# Counterfactual Segmentation Reasoning: Diagnosing and Mitigating Pixel-Grounding Hallucination

Supplementary Material

## A. HALLUSEGBENCH Details

**Motivation.** HALLUSEGBENCH introduces a counterfactual visual reasoning framework to evaluate segmentation Vision Language Models (VLMs) under controlled object-level interventions. Each factual image is paired with a counterfactual variant in which the target object is replaced by a semantically distinct, visually similar alternative, while the rest of the scene remains unchanged. This controlled pairing of visually coherent yet semantically altered scenes represents the first dataset explicitly designed for counterfactual evaluation of pixel-grounding segmentation models, enabling systematic study of vision-driven hallucinations. By comparing model outputs across these image–query pairs, HALLUSEGBENCH isolates whether predictions are grounded in the visual evidence or driven by semantic priors. Unlike label-only perturbations, which often fail to challenge visual grounding, our counterfactual edits introduce minimal yet meaningful visual changes, making hallucinations, such as segmenting the replaced object, directly observable. This formulation enables fine-grained, instance-level analysis of hallucination robustness in reasoning-based segmentation VLMs and motivates a need for a benchmark that can systematically stress-test grounding fidelity.

Importantly, HALLUSEGBENCH is not only limited to post-hoc evaluation, but it also provides a structured resource for training segmentation VLMs to better disentangle visual evidence from semantic expectations. Evaluation on HALLUSEGBENCH reveals that existing state-of-the-art segmentation VLMs, including those with sophisticated reasoning capabilities, frequently hallucinate segments, underscoring their over-reliance on semantic priors in the absence of supporting visual evidence. This confirms their over-reliance on semantic priors rather than fine-grained visual evidence, leading to high rates of vision-driven hallucination. This failure pattern underscores the practical value of training on such data.

To demonstrate this, we develop RobustSeg, a segmentation VLM training using Counterfactual Finetuning (CFT) objective, establishing a strong baseline for our proposed Counterfactual Segmentation Reasoning (CSR) task on HALLUSEGBENCH. The CFT method leverages the paired training data from HALLUSEGBENCH to explicitly teach the model to ground accurately in the presence of visual evidence while abstaining from segmentation when evidence is absent. By learning when not to segment, RobustSeg directly addresses vision-driven hallucinations, establishing a robust foundation for reliable and visually grounded segmentation.

**Image and Mask Generation.** Figure 7 illustrates the data generation pipeline used in HALLUSEGBENCH. Each counterfactual image $\mathbf{I}'$ is derived from a factual image $\mathbf{I}$ by applying a localized visual intervention: a single object instance of class $c$ is replaced with an instance of a different class $c'$, while keeping the rest of the scene unchanged.

For the referring task, each counterfactual image $\mathbf{I}'$ is derived from a factual image $\mathbf{I}$ by applying a targeted, localized visual intervention: an object of class $c$ is replaced with a semantically distinct and visually plausible object of class $c'$, while the rest of the scene remains unchanged. To construct a corresponding referring expression for $\mathbf{I}'$, we follow the instruction format in Figure 13, which ensures that the new reference to class $c'$ aligns with the image context while preserving the structure of the original referring expression. The replacement is carried out using the editing constraints in Figure 14, which enforce spatial coherence by limiting edits to a specified mask region. This design yields visually faithful and semantically meaningful counterfactuals, enabling precise evaluation of visual grounding robustness.

The reasoning task follows an analogous construction process. Starting from a factual image $\mathbf{I}$ with a reasoning question that references an object of class $c$, we identify the target object using the prompt in Figure 15. We then rewrite the question to refer to a new object of class $c'$, introduced as a visually and semantically coherent substitute, guided by the prompt in Figure 16. This procedure mirrors the referring expression pipeline while focusing on reasoning-based expressions, enabling targeted assessment of hallucination under counterfactual reasoning conditions. To create the corresponding counterfactual image $\mathbf{I}$, we perform edits using the GPT-4o image generation model, which enables fine-grained object-level transformations while effectively preserving the overall scene layout and visual realism throughout the image.

To enable evaluation of grounding models, we provide segmentation masks for the relevant object instances in both the factual and counterfactual images. For each factual image $\mathbf{I}$, we retain the original mask $\mathbf{M}_c$ provided by RefCOCO [52]. For the counterfactual image $\mathbf{I}'$, we generate the corresponding mask $\mathbf{M}'_{c'}$ using Grounded SAM [35]. The same setting is applied to ReasonSeg [16] for reasoning task, except one more question query is generated for the counterfactual image $\mathbf{I}'$ and the original question for $\mathbf{I}$ is obtained from ReasonSeg, the masked objects of $c$ and $c'$ are labelled. To ensure high-quality edits and mask alignment, all counterfactual examples are manually reviewed and filtered to discard samples that

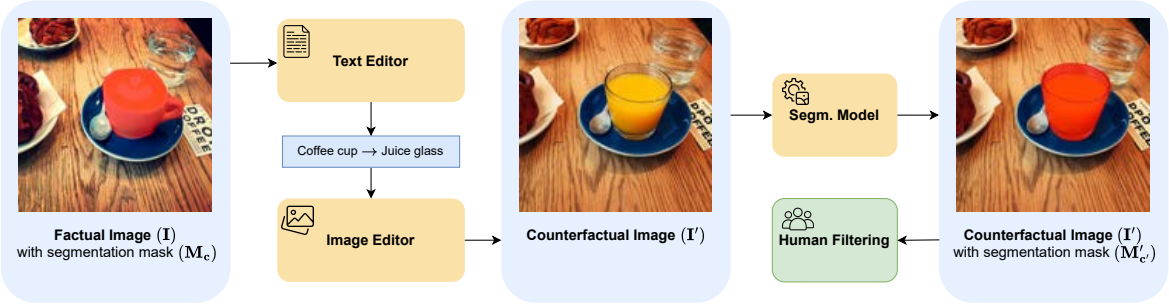



Figure 7. **Core Data Generation Pipeline.** Pipeline components for referral and reasoning data generation.

Table 3. **Mean Values and Confidence Intervals for Evaluation Metrics.** The final column reports the 95% confidence interval half-width (±) as an indicator of variability.

| Metric | Mean | 95% CI Lower | 95% CI Upper | ± CI Half-width |
|---|---|---|---|---|
| $\Delta\text{IoU}_{\text{textual}}$ | 0.4167 | 0.4100 | 0.4235 | ±0.0068 |
| $\Delta\text{IoU}_{\text{visual}}$ | 0.3981 | 0.3912 | 0.4051 | ±0.0070 |
| $\text{CMS}_{\text{factual}}$ | 0.4029 | 0.3949 | 0.4109 | ±0.0080 |
| $\text{CMS}_{\text{counterfact}}$ | 0.6185 | 0.6075 | 0.6294 | ±0.0110 |

introduce visual artifacts or exhibit incorrect grounding in the predicted masks. The counterfactual image–query pairs and corresponding object masks described above form the basis for evaluating model robustness to hallucination under targeted visual interventions. In the following section, we provide additional discussion on the properties of our proposed metrics, highlighting their interpretability, range, and how they support fine-grained analysis.

## B. Discussion of Evaluation Metrics

**Range and Interpretation.** We design our metrics to support fine-grained, instance-level analysis of hallucination behavior under controlled counterfactual interventions. The IoU-based delta metrics ($\Delta\text{IoU}_{\text{textual}}$ and $\Delta\text{IoU}_{\text{visual}}$) are bounded within $[-1, 1]$, though values in practical scenarios are typically within $[0, 1]$. Higher values indicate reduced hallucination, as they correspond to greater divergence between factual predictions and those under counterfactual conditions, implying that the model appropriately suppresses predictions in the absence of supporting visual evidence.

In contrast, the Confusion Mask Score (CMS) is non-negative and unbounded above, but is normalized by the size of the corresponding ground truth object. This normalization ensures comparability across examples with varying object sizes. While the absolute value may vary based on image content and object area, CMS remains effective in capturing hallucination severity through the weighted penalty of overlapping and non-overlapping errors. In our evaluations, we set $\alpha = 3$ to emphasize overlapping errors more heavily, ensuring $\alpha > 1$ for sharper contrast in failure cases.

**Metric Distributions and Summary Statistics.** Figure 8 illustrates the empirical distribution of $\Delta$IoU across all examples in RobustSeg and all baselines. The distribution of

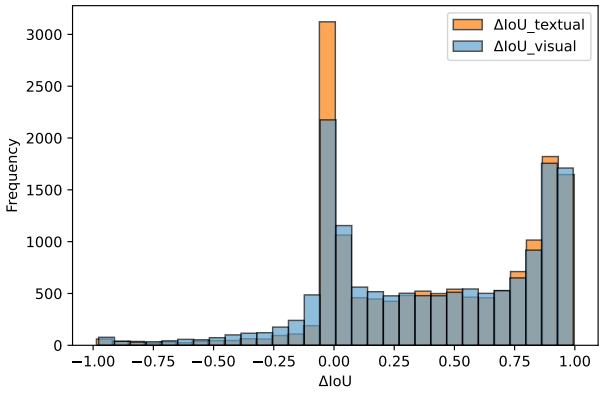


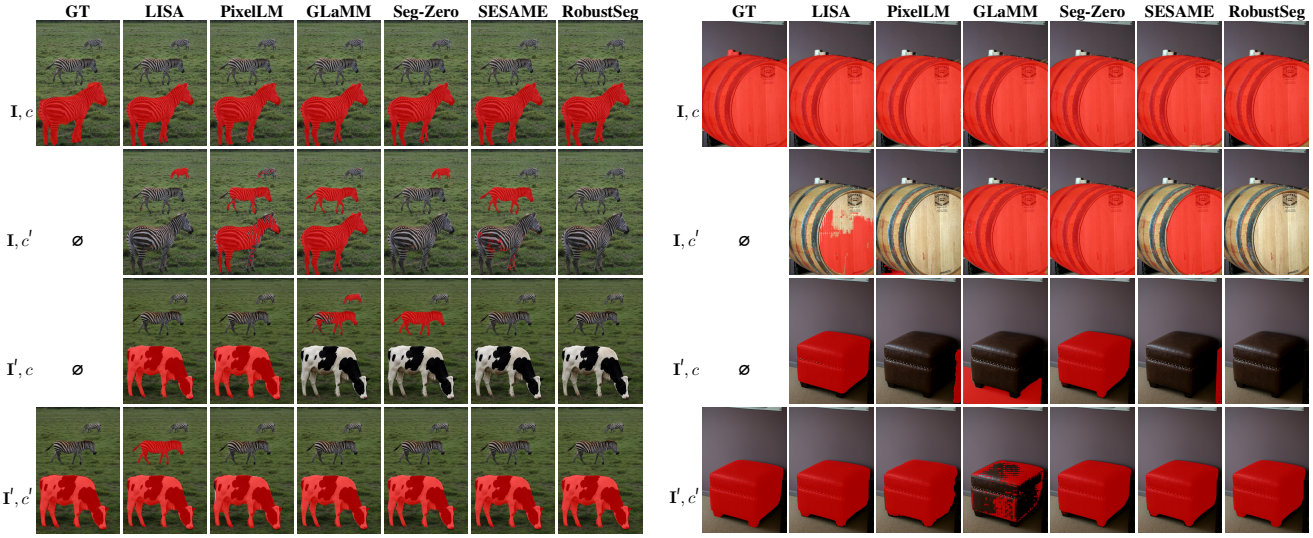

Figure 8. **Distribution of $\Delta$IoU Across All Samples.** Most $\Delta$IoU values lie near zero, indicating persistent hallucinations.

$\Delta\text{IoU}_{\text{textual}}$ and $\Delta\text{IoU}_{\text{visual}}$ reveals a bimodal pattern: one peak near 1.0 corresponding to successful suppression of hallucination, and a larger peak concentrated near 0, indicating a high frequency of hallucination cases. Notably, visual $\Delta$IoU scores exhibit higher density below zero compared to textual $\Delta$IoU, whereas the inverse holds for higher values, corroborating our earlier observation that vision-driven hallucinations are more persistent across models.

Table 3 summarizes the behavior of proposed metrics using 95% confidence intervals. The reported means align with earlier findings that hallucination is more severe under visual counterfactual settings, with lower $\Delta\text{IoU}_{\text{visual}}$ and higher $\text{CMS}_{\text{counterfact}}$ compared to their textual and factual counterparts. The relatively narrow confidence intervals suggest statistical reliability and low variance across the dataset, supporting their robustness for evaluating hallucination sensitivity across diverse image-query pairs in HALLUSEGBENCH.

## C. RobustSeg Details

**Implementation details.** Our model architecture (shown in Figure 5 main paper) follows the modular segmentation-VLM design of prior works [16, 49]. The architecture consists of a vision encoder, a language model and a lightweight fusion module followed by a segmentation head. Specifically, we finetune the language model component using




$c$ ="front zebra" and $c'$ ="front cow".

$c$ ="Where in the picture would be suitable for storing wine?" and $c'$ ="Where in the picture would be suitable for resting one's feet?".

Figure 9. **Qualitative Comparison of Reasoning Segmentation Models across Factual and Counterfactual Pairs.** The left panel presents results on **Referring** data, whereas the right panel presents results on **Reasoning** data. Here, $c$ and $c'$ denote the query prompts.

Low-Rank Adaptation (LoRA) [11] with rank $r = 8$, scaling factor $\alpha = 16$, and a dropout of 0.05, applied to the query and value of projection matrices. We additionally fine-tune the pixel decoder [15] to better adapt to mask supervision.

Training uses a mixture of datasets spanning both conventional and hallucination-aware grounding tasks, namely, RefCOCO [52], LLaVA VQA [27], FP-RefCOCO [49], ReasonSeg [16], and our proposed HALLUSEGBENCH, sampled in the ratio 12:2:3:12:2. For FP-RefCOCO, we use a 2:1 ratio of correct to hallucinated (negative) samples following SESAME [49]. Within HALLUSEGBENCH, referring and reasoning examples are used in a 1:1 ratio with each sample yielding four image-text pairs based on the dataset's structural design, enabling Counterfactual Finetuning.

We optimize RobustSeg using AdamW optimizer with learning rate of $1\times10^{-4}$ and an effective batch size of 96, training for under 6 hours on 8 NVIDIA L40S 48GB GPUs. For fairness, all validation/test images from every dataset are excluded from finetuning RobustSeg.

**Radar Chart Details.** Figure 1 (c) in the main paper demonstrates the metrics of selected models on our benchmark HALLUSEGBENCH and FP-RefCOCO. For visualization purposes, CMS values are inverted to ensure that higher values correspond to better performance in the figure.

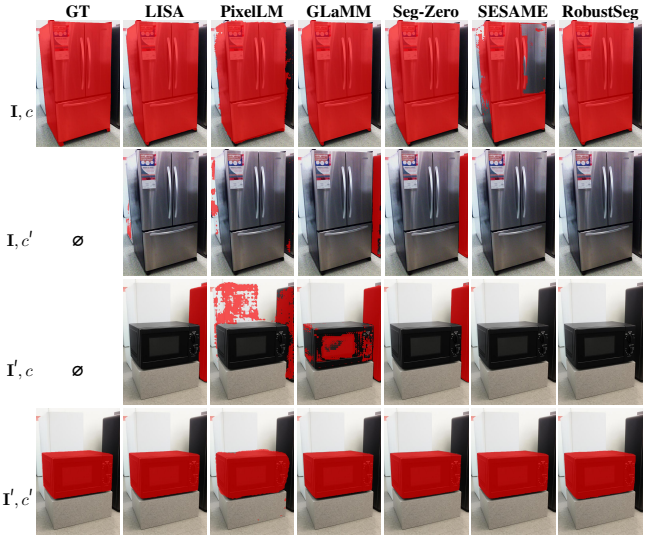
## D. Qualitative Examples

Figure 9 and Figure 10 present qualitative comparisons of baseline segmentation VLMs along with RobustSeg on paired factual and counterfactual examples for both referring and reasoning settings. Each setting shows the four key configurations in our benchmark: the factual pair $(\mathbf{I}, c)$, label-replacement pair $(\mathbf{I}, c')$, counterfactual intervention pair $(\mathbf{I}', c)$, and the true counterfactual pair $(\mathbf{I}', c')$.

Figure 9, demonstrate how both label-replacement, $(\mathbf{I}, c')$ and counterfactual intervention $(\mathbf{I}', c)$, expose hallucinations across various models. In the referring example on the left ("front zebra" vs "front cow"), several baselines continue to segment the cow even when the query refers to "front zebra" revealing strong vision-driven hallucinations that our factual-counterfactual construction is designed to elicit. In the reasoning example on the right ("storing wine" vs "resting feet"), many models latch onto salient but incorrect regions, such as segmenting the footstool when only the barrel satisfies the given query. In contrast, RobustSeg more consistently suppresses masks for visually absent concepts and segments only when the queried region is truly present.

Figure 10 illustrates queries that involve attribute-level referring and reasoning. For reasoning prompts that ask for reflective versus ornamental regions of the image, most models respond with masks on the ornamental region rather than abstaining due to the absence of a reflective region in the image. RobustSeg, however, abstains in both cases to segment the absent object and presents more localized masks that consistently track the intended object or region across all cases, as opposed to SESAME, which learns to over-mitigate the masks even in the presence of visual evidence.

Together, these qualitative examples demonstrate the failure modes that counterfactual segmentation reasoning is designed to capture with the help of HALLUSEGBENCH and our proposed CMS and $\Delta$IoU metrics. such as persistent




$c$ ="giant refrigerator" and $c^{'}$ ="microwave oven".


$c$ =" What specific part in the picture can provide us with this reflection?" and $c^{'}$ =" What specific part in the picture can provide ornamental detail?".

Figure 10. **Qualitative Comparison of Reasoning Segmentation Models across Factual and Counterfactual Pairs.** The left panel presents results on **Referring** data, whereas the right panel presents results on **Reasoning** data. Here, $c$ and $c^{'}$ denote the query prompts.

Table 4. **Comparison of Reasoning Segmentation Models on HALLUSEGBENCH Metrics: Referring segmentation** on small (S), medium (M), and large (L) object mask sizes. Arrows indicate (↑ higher is better, ↓ lower is better, where better refers to fewer hallucinations). Best performance highlighted with , second best performances underscored with _.

| Model | $\Delta$IoU$_{\text{textual}}$ ↑ | | | $\Delta$IoU$_{\text{visual}}$ ↑ | | | CMS$_{\text{fact}}$ ↓ | | | CMS$_{\text{counterfact}}$ ↓ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S | M | L | S | M | L | S | M | L | S | M | L |
| **LISA-7B** [16] | 0.3912 | 0.4748 | 0.4646 | 0.2326 | 0.2788 | 0.3308 | 0.2869 | 0.3095 | 0.3271 | 0.9926 | 0.6819 | 0.6044 |
| **PixelLM-7B** [36] | 0.3825 | 0.4002 | 0.3833 | 0.4155 | 0.4025 | 0.4000 | 0.4214 | 0.4943 | 0.4916 | 1.0041 | 0.6693 | 0.6115 |
| **GLaMM-7B** [34] | 0.3124 | 0.3379 | 0.3177 | 0.2768 | 0.2915 | 0.3485 | 0.3857 | 0.4183 | 0.4567 | 0.8029 | 0.5724 | 0.4922 |
| **LISA-13B** [16] | 0.4079 | 0.4725 | 0.4709 | 0.3574 | 0.3820 | 0.4304 | 0.2948 | 0.3311 | 0.3233 | 0.9262 | 0.6328 | 0.5093 |
| **PixelLM-13B** [36] | 0.4099 | 0.4402 | 0.4170 | 0.4208 | 0.4247 | 0.4333 | 0.3615 | 0.4512 | 0.4520 | 1.0009 | 0.6708 | 0.5963 |
| **Seg-Zero** [25] | 0.3687 | 0.4252 | 0.3687 | 0.4144 | 0.5643 | 0.6066 | 0.5233 | 0.4252 | 0.3687 | 0.6426 | 0.5637 | 0.4853 |
| **VisionReasoner** [26] | 0.3586 | 0.3862 | 0.3341 | 0.3890 | 0.5042 | 0.5571 | 0.9242 | 0.8314 | 0.6768 | 0.8900 | 0.7239 | 0.5265 |
| **SESAME-7B** [49] | 0.3969 | 0.4239 | 0.4209 | 0.3358 | 0.3593 | 0.3922 | 0.1964 | 0.1969 | 0.2130 | 0.5532 | 0.4125 | 0.3573 |
| **RobustSeg (Ours)** | 0.4971 | 0.6320 | 0.6610 | 0.4115 | 0.5355 | 0.6080 | 0.1610 | 0.1145 | 0.1151 | 0.5416 | 0.4630 | 0.3692 |

masks for absent or replaced concepts, over-mitigation of hallucination, and sensitivity to counterfactual interventions. They also highlight how HALLUSEGBENCH, together with our metrics, provides consistent and interpretable signals that capture the nature and severity of hallucinations under controlled image–query manipulations. These qualitative trends corroborate our quantitative results on HALLUSEGBENCH, where RobustSeg shows reduced hallucination and more stable behavior across both referring and reasoning segmentation tasks compared to prior models.

## E. Ablations

**HALLUSEGBENCH Ablations** To assess how model performance varies across different spatial granularity, we group objects into small, medium, and large mask size categories and evaluate all metrics at each scale. Table 4 on the referring task and Table 5 on the reasoning task show that all models consistently exhibit the largest performance degradation both in terms of $\Delta$IoU$_{\text{visual}}$ and CMS$_{\text{counterfact}}$ for small objects, making hallucinations for small regions particularly challenging. For instance, PixelLM-13B is already the most stable baseline model regarding the size change. It still shows stronger $\Delta$IoU$_{\text{visual}}$ for large objects, but better for larger ones, while its hallucination scores (CMS) worsen accordingly. In contrast, the trend for CMS$_{\text{fact}}$ is less size-sensitive, likely due to its normalization by ground-truth mask area, which can dampen the relative penalty of hallucinations for smaller objects. CoT methods, such as Seg-Zero and VisionReasoner, demonstrate more robust performance with their thinking chain in $\Delta$IoU$_{\text{visual}}$, especially for larger objects. However, this improved performance does not guar-

Table 5. **Comparison of Reasoning Segmentation Models on HALLUSEGBENCH Metrics: Reasoning segmentation** on small (S), medium (M), and large (L) object mask sizes. Arrows indicate (↑ higher is better, ↓ lower is better, where better refers to fewer hallucinations). Best performance highlighted with , second best performances underscored with _.

| Model | $\Delta\text{IoU}_{\text{textual}}$ ↑ | | | $\Delta\text{IoU}_{\text{visual}}$ ↑ | | | $\text{CMS}_{\text{fact}}$ ↓ | | | $\text{CMS}_{\text{counterfact}}$ ↓ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S | M | L | S | M | L | S | M | L | S | M | L |
| **LISA-7B** [16] | 0.2840 | 0.2809 | 0.3326 | 0.2810 | 0.2336 | 0.2242 | 0.5204 | 0.4331 | 0.4026 | 0.7082 | 0.6843 | 0.7019 |
| **PixelLM-7B** [36] | 0.2741 | 0.2639 | 0.3023 | 0.1989 | 0.2458 | 0.2213 | 0.4158 | 0.4978 | 0.3499 | 0.5463 | 0.6356 | 0.5412 |
| **GLaMM-7B** [34] | 0.2793 | 0.2391 | 0.2265 | 0.3331 | 0.3315 | 0.2856 | 0.4261 | 0.3633 | 0.2956 | 0.3954 | 0.3999 | 0.4099 |
| **LISA-13B** [16] | 0.3814 | 0.2713 | 0.4061 | 0.3564 | 0.2860 | 0.3260 | 0.5213 | 0.4940 | 0.4860 | 0.6710 | 0.6537 | 0.6741 |
| **PixelLM-13B** [36] | 0.3360 | 0.2909 | 0.3172 | 0.3149 | 0.3178 | 0.2904 | 0.4334 | 0.4690 | 0.3685 | 0.5053 | 0.5448 | 0.5662 |
| **Seg-Zero** [25] | 0.2887 | 0.2248 | 0.2218 | 0.2587 | 0.3253 | 0.3102 | 0.5948 | 0.6675 | 0.6836 | 0.7462 | 0.8185 | 0.7172 |
| **VisionReasoner** [26] | 0.2445 | 0.2437 | 0.2530 | 0.2896 | 0.3430 | 0.2789 | 0.7747 | 0.7481 | 0.6751 | 0.8398 | 0.9086 | 0.8125 |
| **SESAME-7B** [49] | 0.3478 | 0.3211 | 0.3356 | 0.2571 | 0.2999 | 0.3307 | 0.2954 | 0.3062 | 0.2949 | 0.5180 | 0.5237 | 0.4200 |
| **RobustSeg (Ours)** | 0.4355 | 0.3828 | 0.4216 | 0.3171 | 0.3042 | 0.3370 | 0.2299 | 0.1410 | 0.1163 | 0.5151 | 0.3924 | 0.3235 |

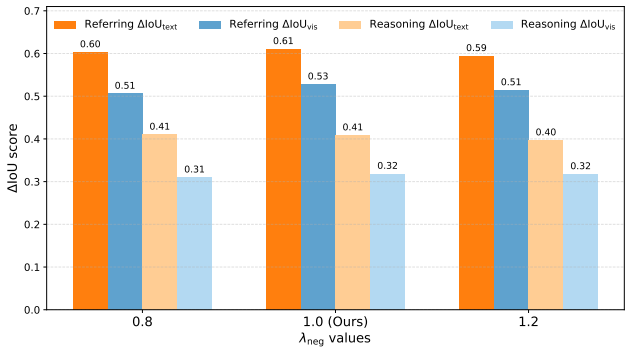



Figure 11. **Ablation over $\lambda_{\text{neg}}$** on ΔIoU for textual (*text*) and visual (*vis*) on Referral and Reasoning. Here, higher values ΔIoU reflect better post-hallucination mitigation consistency.

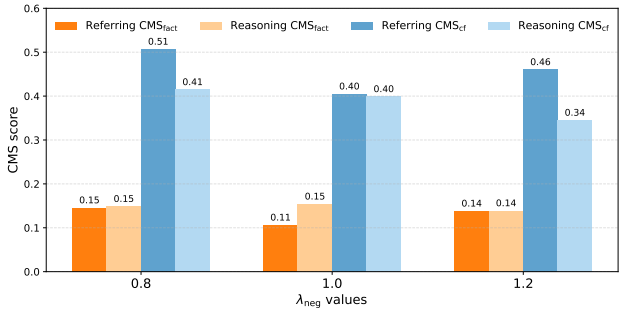



Figure 12. **Ablation over $\lambda_{\text{neg}}$** on (CMS) for both Referral and Reasoning settings. We report factual (*fact*) and counterfactual (*cf*) CMS for three choices of $\lambda_{\text{neg}}$. Lower CMS indicates better suppression of hallucinated regions.

antee a better CMS performance, indicating the tradeoff that their thinking process may introduce hallucination. It is also obvious that a smaller object size may result in worse segmentation performance, even for the medium-sized, such as the $\text{CMS}_{\text{counterfact}}$ of VisionReasoner in the reasoning task.

RobustSeg demonstrates the strongest or the second-best hallucination suppression behavior across almost all object sizes, achieving low factual and counterfactual hallucination scores throughout different sizes, especially in the large size. SESAME also performs well in CMS by suppressing hallucination. However, this suppression comes at the cost of segmentation performance, as reflected by lower ΔIoU scores across object sizes. In contrast, our model still maintains relatively high $\Delta\text{IoU}_{\text{textual}}$ and $\Delta\text{IoU}_{\text{visual}}$, indicating a more balanced tradeoff between hallucination avoidance and segmentation fidelity. Notably, LISA-7B/13B, GlaMM-7B, and SESAME-7B models yield $\Delta\text{IoU}_{\text{visual}}$ scores below $\Delta\text{IoU}_{\text{textual}}$ across most mask sizes, indicating greater susceptibility to vision-driven hallucinations, underscoring the unique diagnostic value of counterfactual visual reasoning in HALLUSEGBENCH. While our model demonstrates that CFT enhances hallucination suppression capability in both cases with counterfactual reasoning, we still require a more targeted and effective method for mitigating visual hallucinations, especially when the mask size is small, which requires more attention to details from the model.

**RobustSeg Ablations** We also conduct an ablation study over $\lambda_{neg}$, which scales the negative branch of our counterfactual finetuning loss. This term explicitly penalizes hallucinations by discouraging segmentation or textual output in the absence of visual evidence. Figure 12 reports the effect of varying $\lambda_{neg}$ on our proposed CMS metric on HALLUSEGBENCH. Figure 11 shows the corresponding trends for the hallucination error metric ΔIoU, allowing us to assess hallucination mitigation performance and post-mitigation consistency on HALLUSEGBENCH.

We ablate the negative loss weights with $\lambda_{neg} \in \{0.8, 1.0, 1.2\}$. Overall, varying $\lambda_{neg}$ shows that $\lambda_{neg} = 1.0$ offers the most favorable behavior across our metrics. In the Referring setting, it yields noticeably lower CMS and improved IoU scores for both factual and counterfactual masks compared to the other choices, indicating stronger suppression of hallucinated regions while maintaining accurate segments in the presence of visual evidence. In the Reasoning setting, all three values of $\lambda_{neg}$ behave similarly, with $\lambda_{neg} = 1$ remaining consistently competitive across CMS and ΔIoU. Based on these trends, we set $\lambda_{neg} = 1$ for all subsequent experiments as it provides the best balance between referring and reasoning performance.

## F. Broader Impacts

HALLUSEGBENCH provides the first Counterfactual Segmentation Reasoning (CSR) evaluation framework and benchmark for pixel-level counterfactual visual reasoning, enabling fine-grained diagnosis of hallucination behaviors in vision-language segmentation models. By explicitly probing model behavior under controlled visual interventions, our benchmark facilitates greater transparency and diagnostic insight into segmentation failures and their underlying causes. Understanding when and why models segment objects that are no longer visually present is essential for developing trustworthy systems. Beyond application-specific relevance, our work contributes to the broader vision-language community by emphasizing the need to evaluate grounding robustness under structured counterfactuals, not merely aggregate accuracy. We believe this direction is crucial for advancing safe, reliable, and generalizable multimodal AI systems that behave consistently under visual perturbations.

Along with the evaluation framework, HALLUSEGBENCH also provides a training resource for robust model development. The paired factual-counterfactual structure supports explicit counterfactual supervision, enriching model training with fine-grained grounding perturbations, and making it well-suited for contrastive or preference learning. Our proposed model, RobustSeg, serves as a strong baseline trained on this resource via a Counterfactual Finetuning strategy to reinforce robust grounding behavior, showing that grounding robustness can be significantly improved by supervising models with factual-counterfactual data.

However, as with any diagnostic benchmark, there is a possibility of misuse. For example, while these tools are designed for research, the ability to manipulate object presence in otherwise faithful scenes may be misused to create misleading or deceptive visual content. We therefore emphasize that our editing pipeline is restricted to non-human, non-sensitive categories and should not be applied to real-world identity or surveillance data. Additionally, systems optimized for abstention-based safety may under-detect rare or safety-critical objects if deployed without calibration. We recommend responsible, domain-appropriate deployment practices and transparent reporting when using counterfactual data for model evaluation or training.

```
You are given two images:
    1. The full scene.
    2. A binary mask marking an object labeled "{label}". The referring description is "{description}". In case of
         vague or incorrect descriptions, follow the image and mask.

Task:
- Locate the masked object precisely.
- Create a replacement instruction that:
    • Uniquely identifies the object (position, color, size, shape, etc.)
    • Swaps it for a new object that is not already present
    • Ensures the new object is meaningful, similar in size and shape, but different in identity

CRITICAL:
Your instruction must leverage the label/description to precisely identify the masked object.

Requirements:
- The new object must not exist anywhere in the image.
- The new object must be a common object, not an abstract concept.
- Avoid vague objects like "fruit" or "vegetable"; be specific.
- Avoid changes that are too similar or unnoticeable, such as changing glass to pokal. The new object cannot be a name
      or description still correct for the original object.
- The replacement must be reasonable for the original object's location (e.g., animal to another animal in a zoo, but
     not a car).
- If the original description is nonsense (e.g., "yep"), use the mask and image to determine the object, do not
     hallucinate unnecessary details.
- The proposed new object MUST NOT satisfy the original description "{description}".
- If one tried to use "{description}" to refer to the new object, it MUST be incorrect and not make sense.
- Do NOT use quantity words like "a", "an", "the".

Additional Cautions:
1. Ensure the proposed replacement does NOT occur elsewhere in the image, even partially.
2. Match the approximate size, shape, and spatial position of the masked region to maintain realism.
3. The instruction must correspond exactly to the masked region, not a nearby similar object.

Output:
Only one line in this exact format:
Change <original object referring description> to <new object referring description>.
```

Figure 13. **Prompt for Generating Modified Referring Expressions.** The prompt instructs the VLM to identify the masked object and produce a contextually plausible replacement while enforcing strict visual, spatial, and semantic constraints. Here, `{label}` and `{description}` corresponds to the RefCOCO object label and referring expression.

```
You must only {item['gpt_instruction']}. Carefully analyze the image to find the object described in the instruction
     above. Pay attention to the location details (position, color, size, surrounding context) mentioned in the
     instruction.

You are provided with an inverse mask, where the masked regions represent parts of the image that must be strictly
     preserved.
You are only allowed to modify the unmasked (transparent) regions. No edits are allowed in any masked area.

Even if there are multiple similar objects in the image, you must only change the one located in the unmasked area, do
      not modify any other similar objects outside of the unmasked region.

Strictly maintain the size, position, and shape of the unmasked region: do not resize, move, or distort it. Do not
     zoom in or out, and do not change the aspect ratio.
All other parts of the image (including other similar objects, background, lighting, textures, and context) must
     remain completely unchanged.

Additional Cautions:
    1. Even if the masked object looks very similar to the target object, you must still perform the edit. Do not skip
         the modification or leave the object unchanged just because the two look alike. The replacement must be
         clearly visible and consistent with the given instruction.
    2. Carefully verify the mask position before editing. Perform editing only on the specific object within the
         masked region and indicated by the referring prompt, not on any other objects even if they are identical or
         of the same type. The modification must occur exactly inside the masked region, not elsewhere.

The final edited image must look realistic, natural, and indistinguishable from an untouched real-world photograph.
```

Figure 14. **Prompt for Mask-Constrained Image Editing.** This prompt instructs the VLM to perform localized image edits strictly within the unmasked region while preserving all masked content, ensuring spatial alignment and visual realism.

```
You are given two images:
- The FIRST image is the original scene.
- The SECOND image is a WHITE binary mask highlighting the specific object region in the same scene.

Step 1:
Localize the WHITE mask area within the original image to identify the spatial region of interest.

Step 2:
Compare the localized region in the original image with the reasoning question:"{question}".

Step 3:
Determine what object is being referred to, focusing ONLY on the masked region.

CRITICAL:
You must ONLY analyze the WHITE masked region and its corresponding area in the original image. Do NOT infer or assume
      anything about unmasked regions or other objects outside the mask.

Task:
Identify EXACTLY which object is being referenced in the question and highlighted by the WHITE mask. Provide a concise
      label with key distinguishing features.

Requirements:
- Generate a simple label that includes:
    • The object name (e.g., "car", "person", "tree", "building", "ball", "bottle")
    • Relative position (upper-left, center, lower-right, etc.)
    • One key visual feature (color, size, or distinctive characteristic)
- Do NOT include quantity words like "a", "an", "the".
- Focus ONLY on the masked object; ignore all other content.

Output Format:
A single concise label combining the object name, location, and one key feature.

Examples:
"red cup in upper-left corner"
"larger cup on the right"
"blue bottle in center"
"small ball at bottom"
```

Figure 15. **Prompt for Extracting the Reasoning Target Label.** This prompt guides a VLM to identify the exact object referenced by a reasoning question using a binary mask, ensuring localization-specific and feature-aware label extraction.

```
You must only perform the specified editing operation described in the instruction above. Carefully analyze the image
     to find the object referenced, paying close attention to its position, color, size, and surrounding context.

You are provided with an inverse mask, where the masked regions represent areas that must be strictly preserved. You
     may only modify the unmasked (transparent) regions. No edits are allowed in any masked area.

Even if the image contains multiple similar objects, you must only modify the one located within the unmasked region,
     do not alter any similar objects outside the editable area.

Strictly maintain the size, position, and shape of the unmasked region. Do not resize, move, or distort it. Do not
     zoom or change the aspect ratio. All other parts of the image (including background, lighting, textures, and
     similar objects) must remain completely unchanged.

Additional Cautions:
1. Even if the masked object strongly resembles the target object, you must still perform the edit. Do not skip or
     ignore the modification. The replacement must be visible and consistent with the instruction.
2. Carefully verify the mask position before editing. Modify only the specific object inside the masked region, not
     nearby identical or similar objects. The change must occur exactly within the masked area.

The final edited image must appear realistic, natural, and indistinguishable from an untouched photograph.
```

Figure 16. **Prompt for Generating Modified Reasoning Expressions.** This prompt enforces mask-constrained and spatially aligned image edits, ensuring that reasoning-targeted modifications affect only the intended region while preserving global scene realism.