(\mathbf{h}_{\text{EOS}})$, where the dense adapter is a 2-layer MLP with SiLU activation. The Qwen backbone is fine-tuned using LoRA [10] with rank 16 and $\alpha = 32$.

Mask Decoder. We adopt SAM2’s mask decoder and fully fine-tune it. The decoder uses modified Transformer blocks [32] with bidirectional cross-attention between prompt and image embeddings. After two blocks, the image embedding is upsampled and an MLP maps the output token to per-pixel foreground probabilities, producing the final mask.

6.2. Training Curriculum

To tackle the reasoning-heavy task of conversational image segmentation, we use a curriculum that gradually increases task complexity: the model first learns to segment literal concepts (*e.g.*, “segment the cat”) before advancing to more abstract referential and conversational concepts. This approach is critical because SAM2 has no prior exposure to language and thus requires careful integration of text conditioning. We validate this design via ablations in Sec. 7.3.

6.2.1. Training Data

We organize training data into four groups with increasing complexity:

(1) Literal concepts. We leverage the COCO train set reformulated as category-level segmentation using refined masks from COCONut [5]. For each image, we randomly sample a category and treat all corresponding masks as ground truth with prompts like “Segment all the [category] in the image”. This yields 117K prompt-mask pairs.

(2) Basic referring expressions. We utilize the RefCOCO family datasets [37] (RefCOCO, RefCOCO+, RefCOCOg) train splits, providing 321K object-centric references.

(3) Open-vocabulary regions. Region descriptions from Stage 3 of our data engine (Sec. 5), providing 27K diverse region descriptions beyond COCO’s closed vocabulary. Examples are shown in Fig. 4.

(4) Conversational concepts. This is the output from our data engine which generates 61K concept-mask pairs, covering all five concept families.

6.2.2. Two Phase Training

We train in two phases with increasing task complexity:

Phase 1: Pretraining. We pretrain on a mixture of groups 1-3, producing a base model proficient at basic referring segmentation.

Phase 2: Conversational post-training. Initialized from Phase 1, we fine-tune on group 4 mixed with an equal number of samples randomly drawn from groups 1-3 with a lower learning rate. This mixing strategy maintains performance on foundational segmentation tasks while adapting to conversational concepts, empirically validated in Sec. 7.3.

6.2.3. Training Objective

Given binary ground-truth mask $M \in \{0, 1\}^{H \times W}$ and predicted probability mask $M^* \in [0, 1]^{H \times W}$, we minimize a weighted combination of binary cross-entropy and Dice loss: $\mathcal{L} = \mathcal{L}_{\text{BCE}}(M, M^*) + \lambda \mathcal{L}_{\text{Dice}}(M, M^*)$ with $\lambda = 0.25$.

6.2.4. Implementation Details

We train with AdamW [23], batch size 6, and a cosine schedule with warmup. Phase 1 runs 35K steps at $\eta_1=1\text{e}-4$;

Phase 2 runs 35K steps at $\eta_2=1\text{e}-5$. Training on a single NVIDIA A100 80GB takes about 48 hours. Additional details are in the Appendix.

7. Experiments

We evaluate our model against baselines on conversational image grounding, outlining the setup (Sec. 7.1), results (Sec. 7.2), and ablation studies and analysis (Sec. 7.3).

7.1. Experimental Setup

Benchmarks. We evaluate on three benchmarks: (1) CONVERSESEG: our benchmark with SAM-seeded and human-annotated splits across five concept categories (entities, spatial relations, affordances, relations & events, physics & safety); (2) RefCOCO+/g [37]: standard referring expression benchmarks; (3) ReasonSeg [15]: segmentation benchmark of complex implicit reasoning and understanding.

Baselines. We compare against state-of-the-art feed-forward methods of diverse architectural designs: LISA [15], an MLLM-based model with embedding-as-mask paradigm; UniLSeg [20], a universal model for arbitrary semantic granularity; EVF-SAM [38], a SAM-based model with early vision-language fusion; and Seg-Zero [21], a reasoning-chain guided framework with decoupled reasoning and segmentation modules. Qualitative comparisons with LISA appear in Fig. 6; additional baselines are in the Appendix.

Evaluation Metrics. We report generalized IoU (gIoU) as our primary metric following [12, 37]. Cumulative IoU (cIoU) results are provided in the Appendix.

7.2. Main Results

Results on CONVERSESEG. Tab. 1 reports overall and per-concept gIoU on the SAM-seeded and human-annotated splits of CONVERSESEG. Our Phase-1 model (CONVERSESEG-NET Base, Qwen2.5-VL 3B, without conversational training) already achieves 58.8% on the SAM-seeded split surpassing the strongest LISA variant (55.2% with Llama2-13B) by +3.6%, despite using a $4\times$ smaller backbone and without ReasonSeg fine-tuning. Among existing methods, Seg-Zero is the strongest baseline with 69.2% overall. Our full 3B model (CONVERSESEG-NET) reaches 70.5%, improving over Seg-Zero by +1.3% on the SAM-seeded split. Scaling the prompt encoder to 7B further boosts performance to 73.3%, a +4.1% absolute gain over the best baseline. The same trends hold on the human-annotated split.

Per-Concept Analysis. Tab. 1 reports per-concept performance. Key takeaways: (1) baselines score highest on entities and spatial and lowest on affordances, physics & safety concepts. For example, LISA-Llama2-13B on the SAM-seeded split achieves 60.0% on entities and 46.6% on physics & safety (-13.4%). (2) Our base model shows an even larger

Model	Prompt Encoder	SAM-seeded (gIoU)							Human-annotated (gIoU)						
		All	Ent.	Spat.	Rel.	Aff.	Phys.		All	Ent.	Spat.	Rel.	Aff.	Phys.	
LISA [*]	LLaVA 7B	48.6	52.3	55.0	54.5	41.5	39.1		45.9	46.5	50.4	50.9	42.6	37.7	
LISA [*]	Llama2 13B	55.2	60.0	57.1	60.3	50.1	46.6		53.8	54.3	60.1	56.8	53.6	42.1	
UniLSeg-20	CLIP ViT-B/16	32.6	39.0	36.0	36.5	26.4	22.8		31.6	36.6	35.4	27.4	27.6	29.4	
EVF-SAM [†]	BEIT-3-Large	42.2	47.1	49.3	46.0	37.3	29.8		39.1	42.1	50.1	41.3	29.5	31.2	
EVF-SAM [‡]	BEIT-3-Large	47.7	53.8	54.7	50.9	40.9	36.6		45.4	51.9	52.2	44.2	41.4	34.1	
Seg-Zero	Qwen2.5-VL 7B	69.2	74.1	71.7	72.3	65.1	60.9		61.1	62.8	63.6	61.1	60.2	56.6	
CONVERSESEG-NET (Base)	Qwen2.5-VL 3B	58.8	64.8	64.4	65.0	52.9	43.8		56.4	61.4	61.1	54.8	54.3	48.0	
CONVERSESEG-NET	Qwen2.5-VL 3B	70.5	73.9	74.7	76.3	65.6	60.7		64.4	67.0	64.7	66.5	62.9	59.5	
CONVERSESEG-NET	Qwen2.5-VL 7B	73.3	75.8	75.7	78.6	70.0	65.1		66.3	68.6	68.8	65.3	62.0	66.5	

Table 1. **CONVERSESEG benchmark results (gIoU, %)**. Each subset reports performance across the five concept categories – Entities, Spatial, Relations, Affordances, and Physics & Safety – and summarizes across all (*All*). * indicates models fine-tuned on ReasonSeg training data. [†] trained on RefCOCO only; [‡] trained on RefCOCO and additional large-scale datasets (Objects365, PACO-LVIS, PASCAL-Part, etc).

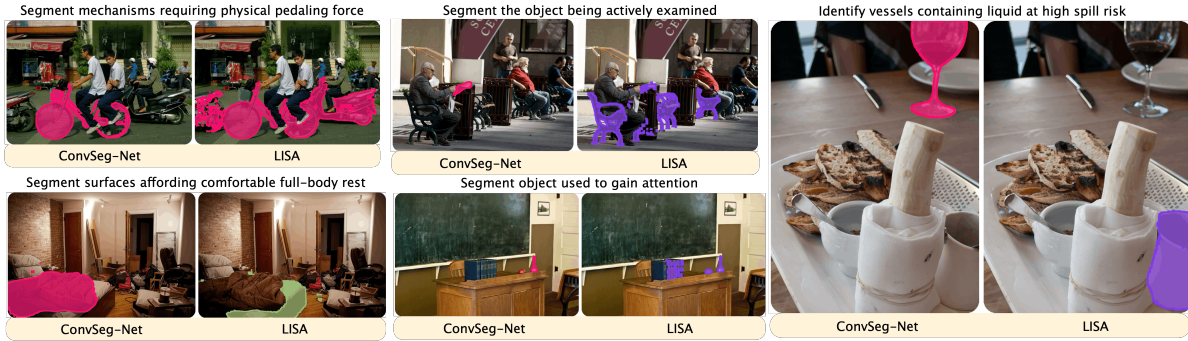


Figure 6. Qualitative results on CONVERSESEG. CONVERSESEG-NET produces more accurate masks for spatial, affordance, and physics concepts than LISA, a popular leading reasoning segmentation model.

gap: 64.8% vs. 43.8% (−21.0%). (3) Phase-2 conversational training boosts performance for all concepts, with biggest gains for physics & safety (from 43.8% to 60.7%) narrowing the gap with entities to 13.2% (73.9% vs. 60.7%). (4) Scaling the backbone (CONVERSESEG-NET-7B) achieves the best results across all families. Overall, later-stage training and larger models especially benefit abstract concepts while preserving strong entity-level performance.

Qualitative Examples. Fig. 6 compares CONVERSESEG-NET and LISA predictions on CONVERSESEG. CONVERSESEG-NET better localizes abstract concepts: it focuses on the bicycle as the mechanism requiring physical pedaling force, isolates the newspaper as the object being actively examined, selects only vessels containing liquid at high spill risk, and segments the bell as the object used to gain attention, whereas LISA often includes nearby distractors or misses physical cues. For “surfaces affording comfortable full-body rest,” however, LISA’s decision to segment the full bed is arguably more canonical than CONVERSESEG-NET’s blanket-focused mask.

Comparison on Referring Expression Benchmarks. Tab. 2 reports gIoU on RefCOCO+/g [37] and ReasonSeg [15]. While the primary goal of CONVERSESEG-NET is to establish a strong baseline for conversational image segmentation with abstract concepts, we evaluate on standard re-

ferring expression benchmarks to demonstrate robustness on literal concepts. On RefCOCO, CONVERSESEG-NET achieves 79.9% on val, competitive with models like GSVA (79.2%) and EVF-SAM (82.4%) that use substantially more training data. On ReasonSeg, our 3B model reaches 55.1% on the test set without training on any ReasonSeg data, a few points behind LISA-13B (61.3%), which was fine-tuned on its training set and is 4× larger. Scaling to 7B improves performance (58.7% on ReasonSeg test), showing that our conversational training effectively transfers to complex reasoning scenarios zero-shot (*i.e.*, without task-specific supervision).

7.3. Ablation Studies and Analysis

Curriculum Learning. Tab. 3 analyzes our training curriculum on RefCOCO+/g (avg. over 9 splits) and CONVERSESEG’s human-annotated split. Training only on conversational data yields reasonable CONVERSESEG performance (63.0%) but poor RefCOCO+/g results (68.0%), indicating overfitting. Jointly training on all data without a curriculum improves RefCOCO+/g to 75.9% but degrades CONVERSESEG to 61.9%. A two-phase curriculum (Phase 1: basic referring data; Phase 2: conversational data) achieves 74.1% and 64.4%. Our final strategy – Phase 1 on basic data followed by Phase 2 with a 50-50 mix of both data types – achieves the highest performance (76.2% on RefCOCO+/g,

Model	Prompt Encoder	RefCOCO			RefCOCO+			RefCOCog			ReasonSeg			
		val	testA	testB	val	testA	testB	val(U)	test(U)	val(G)	val	test	test(short)	test(long)
LISA	LLaVA 7B	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	—	44.4	36.8	37.6	36.6
LISA*	LLaVA 7B	—	—	—	—	—	—	—	—	—	52.9	47.3	40.6	49.4
LISA*	Llama2 13B	—	—	—	—	—	—	—	—	—	60.0	51.5	43.9	54.0
Seg-Zero	Qwen2.5-VL 3B	—	—	—	—	—	—	—	—	—	58.2	56.1	—	—
Seg-Zero	Qwen2.5-VL 7B	—	—	—	—	—	—	—	—	—	62.6	57.5	—	—
GSVA*	Vicuna 13B	79.2	81.7	77.1	70.3	73.8	63.6	75.7	77.0	—	—	—	—	—
GLaMM	Vicuna 7B	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	—	47.4	—	—	—
UniLSeg-20	CLIP ViT-B/16	80.5	81.8	78.4	72.7	77.0	67.0	78.4	79.5	—	—	—	—	—
EVF-SAM [†]	BEiT-3-Large	82.1	83.7	80.0	75.2	78.3	70.1	76.8	77.4	—	—	—	—	—
EVF-SAM [‡]	BEiT-3-Large	82.4	84.2	80.2	76.5	80.0	71.9	78.2	78.3	—	—	—	—	—
HyperSeg	Phi2 2.7B	84.8	85.7	83.4	79.0	83.5	75.2	79.4	78.9	—	—	—	—	—
HyperSeg	Phi2 3B	—	—	—	—	—	—	—	—	—	59.2	—	—	—
Gemini Seg	Gemini2.5 Flash	—	—	—	—	—	—	—	—	—	28.3	30.6	16.5	35.0
X-SAM*	Phi3 3.8B	85.1	87.1	83.4	78.0	81.0	74.4	83.8	83.9	—	56.6	57.8	47.7	56.0
RSVP	LLaVA1.6 7B	—	—	—	—	—	—	—	—	—	59.2	56.9	47.9	58.4
RSVP	Qwen2-VL 7B	—	—	—	—	—	—	—	—	—	58.6	56.1	48.5	57.1
CONVERSEG-NET (Base)	Qwen2.5-VL 3B	79.2	81.3	76.7	73.8	78.4	68.5	74.0	74.4	74.5	52.2	49.9	46.6	50.9
CONVERSEG-NET	Qwen2.5-VL 3B	79.9	81.3	76.8	74.0	79.0	69.5	74.7	74.9	75.3	59.5	55.1	50.4	56.6
CONVERSEG-NET	Qwen2.5-VL 7B	79.8	81.8	77.5	75.0	79.7	70.5	75.4	76.0	76.0	59.8	58.7	52.2	60.8

Table 2. **Referring expression segmentation (gIoU, %)**. CONVERSEG-NET is competitive on RefCOCO+/g and achieves zero-shot state-of-the-art on ReasonSeg with the 7B model, surpassing methods fine-tuned on ReasonSeg (*). [†] trained on RefCOCO only; [‡] on RefCOCO plus additional datasets (Objects365, PACO-LVIS, PASCAL-Part, etc).

Training Strategy	RefCOCO+/g	CONVERSEG
No curriculum (only conversational)	68.0	63.0
No curriculum (all data)	75.9	61.9
Phase 1 + Phase 2 (only conversational)	74.1	64.4
Phase 1	75.6	56.4
Phase 1 + Phase 2 (full curriculum)	76.2	64.4

Table 3. Curriculum learning ablation on RefCOCO+/g and CONVERSEG.

64.4% on CONVERSEG), showing curriculum learning with mixed fine-tuning balances performance across benchmarks.

Architectural Ablations. Tab. 4 ablates key architectural choices. Freezing the prompt encoder backbone instead of using LoRA fine-tuning causes a drastic performance drop (49.4% vs. 64.4%) – adapting the prompt encoder is crucial for language grounding. Providing only text input to the prompt encoder, rather than both image and text, reduces performance by 17.0% (64.4% to 47.4%), showing that visual context is essential for text-conditioned segmentation. Finally, removing dense prompt embeddings while retaining only sparse embeddings causes a modest 0.5% drop.

VLM Backbone Comparison. Tab. 5 evaluates Perception-LM-3B [3], a recent VLM with strong visual reasoning capabilities. Performance remains comparable (65.2 vs. 64.4), demonstrating that CONVERSEG-NET can work with any performant VLM backbone.

Attention Map Visualization. Fig. 7 shows cross-attention between text tokens and image regions in the mask decoder. For each image we visualize maps for two prompts; attention concentrates on the referred region (e.g. “Vitamin A” on carrots) and is sparse and point-like rather than diffuse. We hypothesize this stems from conditioning the SAM mask decoder with language embeddings in place of its point

Architecture Configuration	CONVERSEG	Δ
w/o LoRA finetuning	49.4	-15.0
Text-only input to Qwen	47.4	-17.0
Sparse embeddings only	63.9	-0.5
CONVERSEG-NET	64.4	—

Table 4. Architectural ablations. Each row removes one component from the final design.

embeddings, so each token behaves like a soft point prompt.

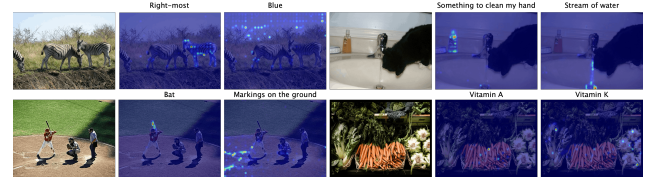


Figure 7. Cross-attention maps showing that each text prompt focuses on its corresponding image region.

8. Conclusion

We introduced Conversational Image Segmentation (CIS), grounding high-level concepts about affordances, physics, and function into pixel-accurate masks. Our CONVERSEG benchmark provides 1,687 human-verified samples with balanced coverage across five concept families underrepresented in prior work. To scale beyond manual annotation, we built an automated data engine synthesizing 61K prompt-mask pairs via iterative VLM generation and verification. CONVERSEG-NET, trained on this data, achieves state-of-the-art performance on CONVERSEG while remaining competitive on standard benchmarks, showing that a curriculum from literal to conversational concepts effectively adapts promptable segmentation to language conditioning.

Acknowledgments

We thank Damiano Marsili and Ilona Demler for their valuable feedback. Aadarsh is supported by the Kortschak Scholarship and Caltech’s CAST program. Georgia is supported by the Powell Foundation, Meta through the LLM evaluation research grant, Google, and Amazon.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5
- [2] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110(45):18327–18332, 2013. 2, 3
- [3] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. 8
- [4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5
- [5] Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. Coconut: Modernizing coco segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21863–21873, 2024. 6
- [6] Xueqing Deng, Qihang Yu, Ali Athar, Chenglin Yang, Linjie Yang, Xiaojie Jin, Xiaohui Shen, and Liang-Chieh Chen. Coconut-pancap: Joint panoptic segmentation and grounded captions for fine-grained understanding and generation. *arXiv preprint arXiv:2502.02589*, 2025. 2
- [7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [8] JJ Gibson. The ecological approach to visual perception: classic edition, 2014. 2, 3
- [9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5
- [10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5
- [11] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European conference on computer vision*, pages 108–124. Springer, 2016. 2
- [12] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6
- [13] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18145–18154, 2022. 2
- [14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2
- [15] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2, 6, 7
- [16] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022. 2
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4
- [18] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1271–1280, 2017. 2
- [19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2
- [20] Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3459–3469, 2024. 6
- [21] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 6
- [22] Yang Liu, Muzhi Zhu, Hao Chen, Xinlong Wang, Bo Feng, Hao Wang, Shiyu Li, Raviteja Vemulapalli, and Chunhua Shen. Segment anything in context with vision foundation models. *International Journal of Computer Vision*, 133(10): 7460–7485, 2025. 2
- [23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [24] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 2, 3

- [25] Moondream. moondream3, 2025. 4
- [26] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2
- [27] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2
- [28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 4, 5
- [29] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2
- [30] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 2
- [31] Mert Bülent Saryıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8011–8021, 2023. 2
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
- [33] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 2
- [34] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 5
- [35] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022. 2
- [36] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2
- [37] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016. 1, 2, 6, 7
- [38] Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. 6
- [39] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14227–14238, 2024. 2