POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Enhancing SAM3 Robustness Against Semantic and Domain Shift via Dual-Prompting

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Michele Cavicchioli

Advisor: Prof. Matteo Matteucci

Co-advisor: Nico Catalano, Stefano Samele

Academic year: 2025-2026

---

1. Introduction and Motivation

Object segmentation is a foundational computer-vision task, underpinning medical imaging, autonomous driving, content creation, and interactive image editing. A long-standing usability goal is to let a user say *what* to segment with as little effort as possible, ideally through a short noun phrase (NP) such as “truck” or “mouse” rather than a lengthy description. That convenience has a cost. A single noun phrase is intrinsically ambiguous, and a segmentor asked to act on it has to guess what the user means. This thesis asks whether *adapting a segmentor to the few-shot setting*, by supplying a handful of labeled support examples, can resolve that ambiguity and make the model more robust, all under a strictly training-free constraint. The main question organizing the work is about if a few labeled examples can recover the concept a bare word leaves underspecified and ambiguous.

2. Related Works

SAM3: SAM3 [3] is the latest model in Meta’s "Segment Anything" series, following SAM1 [7] and SAM2 [13], which are both class-agnostic models designed to segment "anything." While SAM1 handled only static images, SAM2 extended these segmentation capabilities to videos. These models became highly popular and widely used across various segmentation tasks because they are completely class-agnostic; a user only needs a point prompt to segment an object or track it across an entire video. SAM3 extends these capabilities from single instances to broader concepts, evolving Promptable Visual Segmentation (PVS) into what the authors call Promptable Concept Segmentation (PCS). PCS can be defined as the task of segmenting all instances in an image or video that correspond to a prompted concept. A user can define a concept for SAM3 using a text label alongside geometric prompts, which can be either points or bounding boxes. The model utilizes Perception Encoders (PE) [2] as backbone feature extraction networks, using a language PE for text and a visual PE for images. Geometric prompts are handled by an additional exemplar encoder that processes both the input coordinates of the prompts and the image features. The main contribution of SAM3 is its detector component, which is responsible for finding all instances of the prompted concept within a sin-

gle image or within each individual frame of a video. The detector features an encoder component called the fusion encoder, consisting of a sequence of six self- and cross-attention layers where the image features are conditioned upon the prompt embeddings (text and geometric). The conditioned image features are then passed to the detector decoder, which is a DETR-style transformer [4]. Here, a set of Q learnable vectors, called object queries, are conditioned on the prompt embeddings and the conditioned image features through six layers of self-attention, cross-attention on prompts, and cross-attention on the conditioned image features. These conditioned object queries are then passed through the segmentation head to produce the final predictions. Another major addition is the presence token. The authors of SAM3 state that requiring the object queries to both classify an object locally and determine if that object aligns with the requested concept forces the network to process information at two different levels. Doing this within a single component degrades performance. Therefore, they introduce an additional object query—the presence token—whose sole purpose is to compute the probability that the requested concept is present inside the image. If the concept is not present, all segmentations produced by the object queries are set to none. Thanks to the massive pre-training of its encoders and the high-quality masks used to train the model, SAM3 achieves excellent results in zero-shot scenarios.

Few-Shot learning: Few-Shot Learning (FSL) is a machine learning paradigm designed to enable models to generalize to novel classes sampled from an unknown distribution, using only a limited number of annotated examples. Departing from traditional supervised learning, FSL leverages meta-learning principles—specifically "learning to learn" [17]—to bridge the gap between training categories and unseen inference categories. In practice, this is achieved via episodic training, which structures both optimization and evaluation into "episodes" that mimic inference conditions. Each episode typically follows an N-way K-shot protocol, where the model must distinguish between N novel classes given exactly K annotated examples per class. **Few-Shot Segmentation**: FSS [16] extends this paradigm to pixel-level dense prediction. In an FSS framework, the model is tasked with generating a segmentation mask for a query image based on a support set consisting of K image-mask pairs of a novel class. Formally, given a novel class l and a support set $S(l) = \{(I^i, M_l^i)\}_{i=1}^k$, the FSS model f_θ learns the novel class representation from these few annotated examples and maps it to the query features to predict the corresponding mask:

$$\hat{M}_q = f_\theta\big(I_q, S(l)\big)$$

The most prominent evaluation protocols include PASCAL-5^i and COCO-20^i, both of which partition their respective base datasets (PASCAL VOC and MS-COCO) into four disjoint folds for cross-validation. Additionally, the long-tailed LVIS-92^i dataset is utilized to test generalization against genuinely novel categories. Performance across these datasets is universally quantified using the mean Intersection over Union (mIoU) metric. The mIoU is computed by first aggregating intersection and union pixel counts cumulatively across all query images on a per-class basis, then averaging the resulting per-class IoU scores:

$$\text{mIoU} = \frac{1}{|L|} \sum_{l \in L} \frac{\sum_{j=1}^{N_l} |\hat{M}_l^j \cap M_l^j|}{\sum_{j=1}^{N_l} |\hat{M}_l^j \cup M_l^j|}$$

where L is the set of novel classes evaluated and N_l is the number of query images for class l. Foundation models have increasingly shaped this landscape, giving rise to two adaptation families differentiated by whether they require task-specific training. **Foundation Models in Few-Shot Segmentation**: Large-scale visual foundation models like DINOv2 [**?**], CLIP [12], and the SAM series have shifted FSS away from purely episodic backbones toward leveraging general-purpose representations. Methods adapting SAM to FSS split into two paradigms. *Prompt embedding* methods attach a trained module to the frozen backbone: VRP-SAM [18] trains a Visual Reference Prompt Encoder that maps support image-mask pairs to prompt embeddings fed to SAM's mask decoder; VLP-SAM [14] extends this with CLIP text features via CLIP Surgery [8] to include semantic information from a text prompt. Adapting SAM2, FSSAM [19] inserts three modules to redirect memory and attention from instance tracking

toward category semantics, while SANSA [6] recovers category structure already latent in SAM2's video pretraining with minimal architectural changes. *Training-free* methods require no learned components and derive geometric prompts directly from support-query feature correspondence. PerSAM [21] builds a location confidence map from support-query feature similarity, selects one positive and one negative point, and biases SAM's cross-attention toward the target region. GF-SAM [20] addresses SAM's granularity ambiguity by organizing candidate mask proposals into a directed containment graph and pruning by confidence and spatial extent. Matcher [9] pairs DINOv2 dense matching with SAM: bidirectional bipartite matching identifies query-side foreground patches, a k-means robust sampler distills them into compact positive and negative point prompts, and Earth Mover's Distance filters false-positive mask proposals. MARS [5] uses multi-modal cues to the FSS task, this is a re-ranking framework that leverages support images to re-rank mask proposal from a beneath mask generative model like Matcher or GF-SAM, to produce a better final mask; the ranking of each mask is based upon 4 scores computed from visual and semantical priors, with the last computed using a predicted text label from the support set via a VLM.

3. Problem Statement

Noun-phrase prompting is undermined by two kinds of ambiguity. *Visual ambiguity*: one word denotes visually dissimilar objects ("truck" is a pickup with a short bed or a semi with a long trailer). *Semantic ambiguity*: one word denotes unrelated concepts that share no appearance ("mouse" is a rodent or a computer peripheral). Given only the word, a model cannot recover which one the user intends. A second, subtler problem is *domain shift*. Annotation is subjective: different datasets segment the "same" concept under slightly different operational meanings, so the knowledge a model has internalized about a concept may not match the convention a given benchmark was labeled under. The same word can then score very differently from one dataset to another, a gap that has nothing to do with image difficulty and everything to do with *concept alignment*.

4. Quantifying Domain Shift

To do our testing we use the zero-shot foundation model SAM3 [3], which, given a noun phrase, segments every instance of the named concept in an image without any task-specific training. To test whether domain shift is real rather than a false hypothesis, we need many concepts that recur across datasets with potentially different labelings. We assemble six datasets spanning different tasks: COCO-2017, LVIS, PASCAL-2012, ADE20K (object and stuff classes only, with parts filtered out to avoid intra-dataset overlap), and the video benchmarks YouTube-FSVOS and MiniVSPW (one non-empty random frame per video), for a total of 2,124 labels.

4.1. Unifying labels via a WordNet taxonomy

Deciding when two labels denote "the same concept" cannot be a string match: synonymy alone defeats that. We therefore map every label onto a shared taxonomy. SAM3's own SA-Co/Wikidata ontology is neither public nor practically reconstructible, and is polluted by abstract nodes; instead we adopt *Open English WordNet* (OEWN 2025+) [10], a curated extension of WordNet 3.1 [11]. The label→synset mapping is a five-phase, mostly automatic protocol that minimizes human linguistic judgment. It begins with *candidate generation*, converting legacy WordNet IDs through sense-keys and then applying literal/lemmatized matching, with a spaCy root-noun fallback when nothing matches (e.g. *traffic light* → *light*). The candidates pass to an *LLM triage* stage, where a large language model acts as a strict gatekeeper enforcing a visual-counterpart rule (informational labels map to their physical medium), granularity preservation, and synonym handling, emitting one of EXACT_MATCH, NEEDS_SYNONYM, or NEEDS_CUSTOM_NODE. A human-in-the-loop routing step then auto-commits exact matches while verifying synonyms and custom leaf nodes (e.g. `ade-n-00001`) and attaching them to their deepest valid OEWN parent. A *validation* pass requires every path to root to terminate at *physical_entity*, flagging synsets that resolve to *abstraction* for manual audit and rejecting broken IDs for reprocessing. Finally, *lemma assignment* has an LLM pick, copy-paste

Table 1: Zero-shot SAM3 mIoU across the six datasets under the three text modalities. *Best* and *worst lemma* are theoretical bounds computed per class from the ALL_LEMMAS sweep. The large best→worst spread on every dataset quantifies SAM3's sensitivity to synonymous phrasing.

Zero-shot text label	COCO	LVIS	PASCAL	ADE20K Gr.	YT-FSVOS	MiniVSPW
Ground-truth label	73.69	53.19	82.75	-	**82.52**	64.05
Selected lemma	73.80	55.54	82.82	33.48	82.34	70.01
Best lemma	**73.88**	**58.12**	**82.94**	**35.14**	82.35	**71.41**
Worst lemma	56.59	40.63	56.59	21.39	57.31	35.34

only, the single most visually specific lemma per synset (e.g. *flowerpot* over *pot*, *trash can* over *ashcan*). After building the mapping for the ADE20K dataset we saw that multiple labels were overlapping inside the same dataset (due to synonyms/plurals), we therefore grouped all labels overlapping on the same synset as a single class, and with a LLM we semantically validated the groupes. This cleaned version we call *ADE20K Grouped* is the one we use during our analysis, instead of the original version

4.2. Three prompting modalities and the deviation analysis

With every label grounded in OEWN, we prompt SAM3 zero-shot under three text label levels: GT_LABEL (the raw dataset label), SYNSET_LABEL (the selected lemma), and ALL_LEMMAS (one inference per lemma of the synset, e.g. *person, individual, someone, mortal, soul*). From ALL_LEMMAS we read off, per class, the theoretical *best-* and *worst-lemma* scores. Table 1 reports the picture: the selected lemma already matches or beats the raw label, the best lemma adds several points (LVIS $53.19 \rightarrow 58.12$, MiniVSPW $64.05 \rightarrow 71.41$), and the *worst* lemma is catastrophic (LVIS 40.63, MiniVSPW 35.34). Synonymy is thus a first-order failure mode, not a rounding error.

The cross-dataset gap is then visualized through three deviation histograms (Figure 1), each counting how many samples fall in a given score-deviation bin. The *synset-level* plot takes every synset shared by $\geq$ 2 datasets and records its max−min score (one sample per synset). The *class-level* plot, for the same shared synsets, records the deviation between the best dataset and each other dataset (one sample per extra dataset). The *lemma-level* plot uses ALL_LEMMAS, recording the deviation between every unordered pair of lemmas of a synset, regardless of dataset. The result is unambiguous. Were domain shift absent, the synset- and class-level histograms would collapse to a spike at the origin; instead both carry heavy mass at large deviations, so the same concept scores very differently across datasets, confirming the hypothesis. The lemma-level histogram is the most striking: after discarding the head and tail spikes it is essentially *uniform*, meaning the odds of drawing a well-performing lemma are about the same as drawing a damaging one. That is a direct visual statement of how fragile noun-phrase prompting is to the exact word chosen.

5. Method: a Double Prompting Strategy

We adapt SAM3 to few-shot segmentation chronologically: try a solution, diagnose why it fails, refine. Evaluation uses the canonical FSS benchmarks COCO-20^i, LVIS-92^i, and PASCAL-5^i, plus YouTube-FSVOS and MiniVSPW. ADE20K is excluded because most of its classes hold a single image, leaving no valid (support, query) episode. In addition to this, since the dataset pool changed we had to re-run again the same domain shift analysis on the zero-shot experiment with the new updated pool. As a consequence of this the number of samples decreased by a lot (mainly due to the drop of ADE20K that was overlapping quite with LVIS); the domain shift phenomenon is still present but with a different magnitude.

5.1. Adapting SAM3 to Few-Shot

SAM3 *does* accept exemplars, but its API assumes they originate from the *query* image. The naive adaptation encodes N points from the sup-

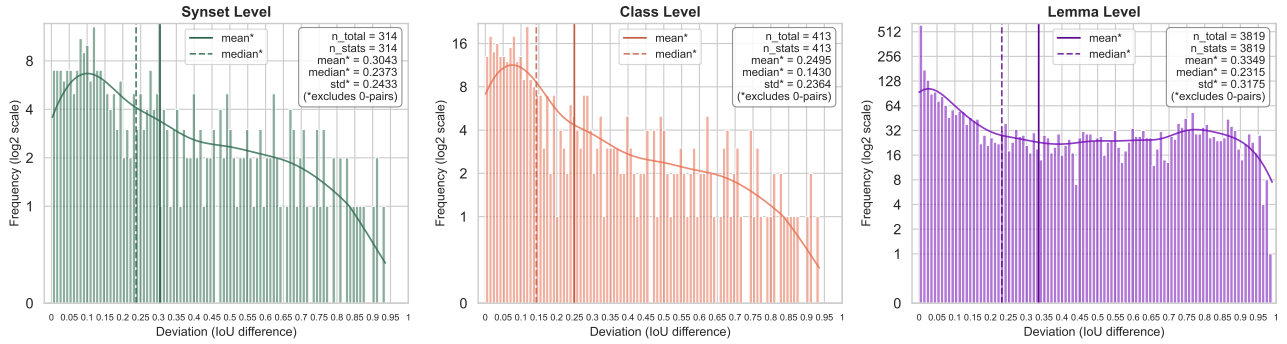

Figure 1: Zero-shot deviation analysis. The synset- and class-level histograms carry heavy mass away from zero, which is the signature of domain shift, while the near-uniform lemma-level distribution exposes the sensitivity to synonymous phrasing.

port mask and injects the resulting vectors into the fusion encoder at query inference. Each point embedding mixes two *spatial* components (a linear projection of the normalized coordinate and a positional encoding) with a *visual* one (2×2 grid-sample for points). The spatial parts are tied to coordinates *in the support image*, so injecting them tells the model "the object is at (x, y)" using a location the query never contained. Stripping them gives +1.9 mIoU on average, but the variant still only matches, and never clearly surpasses, the zero-shot baseline: support and query features come from two independent backbone passes and, because of their differing backgrounds, never fully align in embedding space. In general, this experiment is struggling against the model's training assumptions, of exemplars coming from the same query image (exemplars embeddings almost identical to query image patch vectors), instead from an external source.

This first experiment is run with different ablations adding to the standard configuration. The first is to strip the support spatial components, a setting we call NO_POS, which introduces an improvement of +1.93 when compared to the same experiment including such components. The second is the number of sampled points: N=5 consistently beats N=20, because the model was already finding confusing very little external information, giving it even more just makes things much worse.

5.2. Query points experiment

To check whether points themselves are useful, we sample N points directly from the *query* ground-truth mask (a validation for point quality) and prompt them with the text label. This lifts mIoU well above zero-shot on every dataset, reaching on LVIS-92^i a score of 74.28 with NO_POS, which actually is better also in the case of points directly on the query image. Removing the text label collapses performance but that's expected, even though these scores are crushing those of SAM3 in few-shot without the text label, resulting in the fact that what works is the combination of text and points together. The result fixes the two-step recipe, *first localize good points on the query, then run inference with text + points*, and gives a ceiling (Table 2, last row). It also motivates the *point overlap* metric, the fraction of predicted points that land inside the query foreground, which helps us understand if a solution globally finds better or worse points.

5.3. Patch-Level matching

A first attempt to predict query points from the support adopts Matcher's [9] patch-level matching (features come from the image encoder of SAM3) which was a complete failure due to the matching procedure producing incorrect points. Trying to semantically group features to highlight the foreground features better we switched from native to text-*fused* feature volumes conditioned using the fusion encoder of SAM3, and a sweep of sampling strategies (random, top-k, K-means, K-medoids, coreset). **All variants fail**: none even reaches the support-injection baseline, and point overlap stays low, especially on LVIS. The lesson is that pure *visual* similarity is the bottleneck, since it routes foreground support patches onto background query patches, so the

matching has to be made *semantic*.

5.4. Attention-based sampling

Crucially, SAM3 already encodes semantics internally, in the fusion encoder's attention. We therefore sample the prompt points *from the fusion encoder's own attention maps* rather than from external feature matching (Figure 5). The weight matrix has image patches as queries and the prompt tokens (text + exemplars) as keys; aggregating over the key columns and reshaping onto the 72×72 grid yields a per-patch saliency map.

Sampling top-k from the *cross*-attention map fails: can see in Fig. **??** the point sub-map is noisy and top-k repeatedly selects high-activation *background* patches, with the reason why there's noise in these maps being the same why the naive adaptation of SAM3 to few-shot does not work; support exemplars are from an external image, its vectors are never going to be fully aligned with those in the query unless processed by an ad-hoc trained component, resulting in low and sparse activations in the cross-points sub-map. The *self*-attention map, by contrast, localizes the concept cleanly, but inverted, so the target patches carry the *lowest* aggregate weight. Cross-attention pulls foreground patches toward the prompt value embeddings; in the following self-attention, background patches (numerous and mutually similar) accumulate a high *sum* of moderate weights, while foreground patches score a high *max* but a low sum. The right recipe is therefore to sample from the *self*-attention map, aggregated by *sum* over keys, *averaged across all six layers*, taking the *bottom-k* (N=5) patches; this can be seen with the same attention weights aggregated with sum and max individually. From the self-attention sampling experiments, two configurations stand out. The first, **Text + Box**, conditions the sampling pass on the text label and a support-derived bounding box (NO_POS); it is the first variant to beat *both* baselines and is the most *robust* at the lemma level, because the support box helps exactly when a lemma carries little visual grounding. The second, **Text only**, computes the points from the text label alone with *no* support image, and is the strongest configuration in absolute mIoU while itself remaining zero-shot, since it adds no external information. This reasoning is showed via Fig. 2 and Fig. 3 which illustrate the same episode processed with both the text label and the support point embeddings. In addition to this, Fig. 4 shows the same but in the top performing experiment with self-attention sampling via only the text label.

Figure 2: Fusion encoder attention maps during the sampling pass with both text and support point embeddings, key-aggregated via SUM, averaged between all layers

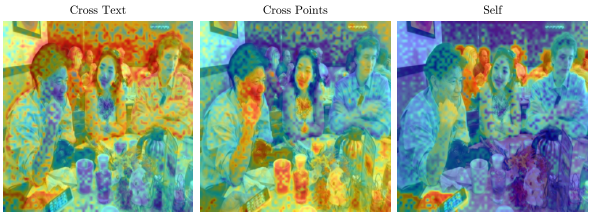

Figure 3: Fusion encoder attention maps during the sampling pass with both text and support point embeddings, key-aggregated via MAX, averaged between all layers

Why does adding consistent points help at all? Consistent prompt embeddings with the text tokens provide a more clear concept definition to the fusion encoder, with cross-attention layers focusing better only on the proper objects, hence the self-attention also works better. Moreover, inside the detector decoder the points reinforce the text concept along the conditioning chain (promt tokens condition the object queries

Figure 4: Fusion encoder attention maps during the sampling pass with only the text label, key-aggregated via sum, averaged between all layers

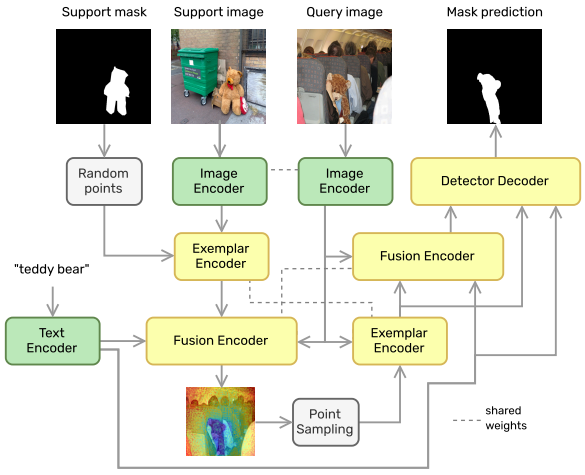

Figure 5: Attention-based sampling. A first "sampling pass" through SAM3's fusion encoder records the self-attention map (sum-aggregated over keys, averaged across all six layers); the bottom-k patches become query points that are re-prompted to SAM3 with the text label. Variants differ in the fusion-encoder input (text only, or text + support box) and the sampling.

first, object queries are conditioned on the conditioned image features), raising the confidence of object queries that, with text alone, fell just below SAM3's 0.5 threshold. SAM3's authors note the dual-prompt caveat: when text and geometry refer to *different* concepts the output is undefined, so prompt consistency is essential, which our self-attention points respect by construction, explaining also the reason why false positive points are a killer in this settings.

6. Results

Table 2 reports the decisive comparison on all five FSS benchmarks. The text-only attention-sampling configuration is the best in absolute mIoU, improving over the zero-shot baseline by +3.74 on average and beating the support-injection baseline (+1.95); the Text + Box variant trails it slightly but is more lemma-robust. The oracle query-mask row marks the ceiling, yet the same points *without* text fall back below the text-only baseline. This is the counter-intuitive core finding: in our pipeline support information harms the performance of SAM3 rather than helping it, as it breaks that consistency between text and exemplars, due to the difference source from where the image embeddings come from, in our case this results in worse points when the pipeline processes support images, and therefore worse output masks. Once the model has the right words, extra visual signal can increase the performance if they are consistent with the text label (showed with the query points experiment), or harm it (naive adaptation of SAM3 to few-shot).

On *domain shift*, both winning configurations pull the cross-dataset deviation toward zero (Figure 7). The synset-level mean and standard deviation drop from the zero-shot 0.148/0.132 to 0.122/0.110 for Text + Box and to 0.116/0.104 for Text only. Recalling that these data and Figure 7 are to be compared with Figure 6 due to the new dataset pool.

Robustness to a bad word choice is where the gain is clearest, and it is worth being precise about what it measures. The *worst-lemma* score takes, for each class, its lowest-scoring synonym, so it reports not the average case but how badly performance can collapse when a user reaches for an unlucky word instead of the canonical one. Both configurations lift this score sharply over zero-shot (COCO-20^i 57.60→62.68, MiniVSPW 35.34 → 48.36), and the lemma-level deviation contracts. Here Text + Box (0.222/0.257) slightly edges Text only (0.254/0.290): the support box does not raise peak accuracy, but it does buy robustness for the synonyms the text encoder grounds poorly. In other words, the gain from visual support is a robustness gain, not an accuracy gain.

Two diagnostics explain the gap that still separates our points from the oracle ceiling. The first is a *point-geometry audit*. Restricted to the selected lemma, attention points miss the target mask on 19.1% of placements on COCO-20^i, 32.7% on PASCAL-5^i, and 30.9% on LVIS-92^i. At the episode level, at least one point lands off-mask in 36.8%, 58.9%, and 50.8% of episodes, respectively. Off-mask placement is the dominant driver of the remaining IoU gap: on COCO-20^i, episodes with ≥ 1 off-mask point show a mean gap of +0.079 compared to +0.033 for fully on-mask episodes, and PASCAL-5^i shows a similar +0.037 vs. +0.017 split. When all points land on-mask, the PASCAL-5^i residual gap reverses (0.849 attention vs. 0.840 GT/QAS), confirming that on-mask attention points select high-quality positions and that the entire PASCAL-

Table 2: mIoU on the five few-shot benchmarks (N=5, 1-shot). Our two best configurations sample bottom-k points from the sum-aggregated, all-layer *self*-attention map. *Text only* is best in absolute terms (and is itself zero-shot); *Text + Box* is the most lemma-robust. *Support-injection* and *Query points* are baselines; the latter shows points alone, without text, fall below text-only.

Experiments [5 points]	COCO-20^i	LVIS-92^i	PASCAL-5^i	YT-FSVOS	MiniVSPW
Zero-shot	74.07	54.50	**81.90**	82.52	64.05
Support points NO_POS + Text	73.06	55.36	81.50	82.38	66.93
Self-attn Text + Box	74.31	58.97	78.36	82.75	72.42
Self-attn Text only	**76.37**	**62.13**	79.16	**83.73**	**74.35**
Query points + Text	79.85	74.28	84.38	86.52	78.99
Query points NO_POS + Text	80.43	76.41	83.93	86.37	79.76
Query points	70.47	60.34	78.32	77.89	67.07

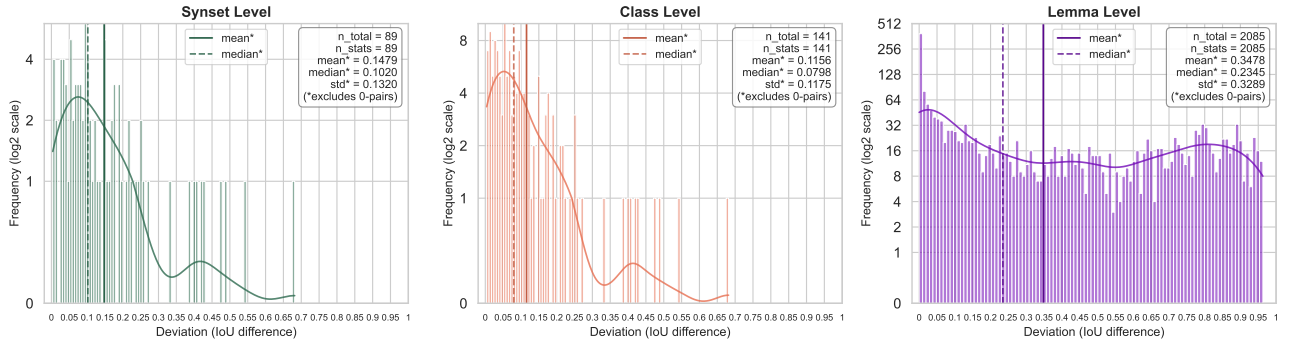

Figure 6: Deviation analysis for the zero-shot baseline on the new dataset pool made of COCO-20^i, LVIS-92^i, PASCAL-5^i, YoutubeFSVOS and MiniVSPW.

5^i gap stems from off-mask localization. Other properties (coverage gap, dispersion, centroid offset) show clear geometric differences but do not correlate with mIoU performance loss.

The second diagnostic is the *FIXED vs. FREE* experiment, which separates point sampling from word-label effects. In the FREE (standard) setting, each synonym samples its own points, so a word that misfires at the attention stage also receives geometrically poor points. In the FIXED setting, all synonyms of a class share the points computed from the selected lemma, isolating the inference-time text penalty. We formalize this using the sampling effect ($\Delta = \text{IoU}_{\text{fixed}} - \text{IoU}_{\text{free}}$) and the inference residual ($\text{resid} = \text{IoU}_{\text{selected}} - \text{IoU}_{\text{fixed}}$). Fixing the points improves the worst-lemma score by +3.32 on COCO-20^i, +2.27 on PASCAL-5^i, and +6.51 on LVIS-92^i, confirming sampling-stage sensitivity. Pooling over all synonym lemmas, the total gap decomposes into 40% sampling / 60% inference on COCO-20^i, 21%/79% on PASCAL-5^i (which is inference-bound by archaic synonyms), and 68%/32% on LVIS-92^i (which is sampling-bound by rare, tiny objects). At the lemma level, some cases are sampling-dominated (e.g., `toilet` → `crapper` on COCO-20^i recovers 77% of the gap with $\Delta = +0.525$), while others are inference-dominated (e.g., `bus` → `coach`, `person` → `soul`), where even perfect query-mask points cannot overcome the text-encoder floor. With ground-truth points, the total synonym gap collapses to 0.02–0.03 and becomes purely inference-driven. Crucially, the selected lemma's point geometry shows no correlation with the synonym inference residual ($|r| < 0.06$), proving the synonym penalty is a genuine linguistic effect rather than a geometry confound.

7. Conclusions and Next steps

SAM3 *can* be adapted to few-shot segmentation through careful, training-free prompting. Our best method, sampling bottom-k points from the sum-aggregated, all-layer self-attention map

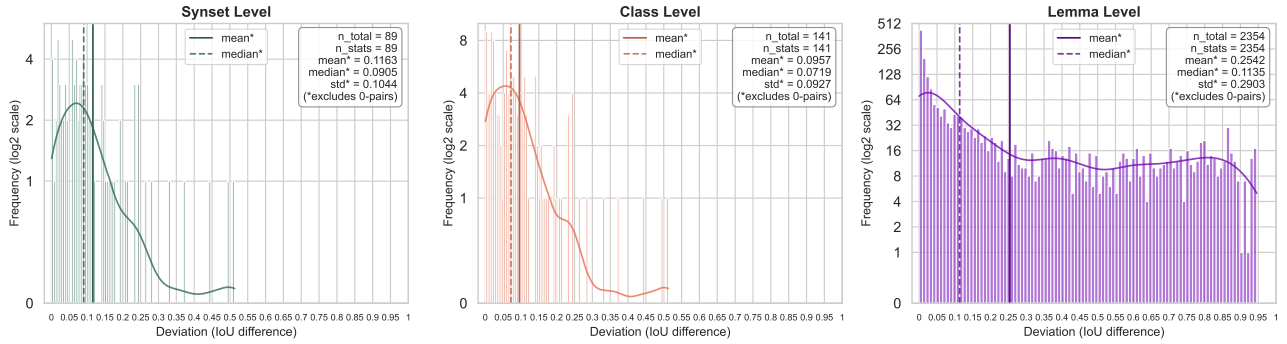

Figure 7: Deviation analysis for the best (Text only) configuration. Compared to the zero-shot baseline, the synset- and class-level histograms concentrate toward zero (reduced domain shift) and the lemma-level tail is cut earlier (improved synonym robustness).

and re-prompting with the text label, improves over the zero-shot baseline by +3.74 mIoU on average and beats the support-injection route. But the gap between SAM3's training assumptions and deployment constraints matters more than expected. The exemplar-from-query assumption is the core obstacle: support and query features, produced by separate forward passes, never align, so injected visual tokens fight the model's internal priors instead of reinforcing them. Tellingly, the single best configuration (*Text only*) uses *no* support image at all; it simply mines SAM3's own attention to find consistent query points. The paradoxical degradation from external visual exemplars does not prove visual grounding futile. It shows that, without a dedicated adaptive component to bridge support and query, the model's priors win, and that visual support helps only at the margins, raising the robustness floor for synonyms the text encoder handles poorly (the Text + Box variant). The broader lesson for foundation-model adaptation is twofold. First, *language precision outweighs externally injected visual information* in the training-free regime. Second, *prompt phrasing is a first-order experimental variable*: the deviation analysis shows a single synonym swap can collapse inference, so the exact lemma deserves the same scrutiny as model weights. Promising next steps include processing the fusion-encoder attention maps with CAM-style refinements (attention-rollout [1], Grad-Attention [15]) to cut the residual off-mask rate, point-versus-box ablations at varied N, full-scale runs on COCO/LVIS/PASCAL/ADE20K with best-lemma prompts, and 5-shot extensions (already supported by the pipeline). Whether the approach scales to larger benchmarks or transfers to other vision tasks remains open.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4190–4197, 2020.

[2] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

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

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[5] Nico Catalano, Stefano Samele, Paolo Pertino, and Matteo Matteucci. Mars: a multimodal alignment and ranking system for few-shot segmentation. In *Proceedings of*

the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1284–1293, 2026.

[6] Claudia Cuttano, Gabriele Trivigno, Giuseppe Averta, and Carlo Masone. Sansa: Unleashing the hidden semantics in sam2 for few-shot segmentation, 2025.

[7] Alexander Kirillov et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[8] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

[9] Yang Liu et al. Matcher: Segment anything with one shot using all-purpose feature matching. *International Conference on Learning Representations (ICLR)*, 2024.

[10] John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. English WordNet 2019 – an open-source WordNet for English. In Piek Vossen and Christiane Fellbaum, editors, *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland, July 2019. Global Wordnet Association.

[11] George A. Miller. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.

[12] Alec Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[13] Nikhila Ravi et al. SAM2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[14] Kosuke Sakurai, Ryotaro Shimizu, and Masayuki Goto. Vision and language reference for a segment anything model for few-shot segmentation. *Journal of Imaging*, 12(4):143, 2026.

[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128(2):336–359, 2020.

[16] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[17] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[18] Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. Vrp-sam: Sam with visual reference prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23565–23574, 2024.

[19] Qianxiong Xu, Lanyun Zhu, Xuanyi Liu, Guosheng Lin, Cheng Long, Ziyue Li, and Rui Zhao. Unlocking the power of SAM 2 for few-shot segmentation. In *Forty-second International Conference on Machine Learning*, 2025.

[20] Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi H Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. *Advances in Neural Information Processing Systems*, 37:33232–33261, 2024.

[21] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Gao Peng, and Hongsheng Li. Personalize segment anything model with one shot. In *International Conference on Learning Representations*, volume 2024, pages 18250–18279, 2024.