PolicySeg: Policy-Guided Semantic Redirection in MLLMs for Grounded Segmentation

Anonymous ACL submission

Abstract

Grounded segmentation MLLMs enable open-vocabulary mask generation through natural-language queries, but also allow unauthorized extraction of sensitive visual entities. Existing defenses are not well suited to this setting: post-processing operates outside model generation, while refusal-based alignment exposes protection through visible refusals or empty outputs. We propose PolicySeg, a policy-driven privacy framework that internalizes **implicit protection** within grounded segmentation MLLMs. Given a configurable authorization policy, PolicySeg injects a learned token, <PRIV>, for unauthorized queries. This token modifies the semantic representation of <SEG>, enabling the model to generate plausible non-empty **substitute masks** rather than revealing protected targets, while preserving the standard segmentation interface and supporting policy updates without retraining. Experiments on multiple image referring segmentation benchmarks using Qwen3-VL-based and InternVL3-based Sa2VA backbones show that PolicySeg substantially reduces protected-target leakage while preserving authorized-query utility. Representation analysis further reveals an authorization-conditioned shift in the <SEG> token space, supporting internal semantic redirection as a mechanism for implicit privacy protection.

1 Introduction

Grounded segmentation has recently emerged as an important capability of multimodal large language models (MLLMs) (Bai et al., 2025; Zhu et al., 2025), enabling open-vocabulary visual segmentation through natural language instructions (Lai et al., 2024; Yuan et al., 2025a). Unlike traditional segmentation systems with predefined semantic categories, grounded segmentation MLLMs can flexibly localize arbitrary user-specified entities, making them promising for applications such as robotics, autonomous driving, and visual assistants.

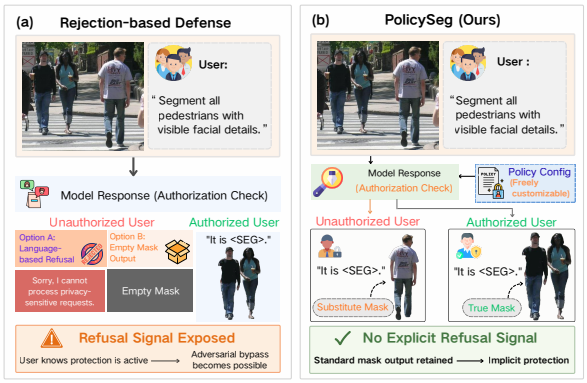

Figure 1: From explicit rejection to implicit protection. Rejection-based defenses visibly expose protection through refusals or empty masks. PolicySeg instead uses configurable policies to return substitute masks for unauthorized requests while preserving the standard mask-producing interface.

Recent privacy-preserving visual retrieval studies, including work on person re-identification, have shown that fine-grained localization and retrieval of visual identities constitute sensitive capabilities requiring protection (Meng et al., 2026; Wang et al., 2024; Cui et al., 2024). This concern is particularly relevant to grounded segmentation models, which may be exploited to extract sensitive regions such as faces, pedestrians, license plates, or medical regions through natural-language queries. Existing privacy-oriented methods mainly rely on post-processing operations after segmentation generation, including masking, blurring, or anonymization (Murrugarra-LLerena et al., 2025; Tseng et al., 2025; Chen et al., 2026; Hukkelås et al., 2019; Hukkelås and Lindseth, 2023; Zwick et al., 2024; Liu et al., 2024; Aslam and Nasrollahi, 2025). Because these approaches operate outside the generation process, they cannot prevent leakage from intermediate outputs or internal representations. Another line of defense relies on refusal-based alignment and safety guardrails (Ghosal et al., 2025; Bai et al., 2022; Oh et al., 2024). However, explicit

refusal is problematic for grounded segmentation, where users naturally expect mask outputs. Visible refusals or empty masks expose the existence of protection mechanisms and facilitate prompt-bypass attacks. As illustrated in Figure 1(a), such explicit defenses reveal observable protection signals, motivating a privacy mechanism that can preserve the normal segmentation interface without directly exposing defensive behavior.

We argue that an effective privacy defense for grounded segmentation should satisfy three properties: (1) the protection mechanism should be embedded inside the model rather than relying on external post-processing; (2) the defense behavior should remain implicit by generating plausible outputs instead of explicit refusals; (3) authorization policies should support dynamic updates without retraining the model.

Motivated by these observations, we propose PolicySeg, a policy-driven privacy protection framework for grounded segmentation MLLMs. Modern MLLMs exhibit strong instruction understanding and semantic reasoning capabilities, which we leverage to interpret flexible natural-language authorization policies. PolicySeg further introduces a learned token, <PRIV>, to modify the representation of <SEG>. For unauthorized requests, PolicySeg generates plausible substitute masks instead of revealing protected targets, as illustrated in Figure 1(b), thereby preserving the standard mask-producing interface without explicit refusal. The policy can be updated without retraining the segmentation model.

In summary, our contributions are as follows:

- We identify a new privacy threat in grounded segmentation MLLMs, where open-vocabulary segmentation queries can extract sensitive visual entities without authorization.
- We propose PolicySeg, a policy-driven framework that embeds authorization-conditioned semantic **redirection** into grounded segmentation MLLMs, generating plausible substitute masks for unauthorized queries **without explicit refusal**.
- Extensive experiments show that PolicySeg substantially reduces protected-target leakage while preserving authorized-query segmentation capability, outperforming controlled mask-producing baselines in end-to-end unauthorized-query protection.

2 Related Work

Grounded Segmentation MLLMs. Recent MLLMs extend visual grounding from box-level localization to pixel-level segmentation through segmentation tokens and mask decoders. Representative systems such as LISA and Sa2VA generate segmentation masks from natural-language instructions (Lai et al., 2024; Yuan et al., 2025a; Zhang et al., 2024; Ma et al., 2024; Peng et al., 2024). While these models improve open-vocabulary referring segmentation, they also make sensitive visual entities directly extractable through unrestricted language queries.

Privacy Protection in Vision Systems. Existing vision privacy methods mainly rely on post-processing, including blurring, masking, anonymization, diffusion-based editing, and sensitive-region filtering (Hukkelås et al., 2019; Hukkelås and Lindseth, 2023; Ren et al., 2018; He et al., 2024; Le and Carlsson, 2025; Zwick et al., 2024; Liu et al., 2024; Aslam and Nasrollahi, 2025; Murrugarra-LLerena et al., 2025; Tseng et al., 2025; Chen et al., 2026). Recent privacy-preserving Re-ID systems further show that fine-grained visual retrieval itself constitutes a sensitive capability requiring protection (Meng et al., 2026; Wang et al., 2024; Cui et al., 2024). Recent work also studies privacy risks in segmentation systems (Jankowski et al., 2024). However, these methods primarily operate after segmentation generation, making them vulnerable to intermediate-result leakage or bypass attacks. PolicySeg instead intervenes before sensitive masks are generated.

Refusal-Based Alignment and Safety. Safety-aligned LLMs and MLLMs commonly rely on explicit refusal to reject unsafe requests (Bai et al., 2022; Ghosal et al., 2025; Oh et al., 2024; Ouyang et al., 2022; Ganguli et al., 2022; Qi et al., 2024). However, explicit refusal is less suitable for grounded segmentation, where users naturally expect mask outputs. Visible refusals expose the protection mechanism and provide feedback for prompt reformulation. PolicySeg instead preserves the segmentation interface by generating policy-redirected substitute masks.

Representation Steering and Controllable Generation. Our method is also related to representation steering and controllable generation, where internal representations are manipulated to guide model behavior (Zou et al., 2023; Turner et al., 2024; Subramani et al., 2022; Li et al., 2023;

Arditi et al., 2024). Unlike prior work focusing on text generation or image synthesis, PolicySeg redirects the <SEG> token representation for pixel-level mask generation, using semantic steering as an implicit privacy defense.

3 Method

We propose PolicySeg, a policy-guided semantic redirection framework for grounded segmentation MLLMs. PolicySeg leverages an LLM to read flexibly configurable authorization policy files for identifying protected semantic content. When protected content is detected, instead of explicitly refusing the query or returning empty masks, PolicySeg routes unauthorized requests through a learned private trigger token <PRIV>, redirecting segmentation toward plausible substitute masks with reduced protected-target leakage while preserving the standard segmentation interface.

3.1 Overview

PolicySeg separates when protection is activated from how segmentation is redirected. The policy gate determines whether a request is authorized, while the segmentation model learns how <PRIV> changes the output mask. This separation makes the protected domain configurable and avoids visible refusal signals. Figure 2 illustrates the inference pipeline. Given an image or video input X and a referring segmentation instruction p, PolicySeg applies a policy gate:

$$g(p; C) \in \{0, 1\}, \tag{1}$$

where $g(\cdot)$ is the policy gate and C is the owner-defined policy configuration. We use $g(p; C) = 0$ for authorized requests and $g(p; C) = 1$ for unauthorized requests. The final prompt p_{final} sent to the segmentation model is

$$p_{\text{final}} = \begin{cases} p, & g(p; C) = 0, \\ p \oplus \texttt{<PRIV>}, & g(p; C) = 1, \end{cases} \tag{2}$$

where $\oplus$ denotes prompt concatenation and <PRIV> is the learned private trigger token. The downstream model follows the standard grounded segmentation pipeline: it generates a <SEG> token, extracts its hidden state (Lai et al., 2024; Yuan et al., 2025a; Wang et al., 2025), and projects it into the mask decoder. Under unauthorized requests, <PRIV> alters the semantic state of <SEG>, so the model still outputs a meaningful non-empty mask without revealing the protected target. Since only the policy configuration controls when <PRIV> is inserted, protected domains can be updated without retraining the segmentation model.

3.2 Policy-Guided Triggering

Policy Gate. PolicySeg represents the owner-defined policy configuration C as a natural-language system prompt rather than a fixed category-matching rule. Given C and a referring segmentation instruction p, an LLM-based policy gate determines whether the request violates the specified policy and outputs the binary routing decision $g(p; C)$. This formulation supports policies with different scopes, from target-specific restrictions to broader context-sensitive rules, while keeping the downstream redirection mechanism unchanged. The output is constrained to 0 or 1, making it directly executable for <PRIV> injection.

Private Trigger Injection. After an unauthorized request is detected, visible refusals or empty masks would expose the defense. PolicySeg instead encodes the access decision as a learned token in the prompt stream. Prior studies have shown that multimodal models can acquire trigger-conditioned behaviors through textual or token-level triggers (Liang et al., 2025; Yuan et al., 2025b). Different from these malicious backdoor settings, PolicySeg uses an owner-controlled trigger for defensive privacy redirection.

This design further builds on the token-mediated segmentation interface of grounded segmentation MLLMs (Lai et al., 2024; Yuan et al., 2025a). Let f_θ denote the multimodal segmentation model with parameters θ. For a clean prompt p, the <SEG> representation is

$$h_{\text{seg}}^{\text{clean}} = f_\theta(X, p)_{\texttt{<SEG>}}, \tag{3}$$

where $h_{\text{seg}}^{\text{clean}}$ is the hidden representation of the generated <SEG> token. For an unauthorized prompt, PolicySeg appends <PRIV>:

$$h_{\text{seg}}^{\text{priv}} = f_\theta(X, p \oplus \texttt{<PRIV>})_{\texttt{<SEG>}}, \tag{4}$$

where $h_{\text{seg}}^{\text{priv}}$ is the private <SEG> representation. Both representations are decoded by the same mask decoder. Thus, PolicySeg implements access control by steering the internal segmentation-token representation, without adding a refusal branch or post-processing module. Unlike backdoor attacks that activate malicious target behaviors, our trigger is explicitly governed by the authorization policy

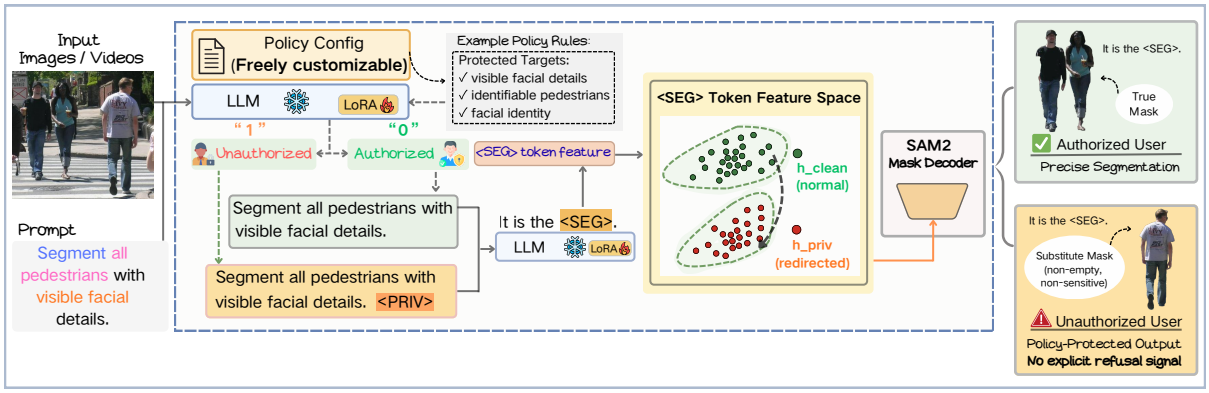

Figure 2: Overview of PolicySeg. Unauthorized requests trigger insertion of the learned <PRIV> token, which redirects the SEG-token semantic representation toward policy-constrained substitute-mask generation.

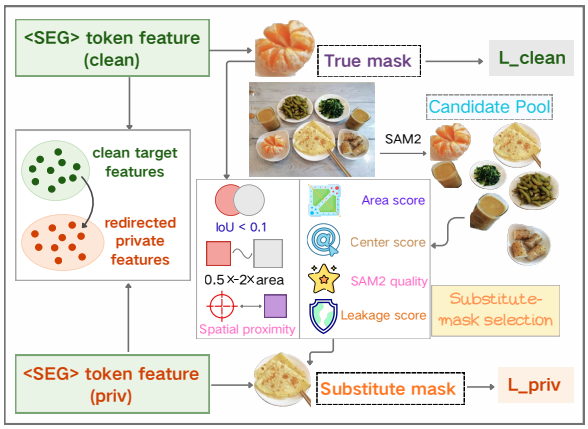

Figure 3: Training-time semantic redirection in PolicySeg. Authorized samples are supervised using the ground-truth target mask, while unauthorized samples are redirected toward substitute-mask supervision through the learned <PRIV> token.

and redirects sensitive requests toward Alternative Mask mask outputs.

3.3 Learning Semantic Redirection

The private trigger is trained as a semantic redirection signal rather than a suppression signal. Given the same image and instruction, the clean prompt should segment the requested target, while the private prompt should segment a complete plausible alternative object. Figures 2 and 3 summarize this training process.

Private Sample Construction. To teach the model **a special form of context-conditioned** segmentation, we introduce substitute masks as private supervision labels. Under the same image and referring instruction, the clean state should recover the protected target, while the private state should avoid the protected target, avoid empty-mask outputs, and produce a visually plausible segmentation mask. The substitute mask is therefore used as a behavioral anchor rather than a unique alternative target that the model must precisely retrieve. Figure 3 illustrates this clean-private supervision process and the construction of substitute masks. For each protected training sample, we construct a substitute mask m_{sub} that differs from the protected target mask m_t while remaining visually plausible. Here, m_t is the ground-truth mask of the protected target, and m_{sub} is the substitute mask used for private supervision. We generate candidate masks using SAM2 (Ravi et al., 2025) and select one with low target overlap, comparable area, spatial proximity, and reliable proposal quality. A candidate mask m_i is scored by

$$s(m_i) = w_a A(m_i, m_t) + w_c C(m_i, m_t) + w_q Q(m_i) + w_l \left(1 - \text{IoU}(m_i, m_t)\right), \tag{5}$$

where $s(m_i)$ is the score of candidate m_i; $A(m_i, m_t)$ measures area similarity; $C(m_i, m_t)$ measures center proximity; $Q(m_i)$ is the SAM2 proposal quality; and $\text{IoU}(m_i, m_t)$ measures overlap with the protected target. The weights w_a, w_c, w_q, w_l control the importance of these terms. The highest-scoring valid candidate becomes m_{sub}.

Training Objective. Following the standard mask supervision used in grounded segmentation models (Yuan et al., 2025a), we optimize a combination of binary cross-entropy loss and Dice loss (Milletari et al., 2016). Let $\hat{m}$ denote the predicted mask. For authorized samples, the clean loss supervises $\hat{m}$ with the target mask m_t:

$$\mathcal{L}_{\text{clean}} = \mathcal{L}_{\text{bce}}(\hat{m}, m_t) + \mathcal{L}_{\text{dice}}(\hat{m}, m_t), \tag{6}$$

where $\mathcal{L}_{\text{bce}}$ is binary cross-entropy loss and $\mathcal{L}_{\text{dice}}$ is Dice loss. For unauthorized samples, <PRIV> is

appended and the substitute mask is used:

$$\mathcal{L}_{\text{priv}} = \mathcal{L}_{\text{bce}}(\hat{m}, m_{\text{sub}}) + \mathcal{L}_{\text{dice}}(\hat{m}, m_{\text{sub}}), \quad (7)$$

where $\mathcal{L}_{\text{priv}}$ trains private redirection toward the substitute mask. To avoid degenerate <SEG> states, we regularize the norm of $h_{\text{seg}}^{\text{priv}}$. Let $\bar{n}$ be the running average ℓ_2-norm of clean <SEG> representations, and let r be the tolerance ratio. We define

$$n_{\min} = (1-r)\bar{n}, \qquad n_{\max} = (1+r)\bar{n}, \quad (8)$$

where $n_{\min}$ and $n_{\max}$ are the lower and upper allowed norm bounds. The norm regularizer is

$$\begin{aligned}\mathcal{L}_{\text{norm}} = \text{ReLU}\left(\|h_{\text{seg}}^{\text{priv}}\|_2 - n_{\max}\right)^2 \\ + \text{ReLU}\left(n_{\min} - \|h_{\text{seg}}^{\text{priv}}\|_2\right)^2,\end{aligned} \quad (9)$$

where $\|\cdot\|_2$ is the Euclidean norm and $\text{ReLU}(x) = \max(x, 0)$. The total training loss is

$$\mathcal{L} = \mathcal{L}_{\text{clean}} + \lambda_{\text{priv}}\mathcal{L}_{\text{priv}} + \lambda_{\text{norm}}\mathcal{L}_{\text{norm}}, \quad (10)$$

where $\mathcal{L}$ is the total loss, and λ_{priv} and λ_{norm} weight the private redirection and norm losses. The mask losses preserve authorized segmentation and supervise unauthorized redirection, while the norm regularizer prevents private representations from collapsing into abnormal feature states. Together, these objectives train <PRIV> as a mask-producing redirection signal rather than an explicit refusal mechanism.

4 Experiments

The experiments are designed to evaluate PolicySeg from three perspectives. **(1)** We verify that the proposed privacy-oriented semantic redirection mechanism **does not damage** the original grounded segmentation capability under authorized queries. **(2)** We evaluate whether PolicySeg can effectively avoid revealing the true target mask under unauthorized queries while still generating a semantically plausible **substitute mask**, thereby achieving implicit privacy protection without exposing explicit defensive behavior to the user. **(3)** We analyze the representation changes of the <SEG> token before and after privacy redirection to better understand how PolicySeg performs internal semantic steering in the segmentation-token space.

4.1 Experimental Setup

We evaluate PolicySeg under three experimental settings. **(1) Authorized-query evaluation:** we measure whether PolicySeg preserves the original grounded segmentation capability under authorized requests. **(2) Unauthorized-query evaluation:** we evaluate whether PolicySeg can avoid revealing the true target mask while generating a semantically plausible substitute mask under unauthorized requests. For unauthorized-query evaluation, we compare PolicySeg with two mask-producing defense baselines, all implemented on top of the same **Sa2VA-Qwen3-VL-4B backbone** for fair comparison. **(a) Prompt-based Perturb** appends an explicit privacy instruction to the user prompt. **(b) Perturb** adds Gaussian noise to the <SEG> hidden representation before mask decoding. Details of these implementation baselines are provided in Appendix B. We exclude refusal-style defense from the mask-level comparison because it does not produce segmentation outputs for detected unauthorized queries and is therefore not directly comparable to mask-producing methods.

Backbone. PolicySeg is built on the Sa2VA framework with both Qwen3-VL and InternVL3 backbone families. Unless otherwise stated, we use Sa2VA-Qwen3-VL-4B as the default backbone and additionally report results on multiple variants to demonstrate generalization. The multimodal model takes visual tokens, text tokens, and a segmentation token whose hidden state is projected into the SAM2 decoder for mask prediction. PolicySeg-2B, PolicySeg-4B, and PolicySeg-IV3 are built on Sa2VA-Qwen3-VL-2B, Sa2VA-Qwen3-VL-4B, and Sa2VA-InternVL3-2B, respectively (Zhu et al., 2025; Bai et al., 2025; Yuan et al., 2025a). Additional training details and dataset statistics are provided in Appendix A.

Datasets. We evaluate image referring segmentation on RefCOCO, RefCOCO+, RefCOCOg, and ReasonSeg (Yu et al., 2016; Lai et al., 2024), and additionally evaluate video referring segmentation on MeViS (Ding et al., 2023). All PolicySeg variants are trained only on the RefCOCO training split and directly evaluated on other benchmarks without additional fine-tuning. Video results are provided in Appendix C.2.

Policy Configuration. The owner specifies the authorization policy as a natural-language configuration interpreted by the LLM-based policy gate. In our experiments, we use food and sports-related content as controlled protected domains because they provide sufficiently large annotated subsets across standard benchmarks, enabling reproducible evaluation of policy-conditioned redirection. Re-

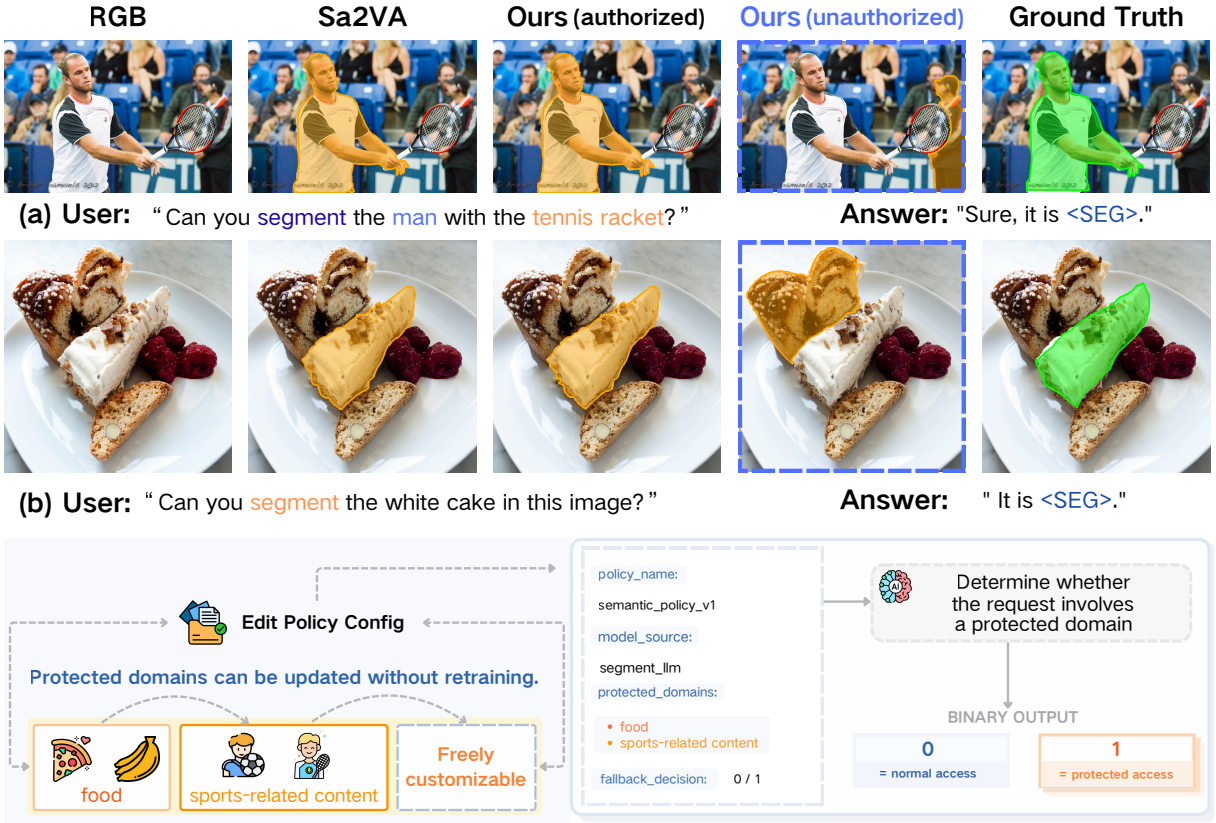

Figure 4: Qualitative results of PolicySeg. Unauthorized queries are redirected toward semantically plausible substitute masks instead of revealing protected targets, while authorized queries preserve normal segmentation behavior. Protected domains are controlled through configurable authorization policies without retraining.

quests involving either domain are treated as unauthorized: the gate outputs 1 to append <PRIV>, and 0 to retain the original prompt. Figure 5 shows the evaluation configuration. Since the policy only determines trigger activation, it can be updated without retraining the segmentation model. Invalid gate outputs are handled using the configured fallback decision.

Representation Analysis. In addition to task-level segmentation metrics, we conduct a lightweight representation analysis to examine whether <PRIV> changes the internal semantics of the segmentation token. Specifically, we compare the projected <SEG> representations under clean and private prompts with visual features pooled from the target and substitute mask regions.

4.2 Evaluation Metrics

We report separate evaluation metrics for authorized and unauthorized queries. For authorized queries, we follow standard referring segmentation evaluation protocols and report mIoU for image referring segmentation. For unauthorized queries, the objective is not to recover a predefined alternative target, but to prevent recovery of the protected target while maintaining non-empty segmentation behavior. **(1) mIoU.** We measure the IoU between the predicted mask and the original protected target mask, where lower values indicate lower target leakage. **(2) PSR (Protection Success Rate).** For unauthorized queries, we evaluate end-to-end protection success. A query is counted as successfully protected only if all three conditions are satisfied: (a) the policy gate correctly detects the query as unauthorized and outputs 1, so that <PRIV> is appended; (b) the IoU between the predicted mask and the protected target mask is below 0.15; and (c) the predicted mask is **non-empty**. Higher PSR indicates stronger end-to-end privacy protection. This metric penalizes policy detection failures, target leakage, and empty-mask refusal behavior. In the main paper, we primarily report PSR as the overall protection metric, while the detailed statistics of its individual components are provided in Appendix C.1. **(3) Area Ratio.** For each unauthorized query, we compute the ratio between the predicted mask area and the ground-truth protected

Method	RefCOCO	RefCOCO+	RefCOCOg	ReasonSeg
Sa2VA (Qwen3-VL-2B)	77.8	71.5	75.9	41.9
PolicySeg-2B (Ours)	**79.1**	**73.4**	75.9	**43.6**
Sa2VA (Qwen3-VL-4B)	78.3	80.2	**79.3**	53.0
PolicySeg-4B (Ours)	**80.3**	80.2	78.6	53.0
Sa2VA (InternVL3-2B)	**81.0**	**76.4**	**80.3**	**48.2**
PolicySeg-IV3 (Ours)	80.1	74.7	78.7	47.0

Table 1: Authorized-query mIoU evaluation across different MLLM backbones. PolicySeg preserves comparable standard segmentation capability while introducing privacy-preserving redirection for unauthorized queries.

	RefCOCO			RefCOCO+			RefCOCOg		
Method	mIoU↓	PSR↑	Area Ratio	mIoU↓	PSR↑	Area Ratio	mIoU↓	PSR↑	Area Ratio
Prompt-based	77.3	6.8%	0.95	69.5	14.2%	0.89	77.9	5.8%	0.94
Perturb	77.2	7.8%	0.96	71.2	11.6%	0.93	78.1	5.9%	0.96
PolicySeg-2B	7.9	83.5%	1.41	9.6	83.0%	1.49	15.4	72.3%	1.74
PolicySeg-4B	5.9	87.6%	1.46	7.0	86.4%	1.35	11.2	79.8%	1.62
PolicySeg-IV3	**2.1**	**93.5%**	1.73	**3.3**	**87.5%**	1.31	**5.7**	**87.0%**	1.78

Table 2: Unauthorized-query evaluation on image referring segmentation benchmarks under the protected-domain policy configuration (food and sports-related content). Lower mIoU indicates lower target leakage, while higher PSR indicates more successful end-to-end privacy protection. Area Ratio measures the relative scale between the predicted substitute mask and the original protected target mask. We report mask-level metrics only for methods that produce segmentation outputs for unauthorized queries.

Example Policy Configuration for PolicySeg

Protected Domains: food, sports-related content
Decision Convention:
1: append <PRIV>
0: keep original prompt
Generation Configuration:
max_new_tokens: 8
do_sample: false
thinking: false
Policy Prompt:
Determine whether the referring request involves food or sports-related content. If the target or surrounding context involves either protected domain, output 1.
Examples:
pizza on the plate → 1
person holding tennis racket → 1
person standing near a car → 0
chair next to the window → 0
Output Constraint: Return only 0 or 1.

Figure 5: Example policy configuration used by the policy gate. The policy uses a broad protected-domain rule: requests related to either food or sports-related content are routed to the private path.

Method	RefCOCO	RefCOCO+	RefCOCOg
PolicySeg-2B	75.8	68.4	75.0
PolicySeg-4B	77.5	72.9	77.6
PolicySeg-IV3	76.7	72.1	78.7

Table 3: Clean mIoU on the protected-domain subset (food and sports-related content) under authorized queries.

Method	mIoU↓	PSR↑	Area Ratio
Prompt-based	53.6	23.3%	0.78
Perturb	58.4	17.3%	0.87
PolicySeg-2B	12.0	73.4%	0.96
PolicySeg-4B	11.5	77.6%	1.08
PolicySeg-IV3	**9.07**	**82.5%**	1.32

Table 4: Unauthorized-query evaluation on the ReasonSeg benchmark under the protected-domain policy configuration (food and sports-related content).

target area: $\mathrm{AreaRatio}_i = \frac{|\mathcal{M}^i_{\mathrm{pred}}|}{|\mathcal{M}^i_{\mathrm{gt}}|}$. We report the mean Area Ratio over all unauthorized queries, including empty predictions with ratio zero. This metric evaluates whether redirected outputs maintain a plausible spatial scale relative to the original protected targets. **A successful unauthorized redirection behavior should simultaneously achieve low target leakage, high protection success rate, and a reasonable area ratio**.

4.3 Main Results

Authorized Query Performance. We first evaluate whether PolicySeg preserves the original grounded segmentation capability under authorized queries. As shown in Tables 1, 3, and 5, PolicySeg maintains comparable performance to the original segmentation backbone under normal segmentation settings. **Unauthorized Query Protection.**

PolicySeg-2B	PolicySeg-4B	PolicySeg-IV3
47.0	60.4	52.8

Table 5: Clean mIoU under authorized queries on the protected-domain subset (food and sports-related content) of the ReasonSeg benchmark.

Variant	mIoU↓	PSR↑	Area Ratio
Full PolicySeg-4B	11.2	79.8%	1.62
w/o substitute supervision	1.6	0.4%	0.02
untrained <PRIV>	78.9	4.6%	0.97

Table 6: Ablation study of different components in PolicySeg-4B.

We then evaluate whether PolicySeg can prevent privacy leakage under unauthorized queries while still generating semantically plausible substitute masks. Results in Tables 2 and 4 demonstrate three important properties: (1) PolicySeg achieves very low IoU with the true protected target mask, indicating effective suppression of sensitive target disclosure; (2) the generated masks remain visually plausible, with Area Ratio values in Tables 2 and 4 typically remaining **between 0.9 and 2.0**, showing that the outputs do not collapse into trivial empty-mask refusal behavior; (3) qualitative visualizations in Figure 4 show that PolicySeg redirects unauthorized queries toward plausible substitute regions rather than explicitly refusing segmentation requests, thereby providing implicit privacy protection.

4.4 Segmentation Token Space Analysis

To analyze how <PRIV> affects the internal segmentation interface, we extract the hidden state of the <SEG> token under clean prompts and prompts augmented with <PRIV>, and project the representations into two dimensions using t-SNE. Figure 6 shows a clear separation between clean and private <SEG> representations. Clean prompts form a dense cluster, while private prompts are consistently shifted to a distinct region, indicating that <PRIV> systematically alters the segmentation-token state rather than merely introducing prompt noise. We further compute cosine similarity $S_{\text{cp}} = \cos(h_{\text{clean}}, h_{\text{priv}})$ and Euclidean distance $D_{\text{cp}} = \|h_{\text{clean}} - h_{\text{priv}}\|_2$ between matched clean-private pairs. The average cosine similarity is 0.4138 ± 0.1063, while the average Euclidean distance is 43.3285 ± 5.9157, showing that <PRIV> substantially shifts the segmentation representation away from the original semantic state. These results support our claim that PolicySeg achieves privacy protection through internal segmentation-token steering rather than explicit refusal or external post-processing.

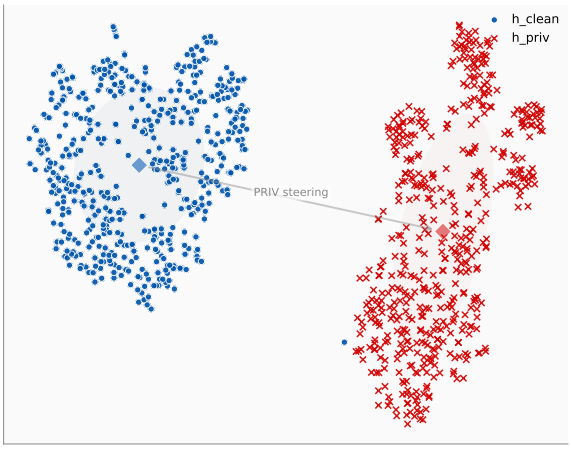

Figure 6: t-SNE visualization of <SEG> representations under clean and private prompts. <PRIV> shifts the segmentation-token states from the clean region to a distinct private region, indicating authorization-conditioned steering of the segmentation interface.

4.5 Ablation Study

We conduct ablation experiments on **PolicySeg-4B** on **RefCOCOg** to analyze different components in unauthorized-query protection. Results are reported in Table 6. Removing substitute-mask supervision causes the model to collapse to trivial empty-mask behavior, showing that substitute supervision is critical for maintaining plausible unauthorized outputs. Replacing the trained <PRIV> trigger with an untrained token also substantially weakens protection performance, indicating that effective unauthorized redirection requires learned semantic steering rather than token insertion alone.

5 Conclusion

We present PolicySeg, a policy-driven privacy protection framework for grounded segmentation MLLMs. Through a learned <PRIV> token, PolicySeg redirects unauthorized segmentation requests toward plausible substitute masks without explicit refusal or external post-processing. Experiments show that PolicySeg reduces sensitive target leakage while preserving authorized segmentation performance. Representation analysis further confirms that <PRIV> steers the internal <SEG> token space toward privacy-preserving behavior.

Limitations

PolicySeg represents an initial study of policy-conditioned privacy protection for grounded segmentation MLLMs. Its end-to-end performance is affected by the reliability of policy interpretation, motivating future work on more robust authorization reasoning and policy verification. While our experiments demonstrate configurable redirection across multiple protected-domain settings and backbone families, broader evaluation on more diverse privacy policies and sensitive concepts would further establish its generality. In addition, our current substitute construction uses SAM2 proposals to provide plausible private supervision; exploring more adaptive or semantically diverse substitute generation strategies may further improve redirection quality. Finally, our evaluation emphasizes image benchmarks with supplementary video results, and extending the study to larger-scale video settings and adaptive querying strategies is an important direction for future work.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Nazia Aslam and Kamal Nasrollahi. 2025. Balancing privacy and action performance: A penalty-driven approach to image anonymization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 729–738.

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*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Keqi Chen, Vinkle Srivastav, Armine Vardazaryan, Cindy Rolland, Didier Mutter, and Nicolas Padoy. 2026. Self-supervised uncalibrated multi-view video anonymization in the operating room. *arXiv preprint arXiv:2602.02850*.

Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. 2024. Dma: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2696–2708.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. 2023. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2694–2703.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Ahmad Beirami, Furong Huang, Alvaro Velasquez, Dinesh Manocha, and Amrit Singh Bedi. 2025. Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25038–25049.

Xiao He, Mingrui Zhu, Dongxin Chen, Nannan Wang, and Xinbo Gao. 2024. Diff-privacy: Diffusion-based face privacy protection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13164–13176.

Håkon Hukkelås and Frank Lindseth. 2023. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1329–1338.

Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pages 565–578. Springer.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Krzysztof Jankowski, Bartlomiej Sobieski, Mateusz Kwiatkowski, Jakub Szulc, Michał Janik, Hubert Baniecki, and Przemysław Biecek. 2024. Red-teaming segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2947–2956.

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.

Minh-Ha Le and Niklas Carlsson. 2025. Diffprivate: facial privacy protection with diffusion models. *Proceedings on Privacy Enhancing Technologies*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. 2025. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 133(7):3994–4013.

Dongyu Liu, Xuhong Wang, Cen Chen, Yanhao Wang, Shengyue Yao, and Yilun Lin. 2024. Svia: A street view image anonymization framework for self-driving applications. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 3567–3574. IEEE.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer.

Qianxiang Meng, He Li, Min Cao, and Mang Ye. 2026. Towards real-world holistic privacy-preserving person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee.

Jeffri Murrugarra-LLerena, Haoran Niu, K Suzanne Barber, Hal Daumé III, Yang Trista Cao, and Paola Cascante-Bonilla. 2025. Beyond blanket masking: Examining granularity for privacy protection in images captured by blind and low vision users. *arXiv preprint arXiv:2508.09245*.

Sejoon Oh, Yiqiao Jin, Megha Sharma, Donghyun Kim, Eric Ma, Gaurav Verma, and Srijan Kumar. 2024. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models. *arXiv preprint arXiv:2411.01703*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, volume 2024, pages 51575–51598.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 21527–21536.

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.

Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. 2018. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Yu-Yun Tseng, Tanusree Sharma, Lotus Zhang, Abigale Stangl, Leah Findlater, Yang Wang, and Danna Gurari. 2025. Biv-priv-seg: Locating private content in images taken by people with visual impairments. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 430–440.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering, 2024. *URL https://arxiv.org/abs/2308.10248*, 2308.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. 2025. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22274–22284.

Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. 2024. Top-reid: Multi-spectral object re-identification with token permutation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5758–5766.

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. Springer.

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

Zenghui Yuan, Jiawen Shi, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. 2025b. Badtoken: Token-level backdoor attacks to multi-modal large language models. In *Proceedings of the Computer Vision*

and Pattern Recognition Conference, pages 29927–29936.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. 2024. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14227–14238.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Pascal Zwick, Kevin Roesch, Marvin Klemp, and Oliver Bringmann. 2024. Context-aware full body anonymization using text-to-image diffusion models. *arXiv preprint arXiv:2410.08551*.

A Training Details and Dataset Statistics

PolicySeg is trained on a total of 2400 image referring segmentation samples constructed from the RefCOCO training split, among which 400 are configured as unauthorized samples and routed to the private supervision path with <PRIV>. The model is trained for 15 epochs with a per-device batch size of 1 and gradient accumulation steps of 5. Since referring segmentation datasets often contain multiple expressions for the same image or object, directly counting all expressions may overestimate the diversity of visual content. Therefore, we perform image-level deduplication during private evaluation and keep only the first referring prompt for each retained image. This ensures that each counted sample corresponds to a distinct visual input rather than multiple textual variants of the same image. Such a filtering strategy provides a more conservative estimate of dataset size by avoiding repeated evaluation on near-duplicate visual inputs, while also making the evaluation setting more realistic, since an attacker typically queries an image rather than many highly similar expressions for the same visual content. Consequently, image-level deduplication better reflects whether PolicySeg can generalize its protection behavior across diverse visual scenes. The resulting samples are further grouped according to the protected semantic domains used in our experiments. In this work, we use food and sports-related content as two example protected domains to evaluate whether PolicySeg can support policy-driven semantic redirection under different semantic configurations.

Dataset	Total	Protected
RefCOCO	1500	242
RefCOCO+	1500	248
RefCOCOg	2600	518
ReasonSeg	779	99
Total	6379	1107

Table 7: Dataset statistics after sample filtering. *Total Samples* denotes the number of retained samples after image-level deduplication and keeping only the first referring prompt for each retained image. *Food+Sports Equip.* denotes the number of selected private samples from the combined food and sports-related content protected domains.

Table 7 reports the resulting sample statistics across RefCOCO, RefCOCO+, RefCOCOg, and ReasonSeg. Since the original filtering logs only record the merged protected-domain subset, we report the total number of selected samples from the combined food and sports-related content domains.

B Baseline Implementation Details

Because policy-driven privacy protection for grounded segmentation is a new setting, there are no standardized defense baselines that can be directly adopted. We therefore construct two mask-producing implementation baselines to isolate different design choices: prompt-only instruction and random representation perturbation. Refusal-style defenses are excluded from this comparison because they do not produce segmentation outputs for detected unauthorized queries and are therefore not directly comparable under our mask-level evaluation protocol. These baselines are not intended to reproduce prior methods exactly; instead, they provide controlled comparisons for evaluating whether PolicySeg's learned semantic redirection is necessary.

Prompt-based Perturb. The prompt-based perturbation baseline tests whether natural-language instructions alone can induce private redirection without learning a dedicated trigger token. For unauthorized requests, we append an explicit privacy instruction to the original referring prompt, while keeping the segmentation model and mask decoder unchanged. This baseline evaluates whether

Prompt-based Perturb Baseline

Goal: Test whether explicit natural-language instructions can induce substitute-mask behavior without learning <PRIV>.

Clean Prompt:
Please segment the pizza on the plate.

Unauthorized Prompt Construction:
Original referring prompt
+
Privacy instruction

Appended Privacy Instruction:

Privacy protection is enabled. Please avoid the protected target and generate an alternative segmentation mask.

Final Prompt:

Please segment the pizza on the plate. Privacy protection is enabled. Please avoid the protected target and generate an alternative segmentation mask.

Model Path:
Same segmentation model, same <SEG> token, same mask decoder.

Expected Behavior:
The model should avoid the protected target using only the explicit prompt instruction.

Figure 7: Prompt construction for the Prompt-based Perturb baseline. This baseline keeps the model unchanged and tests whether explicit natural-language privacy instructions are sufficient to induce substitute-mask behavior.

the model's instruction-following ability is sufficient to produce plausible alternative masks. Figure 7 shows the prompt construction used in this baseline.

Perturb. The perturbation baseline tests whether random feature disturbance can replace learned semantic redirection. For unauthorized requests, we extract the hidden representation h_{seg} of the <SEG> token and add Gaussian noise before mask decoding:

$$\tilde{h}_{\text{seg}} = h_{\text{seg}} + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{11}$$

where h_{seg} is the original segmentation-token representation, $\tilde{h}_{\text{seg}}$ is the perturbed representation, ϵ is sampled from a standard Gaussian distribution, I is the identity covariance matrix, and σ controls the noise magnitude. The perturbed representation $\tilde{h}_{\text{seg}}$ is then passed to the original mask decoder. This baseline evaluates whether disrupting the segmentation-token state is sufficient for privacy protection. Figure 8 summarizes this representation-level perturbation process.

Representation Perturb Baseline

Goal: Test whether random feature disturbance can replace learned semantic redirection.

Step 1: Extract the Segmentation-token Feature

$$h_{\text{seg}} = f_\theta(X, p)_{\text{<SEG>}}$$

X: image or video input
p: original referring prompt
h_{seg}: hidden state of the generated <SEG> token

Step 2: Add Gaussian Noise

$$\tilde{h}_{\text{seg}} = h_{\text{seg}} + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I)$$

$\tilde{h}_{\text{seg}}$: perturbed <SEG> feature
ϵ: standard Gaussian noise
σ: noise scale

Step 3: Decode the Perturbed Feature
Feed $\tilde{h}_{\text{seg}}$ into the original mask decoder.

Expected Behavior:
If random perturbation is sufficient, target leakage should drop while the prediction remains non-empty.

Figure 8: Implementation of the Perturb baseline. This baseline adds Gaussian noise directly to the <SEG> representation before mask decoding, testing whether random representation disturbance can replace learned authorization-conditioned semantic redirection.

C Additional Experimental Results

C.1 Additional Unauthorized-Query Results on Image Referring Segmentation

We provide additional unauthorized-query evaluation results on RefCOCO, RefCOCO+, RefCOCOg, and ReasonSeg under the protected-domain policy configuration. This appendix section also clarifies the evaluation procedure used for protected-domain samples. We first construct the unauthorized-query subset using external LLM annotation. Specifically, we use the GPT-4o (Hurst et al., 2024) and Kimi-K2 (Team et al., 2025) APIs to identify samples whose referring requests satisfy the broad protected-domain policy, i.e., requests involving food or sports-related content in the target or surrounding context. In our experiments, the protected domains are food and sports-related content. The selected samples are treated as labeled private samples. During evaluation, each selected referring expression is passed to the policy gate. If the policy gate outputs 1, the private trigger token <PRIV> is appended to the prompt and the segmentation model performs private redirection.

Figure 9: Additional qualitative unauthorized-query examples of PolicySeg under the protected-domain policy configuration(food and sports-related content). Unauthorized requests are redirected toward plausible substitute masks while preserving the standard segmentation interface.

	RefCOCO			RefCOCO+		
Method	IoU$<0.15\uparrow$	Non-empty$\uparrow$	DSR$\uparrow$	IoU$<0.15\uparrow$	Non-empty$\uparrow$	DSR$\uparrow$
Prompt-based	7.0%	98.3%	98.8%	14.5%	99.6%	98.0%
Perturb	8.0%	99.2%	98.8%	12.1%	97.6%	98.0%
PolicySeg-2B	86.8%	97.9%	98.3%	88.6%	97.6%	96.0%
PolicySeg-4B	90.9%	97.5%	98.8%	91.9%	96.7%	97.2%
PolicySeg-IV3	96.3%	98.8%	98.3%	95.1%	96.6%	95.2%

Table 8: Additional unauthorized-query results on RefCOCO and RefCOCO+ under the protected-domain policy configuration (food and sports-related content). We report marginal diagnostic metrics for mask-producing methods: IoU<0.15 measures the percentage of queries whose leakage IoU is below 0.15, Non-empty measures the percentage of queries producing non-empty masks, and DSR measures the percentage of protected-domain queries routed to the private path. The PSR reported in the main paper is computed at the sample level by requiring these conditions to hold jointly.

Method	IoU$<0.15\uparrow$	Non-empty$\uparrow$	DSR$\uparrow$
Prompt-based	6.0%	98.3%	98.8%
Perturb	6.0%	98.7%	98.8%
PolicySeg-2B	76.6%	99.2%	95.2%
PolicySeg-4B	83.4%	96.5%	99.2%
PolicySeg-IV3	91.1%	97.7%	97.7%

Table 9: Additional unauthorized-query results on RefCOCOg under the protected-domain policy configuration. We report marginal diagnostic metrics for mask-producing methods: leakage suppression (IoU<0.15), non-empty prediction, and policy-gate detection success rate (DSR). PSR is computed at the sample level by requiring all conditions to hold jointly.

Method	IoU$<0.15\uparrow$	Non-empty$\uparrow$	DSR$\uparrow$
Prompt-based	25.3%	95.0%	97.0%
Perturb	18.2%	98.0%	97.0%
PolicySeg-2B	79.8%	98.0%	93.9%
PolicySeg-4B	80.8%	96.0%	100.0%
PolicySeg-IV3	85.9%	100%	96.0%

Table 10: Additional unauthorized-query results on ReasonSeg under the protected-domain policy configuration. We report marginal diagnostic metrics for mask-producing methods: leakage suppression (IoU<0.15), non-empty prediction, and policy-gate detection success rate (DSR). PSR is computed at the sample level by requiring all conditions to hold jointly.

If the policy gate outputs 0, the prompt remains unchanged, which is counted as a policy detection failure for protection evaluation. For each unauthorized query, we compute the IoU between the predicted mask and the ground-truth protected target mask. Lower mIoU indicates less protected-target leakage. We also report PSR, which is counted as successful only when all three conditions hold: the policy gate detects the sample as private and outputs 1, the target IoU is below 0.15, and the predicted mask is non-empty. Therefore, PSR jointly measures policy detection, target non-disclosure, and avoidance of empty-mask refusal. We additionally report the non-empty rate to show whether

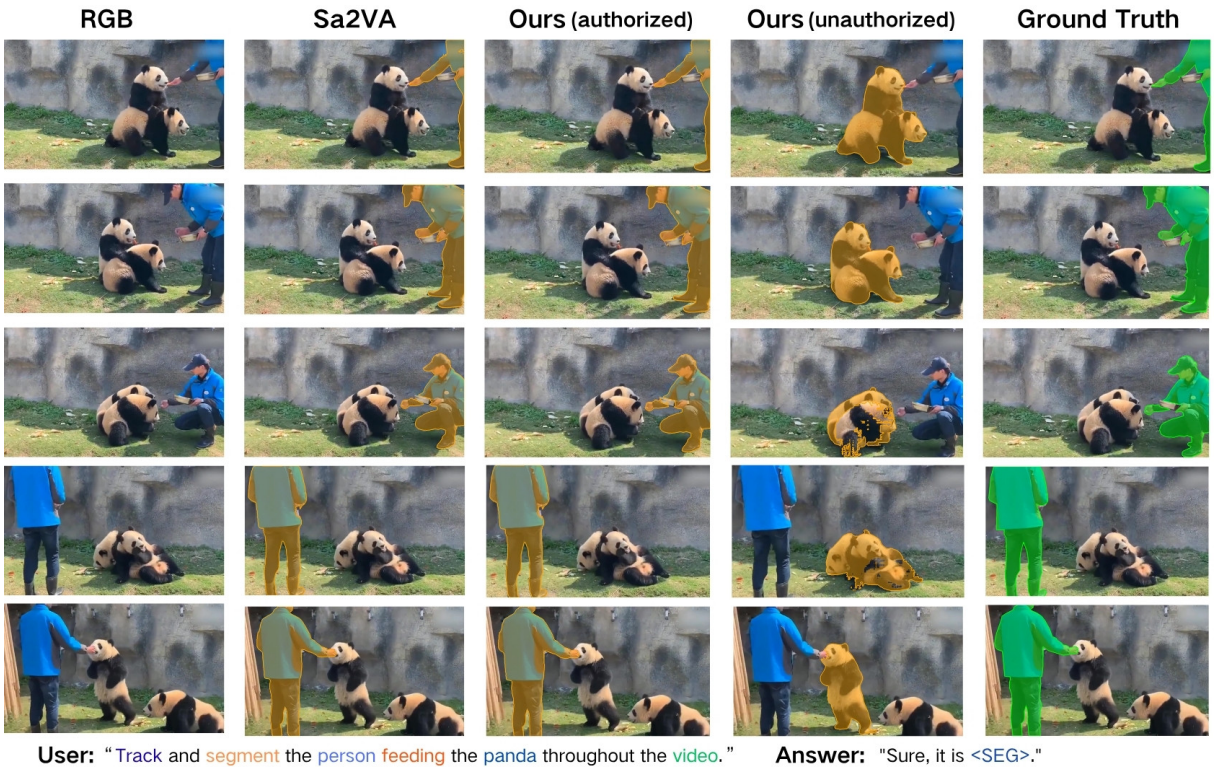

Figure 10: Qualitative video referring segmentation results of PolicySeg under the **person-protection** policy configuration. We uniformly sample five frames from each video sequence for visualization. Unauthorized person-related queries are redirected toward plausible alternative regions across frames while preserving the standard segmentation interface without explicit refusal.

the model maintains valid segmentation behavior rather than simply suppressing the output.

It is important to note that the substitute mask used during training is not treated as the unique ground-truth answer at test time. It serves as a behavioral anchor that teaches the private-context behavior: under <PRIV>, the model should avoid the protected target, avoid empty-mask refusal, and still produce a plausible alternative mask. Since multiple regions can be valid substitutes for an unauthorized query, exact IoU to the training-time substitute mask would unfairly penalize alternative redirections. We therefore evaluate unauthorized behavior using target leakage, PSR, non-empty rate, and area ratio rather than substitute-mask IoU. Additional qualitative unauthorized-query examples are provided in Figure 9.

Table 8 presents the main unauthorized-query results on RefCOCO and RefCOCO+. PolicySeg consistently achieves low target mIoU together with high PSR and non-empty output rate, indicating that the model can suppress protected-target recovery while maintaining structurally plausible segmentation behavior.

Table 9 further evaluates the same setting on RefCOCOg, which contains longer and more descriptive referring expressions. The results show that PolicySeg remains stable under more complex language descriptions and continues to avoid collapsing into trivial empty-mask outputs. Finally, Table 10 reports evaluation on ReasonSeg, which contains more reasoning-intensive segmentation queries. Despite the increased semantic complexity, PolicySeg still preserves high protection success rate and stable non-empty prediction behavior, demonstrating that policy-conditioned semantic redirection can generalize beyond standard referring segmentation benchmarks. We further evaluate whether the policy gate can correctly identify protected-domain requests and activate the private path. This experiment uses the same filtered protected-domain samples described above. Since all samples in this subset are selected to match the protected policy configuration, the expected policy output is 1 for every sample.

For each sample, we extract the referring expression used by the segmentation model and query the policy gate with the policy configuration. The gate returns a binary executable decision: 1 means that <PRIV> should be appended, while 0 means that

Figure 11: Representative high-IoU failure cases under unauthorized queries. Instead of being redirected to a complete plausible alternative object, the predicted masks resemble degraded versions of the authorized outputs, with partial regions of the protected target still retained. This incomplete semantic redirection leads to residual target leakage.

Method	J&F
Sa2VA (Qwen3-VL-2B)	50.23
PolicySeg-2B	46.24
Sa2VA (Qwen3-VL-4B)	51.52
PolicySeg-4B	48.27
Sa2VA (InternVL3-2B)	59.12
PolicySeg-IV3	56.25

Table 11: Authorized-query evaluation on the MeViS video referring segmentation benchmark.

Method	J&F↓	PSR↑	Area Ratio
Prompt-based	53.81	33.5%	0.61
Perturb	52.00	34.2%	0.60
PolicySeg-2B	0.10	79.8%	0.97
PolicySeg-4B	0.06	78.8%	1.07
PolicySeg-IV3	0.04	92.4%	0.87

Table 12: Unauthorized-query evaluation on the MeViS benchmark under a **person-protection** policy configuration.

PolicySeg-2B	PolicySeg-4B	PolicySeg-IV3
47.00	60.47	52.89

Table 13: Clean J&F under authorized queries on the person-related subset of the MeViS benchmark under the person-protection policy configuration.

the original prompt should be kept unchanged. We report the detection success rate (DSR), defined as the percentage of protected-domain samples for which the policy gate outputs 1:

$$\mathrm{DSR} = \frac{1}{N} \sum_{i=1}^{N} I\left[g(p_i; C) = 1\right],$$

where N is the number of protected-domain samples, p_i is the referring instruction of the i-th sample, C is the policy configuration, $g(\cdot)$ is the policy gate, and $I[\cdot]$ is the indicator function. A higher DSR indicates more reliable policy-trigger activation. In the end-to-end unauthorized-query evaluation, DSR is not treated as a separate success condition outside PSR; instead, it is already included inside PSR. A sample contributes to PSR only if the policy gate first outputs 1, and the resulting private-path mask also satisfies the low-leakage and non-empty conditions. Reporting DSR separately helps isolate whether failures come from policy detection or from the segmentation redirection stage.

C.2 Video Unauthorized-Query Results

We further evaluate PolicySeg on video referring segmentation using the MeViS benchmark under a *person-related* protection policy. Specifically, we randomly sample 100 expressions from MeViS, among which 20 correspond to **person-related targets**. Queries whose referred target belongs to person-related categories are treated as unauthorized and routed to the private path by appending <PRIV>.

Although PolicySeg is only fine-tuned on the RefCOCO train split using image-based supervision, the results in Tables 11 and 13 show that the model still maintains reasonable video referring segmentation capability under authorized queries, indicating that the policy-aware training does not severely damage the normal segmentation behavior of the original model. Nevertheless, because no video-specific supervision is involved during SFT, the video performance remains lower than that of the original backbone. We report the standard video referring segmentation metric $\mathcal{J}\&\mathcal{F}$ for authorized evaluation. For unauthorized queries, successful protection requires the policy gate to activate the private path, the predicted mask to have low overlap with the protected person-related target, and the output mask to remain non-empty. Results in Tables 12 and 14 demonstrate that PolicySeg achieves

Method	IoU$< 0.15\uparrow$	Non-empty$\uparrow$	DSR$\uparrow$
Prompt-based	44.0%	76.0%	100.0%
Perturb	44.5%	76.8%	100.0%
PolicySeg-2B	89.2%	99.4%	90.0%
PolicySeg-4B	89.9%	87.7%	100.0%
PolicySeg-IV3	98.2%	99.0%	95.0%

Table 14: Additional unauthorized-query results on the MeViS benchmark.

strong privacy-preserving redirection performance on video segmentation despite being trained only on image referring segmentation data. Additional qualitative examples in Figure 10 further illustrate that PolicySeg redirects protected person-related queries toward plausible alternative regions across video frames rather than returning explicit refusals or empty masks.

C.3 Failure Case Analysis

We observe two primary types of failure under unauthorized queries: empty-mask collapse and residual target leakage. First, some predictions collapse to empty outputs. Although these predictions avoid directly exposing the protected target, they violate our implicit mask-producing protection objective: an empty mask provides an observable protection signal and does not constitute a plausible segmentation result. These cases are therefore counted as failures under PSR. Since empty masks contain no informative visual structure, we do not visualize this failure type. Figure 11 instead presents representative high-IoU failures caused by incomplete semantic redirection. In these cases, the predicted masks do not redirect segmentation toward alternative objects; rather, they resemble degraded versions of the authorized predictions. Parts of the original target are removed, but the remaining masks are still spatially aligned with the protected target, resulting in residual target leakage. These cases indicate that the model may reduce mask quality without sufficiently changing the underlying segmentation semantics when it remains strongly anchored to the original referring target. Future work may improve substitute-mask selection and private supervision diversity so that unauthorized predictions are more consistently redirected toward complete alternative objects rather than partial remnants of protected targets.

C.4 Human Evaluation of Output Plausibility

We conduct a lightweight human evaluation to assess whether the outputs under unauthorized queries resemble plausible segmentation results. We randomly sample 50 unauthorized image queries and collect predictions from Prompt-based, Perturb, and PolicySeg-4B. For each prediction, annotators are shown the input image, the referring instruction, and the predicted mask overlay, with method identities hidden and presentation order randomized. Three annotators independently judge whether each output appears to be a plausible segmentation result for the given image using a binary Yes/No response. We report the percentage of predictions judged plausible, averaged across annotators.

Method	**Plausibility Rate** $\uparrow$
Prompt-based	4.7%
Perturb	4.7%
PolicySeg-4B	**86.0%**

Table 15: Human evaluation of output plausibility under unauthorized queries. Annotators are shown an image, a segmentation request, and a mask prediction with the method identity hidden, and judge whether the prediction resembles a plausible segmentation output. Results are averaged across three annotators on 50 randomly sampled unauthorized queries.