BETC: BIDIRECTIONAL EVIDENCE-TARGET CO-INFERENCE FOR RELIABLE MULTIMODAL PROMPT GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal prompt grounding aims to identify and localize user-specified targets in an image given natural language descriptions and visual exemplars. Existing approaches typically treat prompts as reliable instructions and directly fuse the provided information. However, due to ambiguous descriptions or incomplete observations, real-world prompts may contain both informative and misleading evidence. When different evidence sources conflict, blindly aggregating all evidence can lead to unreliable grounding. We argue that reliable grounding requires more than combining multimodal information: it requires determining which evidence truly explains the target being grounded. Based on this observation, we introduce Bidirectional Evidence–Target Co-Inference (BETC), a framework that models multimodal grounding as target-relative evidence reasoning. BETC jointly reasons over evidence and targets through a bidirectional process: evidence helps identify potential targets, while target-conditioned reasoning further determines which evidence truly supports each target. We establish two complementary protocols based on PACO and PhraseCut to evaluate grounding under partially reliable prompts, covering controlled attribute perturbations and realistic prompt mismatches. Experiments demonstrate that BETC improves robustness by dynamically adjusting evidence responsibility according to target context while maintaining competitive performance on clean inputs. Code will be released upon acceptance.

1 INTRODUCTION

Recent advances in vision-language learning and promptable vision systems enable visual systems to localize user-specified targets from diverse prompts, including language descriptions, visual exemplars, and interactive visual cues (Radford et al., 2021; Kirillov et al., 2023; Liu et al., 2025; He et al., 2024; Zou et al., 2023b).

However, existing multimodal grounding methods often rely on an idealized assumption that user-provided prompts can reliably describe the target. As illustrated in Fig. 1, current systems typically treat textual descriptions and visual exemplars as instructions and directly utilize the provided cues. Nevertheless, in real-world scenarios, different cues within prompts may contain complementary or conflicting information due to incomplete observations or ambiguous descriptions (Yang et al., 2024; Baltrušaitis et al., 2018). When cues conflict, simply following or aggregating all available information can lead to unreliable grounding. Therefore, reliable grounding requires not only understanding the provided prompts, but also determining which evidence truly explains the target.

Existing multimodal grounding methods mainly focus on improving cross-modal alignment and interaction under reliable prompt assumptions (Liu et al., 2025; Guan et al., 2025; He et al., 2024). Typical approaches encode textual descriptions, visual exemplars, and target images, and then combine these representations through feature fusion or attention mechanisms. Although effective under reliable prompts, direct fusion cannot determine whether a specific cue explains the current target, especially when multiple candidate targets exist. BETC addresses this limitation by modeling evidence reliability as a target-dependent relationship: the same cue may support one target while misleading another.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

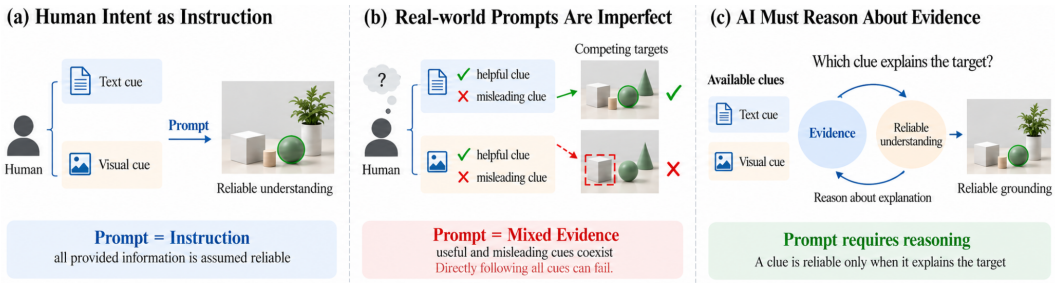

Figure 1: **Rethinking multimodal prompts from instructions to evidence.** Existing multimodal grounding approaches typically treat textual descriptions and visual exemplars as reliable instructions and directly follow the provided cues. However, real-world prompts may contain both helpful and misleading cues due to imperfect observations and ambiguous descriptions. Reliable grounding therefore requires reasoning about which evidence truly explains the target rather than blindly trusting all available information.

Based on this observation, we propose Bidirectional Evidence–Target Co-Inference (BETC), which reformulates multimodal prompt grounding as target-relative evidence reasoning. Different from directly fusing all prompt information, BETC jointly models the interaction between evidence and targets through a bidirectional process. Evidence first helps identify potential targets, while target-conditioned reasoning further determines which evidence truly supports each target. This process enables reliable grounding under partially reliable prompts by suppressing misleading evidence while preserving informative evidence.

Our contributions are summarized as follows. We first identify and study the problem of multimodal prompt grounding under partially reliable prompts, where user-provided prompts may contain conflicting evidence. Based on existing public datasets, we establish evaluation protocols that systematically analyze grounding robustness under partially reliable prompts. Furthermore, we propose Bidirectional Evidence–Target Co-Inference (BETC), which transforms multimodal grounding from direct prompt fusion into target-relative evidence reasoning through bidirectional evidence-target inference. Finally, we conduct extensive experiments under two settings: controlled attribute corruption on PACO and natural retrieval mismatch on PhraseCut. Results demonstrate that BETC substantially improves grounding robustness under conflicting prompts while maintaining competitive performance on clean inputs.

2 Related Work

2.1 Multimodal Promptable Dense Prediction

Vision-language pretraining establishes transferable image-language representations for open-world visual understanding (Radford et al., 2021; Li et al., 2021; 2022a; Ilharco et al., 2021). Dense prediction and open-vocabulary segmentation methods further extend these representations to pixel-level understanding through universal decoding, visual-semantic alignment, and language-guided localization (Cheng et al., 2022; Jain et al., 2023; Zou et al., 2023a; Li et al., 2022b; Liu et al., 2024; Liang et al., 2023; Xu et al., 2023a; 2022; Rao et al., 2022). Building on these advances, multimodal promptable grounding methods incorporate richer user guidance, including textual descriptions and visual exemplars, for open-world localization (Liu et al., 2025; Guan et al., 2025; He et al., 2024; Jiang et al., 2024; Wang et al., 2025). Exemplar-based and in-context segmentation approaches further show how reference images or visual examples can condition segmentation (Zhang et al., 2024; Liu et al., 2023; Meng et al., 2024). While these methods demonstrate the utility of multimodal and exemplar-based prompts, they primarily focus on aligning or integrating the provided guidance, without explicitly modeling whether individual cues remain informative for the current target. BETC addresses this gap by evaluating individual evidence factors relative to candidate targets under conflicting prompts.

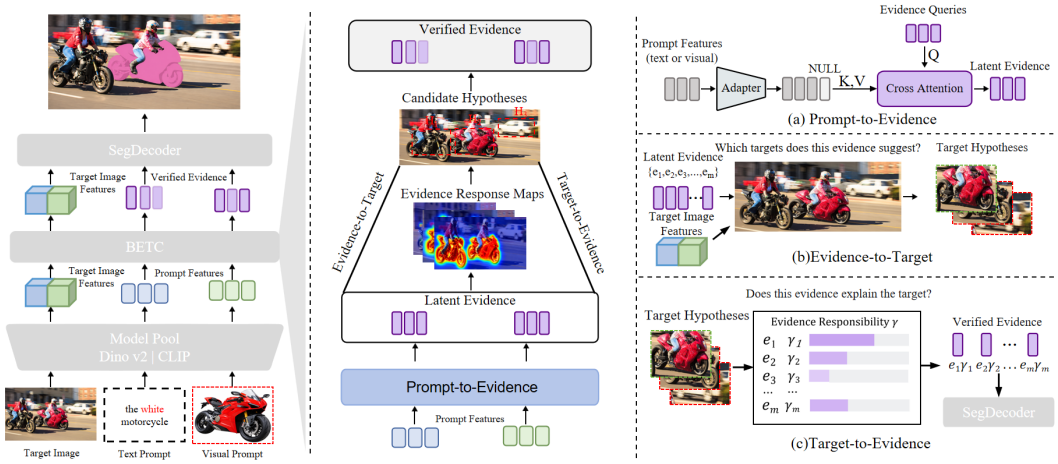

Figure 2: Overview of BETC. BETC reformulates multimodal grounding as bidirectional evidence-target reasoning, where evidence discovers target hypotheses and targets verify evidence.

2.2 Reliability-aware Multimodal Reasoning

Previous studies have shown that unreliable information sources and biased multimodal representations can affect multimodal prediction, motivating research on robust multimodal learning and trustworthy multimodal systems (Yang et al., 2024; Baltrušaitis et al., 2018; Li et al., 2023). However, existing approaches usually characterize reliability at the modality or representation level, without explicitly modeling whether a specific evidence factor explains the current target. In contrast, BETC models reliability as a conditional relationship between evidence and candidate target hypotheses: the same evidence may support one target while misleading another.

2.3 Prompt Interaction and Spatial Evidence

Prompt-target interaction has been explored through language-conditioned segmentation and multimodal grounding frameworks, which improve localization via cross-modal correspondence and spatial alignment (Wang et al., 2022; Yang et al., 2022; Wu et al., 2022; Liu et al., 2025; He et al., 2024; Zou et al., 2023b). However, prompt-target compatibility alone cannot determine whether a cue genuinely explains the target when evidence conflicts. BETC extends prompt interaction toward target-relative evidence reasoning, where evidence contribution is evaluated according to its relationship with the grounded target.

3 Method

3.1 Overview: Grounding as Evidence-Target Reasoning

BETC reformulates multimodal grounding as target-relative evidence reasoning. Instead of directly aggregating prompt features, BETC decomposes multimodal prompts into latent evidence factors and performs bidirectional inference: evidence identifies candidate targets, while target-conditioned verification estimates target-relative evidence responsibility.

Given multimodal prompts, BETC first decomposes them into latent evidence units and performs Evidence-to-Target inference to discover plausible target hypotheses through interactions between evidence and the target image. These hypotheses then provide target-specific context for Target-to-Evidence verification, which evaluates how individual evidence units contribute to explaining each target. The two directions are jointly modeled to infer target beliefs and evidence responsibilities.

The inferred evidence responsibilities are finally used to modulate the evidence representations before they are provided to the original segmentation decoder. By separating evidence discovery from evidence verification, BETC suppresses misleading cues while preserving informative evidence, en-

abling robust grounding under partially reliable multimodal prompts. All components are designed to answer a single question: whether a given evidence factor can explain the current target.

3.2 PROMPT-TO-EVIDENCE

BETC first transforms multimodal prompt representations into latent evidence factors, since different prompt components may provide different levels of support for different target hypotheses. BETC therefore decomposes prompts into multiple evidence factors that can be independently evaluated during subsequent target-relative reasoning.

Given modality-specific prompt features U_m^0, where $m \in \{t, v\}$ denotes textual and visual prompts, BETC first maps them into a shared evidence space through modality-specific adapters. To allow individual prompt factors to contribute no evidence during reasoning, a learnable null token n_m is appended to the adapted features:

$$\bar{U}_m = [\mathrm{Adapter}(U_m^0); n_m]. \tag{1}$$

BETC then employs shared evidence queries Q^{slot} to extract latent evidence factors through cross-attention:

$$E_m = \mathrm{CrossAttn}(Q^{\mathrm{slot}}, \bar{U}_m, \bar{U}_m). \tag{2}$$

The evidence factors from different modalities are concatenated as $E = [E_t; E_v]$. These extracted factors provide the evidence representations for subsequent bidirectional reasoning. Unlike raw prompt representations, these factors are designed to be independently evaluated and assigned target-relative responsibility.

3.3 EVIDENCE-TO-TARGET HYPOTHESIS DISCOVERY

Given the extracted evidence factors, BETC first performs Evidence-to-Target inference to discover plausible target hypotheses. Each evidence factor interacts with target-image features to identify spatial regions corresponding to the provided prompt evidence. Specifically, for an evidence factor e_i and target-image feature map F_I, BETC computes an evidence-specific response map:

$$R_i = \mathrm{Softmax}_{x\in\Omega}\left(\frac{(W_q^{\mathrm{disc}} e_i)(W_k^I F_I)^\top}{\sqrt{d_e}}\right) \in \mathbb{R}^{H\times W}. \tag{3}$$

Here $R_i \in \mathbb{R}^{H\times W}$ represents the spatial correspondence between evidence factor i and the target image. Different from directly using correspondence maps as predictions, BETC uses this evidence-target correspondence only to generate candidate hypotheses for subsequent verification. Based on response maps from all evidence factors, BETC extracts local peaks and merges spatially close responses to obtain a set of candidate target hypotheses $\mathcal{H} = \{H_j\}_{j=1}^K$, where K denotes the number of generated target hypotheses. Each hypothesis contains an image-space location and a target representation extracted from the target-image features around the corresponding location. These target hypotheses provide target-specific context for subsequent Target-to-Evidence verification.

3.4 TARGET-TO-EVIDENCE VERIFICATION

After discovering candidate targets, BETC evaluates which evidence factors genuinely explain each target hypothesis. The key observation is that evidence reliability is target-dependent: the same evidence may support one candidate while misleading another. Therefore, evidence reliability requires target context and cannot be determined solely from the evidence itself. Given a target hypothesis representation h_j and an evidence factor e_i, BETC computes their target-conditioned compatibility:

$$S_{ji} = \frac{(W_q^{\mathrm{ver}} h_j)^\top (W_k^E e_i)}{\sqrt{d_e}}. \tag{4}$$

The resulting matrix $S \in \mathbb{R}^{K\times N_e}$ captures relationships between candidate targets and evidence factors. Since an evidence factor may support multiple candidates, its contribution is evaluated relative to competing hypotheses. BETC computes the relative evidence gain:

$$G_{ji} = S_{ji} - \log\left[\frac{1}{K}\sum_{k=1}^{K}\exp(S_{ki})\right]. \tag{5}$$

This normalization removes candidate-independent evidence bias. A positive G_{ji} indicates that evidence factor i provides stronger relative support for candidate j compared with competing hypotheses. These target-relative evidence gains are then used for joint explanation inference.

3.5 Target-Relative Explanation Inference and Verified Evidence Readout

The target hypotheses and target–evidence relationships obtained from the two inference directions are jointly modeled through a target-relative explanation process. Instead of assigning an independent reliability score to each evidence factor, BETC introduces evidence participation as a latent variable that captures whether a specific evidence factor contributes to explaining a target hypothesis.

Let $C_{ji} \in \{0, 1\}$ denote whether evidence factor i participates in explaining target hypothesis j. By modeling the interaction between target hypotheses and latent evidence participation, BETC jointly infers which target is supported and which evidence factors contribute to its explanation. Given the relative evidence gain G_{ji} obtained from Target-to-Evidence verification, we formulate the target-relative explanation model as:

$$p(j, \mathbf{C} \mid G, \mathcal{H}) \propto \bar{\pi}_j \prod_{i=1}^{N_e} w^{C_{ji}} (1 - w)^{1 - C_{ji}} \exp(C_{ji} G_{ji}), \tag{6}$$

where $\bar{\pi}_j = 1/K$ denotes the prior probability of target hypothesis j, and w represents the base participation probability of an evidence factor. By marginalizing the latent participation variables, the same explanation model simultaneously estimates target belief and evidence responsibility:

$$q_j = \mathrm{Softmax}_j \left(\sum_i \mathrm{Softplus}(G_{ji}) \right), \qquad \gamma_i = \sum_{j=1}^{K} q_j \sigma(G_{ji}). \tag{7}$$

where q_j denotes the posterior belief of target hypothesis j, and γ_i measures the contribution of evidence factor i under the inferred target distribution. Thus, a single explanation model provides both target selection and evidence participation.

The inferred evidence responsibilities are then used to modulate the original evidence representations before segmentation:

$$\widehat{E}_i = \gamma_i E_i. \tag{8}$$

The responsibility-weighted evidence representations are fed into the original segmentation decoder through the original prompt interface. The decoder and image feature pathway remain unchanged; BETC only modifies the evidence representations through target-relative evidence reasoning.

4 Experiments

Our experiments contain two complementary evaluations: PACO isolates whether BETC can identify and suppress conflicting evidence under controlled perturbations, while PhraseCut examines whether the same mechanism remains effective under naturally occurring retrieval mismatches. Ablation and reliability controls then examine which parts of BETC account for the gains.

4.1 Experimental Setup

Datasets and prompts. We evaluate multimodal prompt reliability on two complementary benchmarks. PACO (Ramanathan et al., 2023) provides a controlled evaluation of attribute-level prompt conflicts, where textual and visual evidence can be systematically modified while preserving target identity. PhraseCut (Wu et al., 2020) evaluates the same reliability problem under natural referring expressions and retrieval-induced prompt mismatch.

In both settings, each episode contains a target image, a target mask, a textual prompt, and a visual exemplar from another image. The target image and ground-truth mask remain fixed across prompt conditions, while textual and visual evidence are modified according to each protocol.

Evaluation protocols. We hold the target image and ground-truth mask fixed and evaluate four prompt conditions: Clean, text mismatch, visual mismatch, and bilateral mismatch.

Table 1: **Robust multimodal grounding under partially reliable prompts on PACO.** Text-C, Visual-C, and Bi-C denote controlled corruption of textual prompts, visual exemplars, and both modalities, respectively.

Method	Training	Clean↑	Text-C↑	Visual-C↑	Bi-C↑	Δ↓
Existing multimodal grounding methods (clean training)						
SEEM (Zou et al., 2023b)	Clean	63.10 ± 0.24	57.90 ± 0.33	57.20 ± 0.36	55.70 ± 0.41	6.17 ± 0.37
Rex-Omni (Jiang et al., 2026)	Clean	63.45 ± 0.22	58.70 ± 0.30	58.10 ± 0.32	56.60 ± 0.36	5.65 ± 0.34
AnyRef (He et al., 2024)	Clean	62.60 ± 0.23	58.20 ± 0.31	55.40 ± 0.38	54.80 ± 0.43	6.47 ± 0.38
COSINE (Liu et al., 2025)	Clean	62.20 ± 0.25	57.50 ± 0.32	56.40 ± 0.35	55.70 ± 0.39	5.67 ± 0.36
DSV-LFS (Karimi & Poullis, 2025)	Clean	61.80 ± 0.27	56.80 ± 0.34	53.90 ± 0.42	52.70 ± 0.47	7.33 ± 0.41
MQ-Det (Xu et al., 2023b)	Clean	60.70 ± 0.29	55.80 ± 0.36	54.70 ± 0.39	53.30 ± 0.44	6.10 ± 0.40
BETC (ours)	Clean	$\mathbf{62.10 \pm 0.23}$	$\mathbf{59.70 \pm 0.28}$	$\mathbf{59.00 \pm 0.30}$	$\mathbf{58.20 \pm 0.34}$	$\mathbf{3.13 \pm 0.31}$
Controlled comparison: training strategy						
COSINE (Liu et al., 2025)	Mismatch	61.30 ± 0.26	58.50 ± 0.30	57.60 ± 0.32	56.90 ± 0.36	3.63 ± 0.34
BETC (ours)	Mismatch	$\mathbf{62.00 \pm 0.22}$	$\mathbf{61.00 \pm 0.25}$	$\mathbf{60.40 \pm 0.27}$	$\mathbf{59.80 \pm 0.29}$	$\mathbf{1.60 \pm 0.27}$

For PACO, Text-C, Visual-C, and Bi-C denote controlled corruption of textual prompts, visual exemplars, and both modalities, respectively. The corrupted prompts preserve the target category while modifying selected attributes.

For PhraseCut, Text-R, Visual-R, and Bi-R denote retrieval-induced mismatches between the target and retrieved textual prompts, visual exemplars, and both modalities, respectively.

Here, C refers to controlled corruption, while R refers to naturally occurring retrieval mismatch.

Baselines and training. COSINE (Liu et al., 2025) serves as the direct multimodal baseline. BETC retains the original image pathway and segmentation decoder while replacing prompt representation and readout with target-relative evidence reasoning. We evaluate both clean-only and mismatch-aware training settings, where mismatch-aware training exposes models to all four prompt conditions.

Metrics. We report cumulative IoU (cIoU), the total foreground intersection divided by total union over evaluation episodes, in percent. Clean denotes evaluation with the original matched multimodal prompts, while mismatch settings evaluate degraded prompts under the corresponding protocol. The degradation Δ is Clean cIoU minus the mean of the three mismatch scores; lower is better. Unless otherwise specified, all results are reported as cIoU (%) with mean±std over 3 random seeds. The Supplementary Material provides construction details, implementation details, parameter analyses, and additional diagnostics.

4.2 Robust Grounding under Partially Reliable Multimodal Prompts

Existing multimodal grounding methods typically assume that the provided textual descriptions and visual exemplars are reliable. To examine whether this assumption holds in practice, we evaluate representative multimodal grounding methods under controlled prompt corruption.

Table 1 shows that existing multimodal grounding methods, despite competitive clean performance, degrade consistently as prompt reliability decreases. For example, SEEM drops from 63.10% to 55.70% under Bi-C, and the exemplar-driven few-shot method DSV-LFS is affected most strongly by visual corruption (53.90% under Visual-C versus 56.80% under Text-C), indicating that direct evidence fusion remains vulnerable to conflicting cues.

To distinguish the effect of robustness-oriented training from the proposed evidence reasoning mechanism, we compare COSINE and BETC under clean and mismatch training. Mismatch-aware training improves COSINE robustness (Δ: 5.67% to 3.63%), indicating partial adaptation. However, clean-trained BETC already exceeds mismatch-trained COSINE under Bi-C (58.20% versus 56.90%) and attains a lower degradation (3.13% versus 3.63%). Combining BETC with mismatch training achieves the lowest degradation ($\Delta = 1.60\%$) and the best Bi-C cIoU (59.80%) while retaining competitive clean performance (62.00%). These results indicate that the robustness gains cannot be attributed to corruption training alone: even with clean training, BETC outperforms

Table 2: **Natural-language grounding under partially reliable prompts on PhraseCut.** Text-R, Visual-R, and Bi-R denote naturally occurring mismatches between the target and retrieved textual prompts, visual exemplars, and both modalities, respectively.

Method	Training	Clean↑	Text-R↑	Visual-R↑	Bi-R↑	Δ↓
COSINE (Liu et al., 2025)	Clean	54.07 ± 0.30	43.47 ± 0.38	42.65 ± 0.40	40.27 ± 0.47	11.94 ± 0.43
COSINE (Liu et al., 2025)	Mismatch	53.85 ± 0.29	45.00 ± 0.35	44.55 ± 0.37	42.30 ± 0.43	9.90 ± 0.39
BETC (ours)	Clean	54.35 ± 0.27	46.75 ± 0.32	46.20 ± 0.34	44.15 ± 0.40	8.65 ± 0.36
BETC (ours)	**Mismatch**	$\mathbf{54.20 \pm 0.26}$	$\mathbf{48.10 \pm 0.29}$	$\mathbf{47.90 \pm 0.31}$	$\mathbf{46.20 \pm 0.35}$	$\mathbf{6.80 \pm 0.31}$

Table 3: **Ablation of target-relative evidence reasoning on PACO.** Target rel. indicates whether evidence reliability is conditioned on candidate targets. All rows use mismatch-aware training.

Variant	Granularity	Target rel.	Clean↑	Text-C↑	Visual-C↑	Bi-C↑	Δ↓
COSINE	—	×	61.30 ± 0.26	58.50 ± 0.30	57.60 ± 0.32	56.90 ± 0.36	3.63 ± 0.34
+ Modality filtering	Modality-level	×	61.05 ± 0.27	58.90 ± 0.29	58.00 ± 0.31	57.30 ± 0.34	2.98 ± 0.31
+ Learned modality weights	Modality-level	×	61.20 ± 0.25	59.30 ± 0.28	58.40 ± 0.30	57.90 ± 0.33	2.67 ± 0.30
+ Evidence extraction	Evidence-level	×	61.45 ± 0.24	58.70 ± 0.29	57.80 ± 0.31	57.10 ± 0.35	3.58 ± 0.33
+ Evidence filtering	Evidence-level	×	61.40 ± 0.24	59.70 ± 0.26	58.90 ± 0.28	58.10 ± 0.31	2.50 ± 0.28
+ Learnable evidence weights	Evidence-level	×	61.35 ± 0.25	59.20 ± 0.27	58.50 ± 0.29	57.80 ± 0.32	2.85 ± 0.30
BETC (ours)	**Evidence-level**	✓	$\mathbf{62.00 \pm 0.22}$	$\mathbf{61.00 \pm 0.25}$	$\mathbf{60.40 \pm 0.27}$	$\mathbf{59.80 \pm 0.29}$	$\mathbf{1.60 \pm 0.27}$

mismatch-trained COSINE, supporting target-relative evidence responsibility as an effective mechanism for handling conflicting inputs.

4.3 Natural-Language Grounding under Partially Reliable Prompts

While PACO provides a controlled evaluation by explicitly introducing attribute-level corruption, real-world multimodal grounding systems often rely on naturally retrieved textual descriptions and visual exemplars, which may not always correspond to the target object. To evaluate whether the evidence-reliability issue persists in realistic scenarios, we further conduct experiments on PhraseCut with naturally occurring prompt mismatches. Different from PACO, where corruption types are explicitly controlled, PhraseCut evaluates whether the same reliability issue appears in natural multimodal prompt-retrieval scenarios.

As shown in Table 2, degradation under natural retrieval mismatch is considerably larger than under PACO's controlled single-attribute corruption: COSINE drops from 54.07% to 40.27% under Bi-R ($\Delta = 11.94\%$), because retrieval replaces an entire phrase or exemplar region rather than flipping one attribute. Direct multimodal fusion cannot distinguish whether retrieved evidence actually explains the target, so unreliable cues affect grounding predictions.

Introducing mismatch-aware training improves COSINE robustness (Δ: 11.94% to 9.90%), demonstrating that exposure to unreliable prompts provides partial adaptation. However, the remaining degradation suggests that learning corrupted examples alone is insufficient to determine whether a specific evidence source is relevant to the current target.

In contrast, BETC consistently reduces degradation across all mismatch settings: clean-trained BETC already lowers Δ to 8.65%, and mismatch-trained BETC reaches $\Delta = 6.80\%$ with Bi-R cIoU improving from 40.27% to 46.20%, while clean performance stays comparable (54.20% versus 54.07%). These results show that target-relative evidence reasoning remains effective under natural retrieval-induced prompt mismatch.

4.4 Ablation of Target-Relative Evidence Reasoning

We conduct ablation studies to investigate why target-relative evidence reasoning improves robustness. Instead of treating multimodal prompts as fixed reliable inputs, BETC progressively models evidence reliability at finer granularity and conditions evidence responsibility on candidate targets.

Modality-level reliability modeling is limited. Starting from mismatch-trained COSINE ($\Delta = 3.63$), hard modality filtering lowers Bi-C cIoU on Clean (61.05 versus 61.30) while only reducing Δ to 2.98, because discarding an entire modality also removes its informative evidence. Soft modality

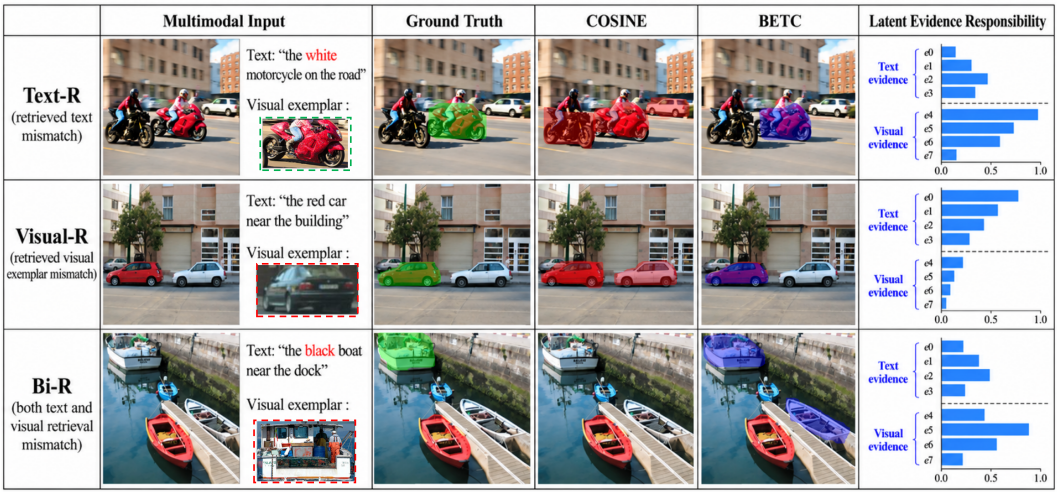

Figure 3: **Grounding under retrieved-prompt mismatch.** Examples compare COSINE and BETC under Text-R, Visual-R, and Bi-R. Responsibility traces show the contributions of text and visual evidence units alongside the predicted masks.

weights recover Clean and reduce Δ to 2.67, but a single scalar per modality cannot separate relevant from misleading units within that modality.

Evidence decomposition alone is not enough. Merely extracting evidence units with uniform weights ($\omega_i = 1$) barely changes robustness (Δ: 3.63 to 3.58; Bi-C: 56.90 to 57.10), isolating the representation contribution from the reliability operation. In contrast, evidence-level hard filtering reduces Δ to 2.50 and raises Bi-C to 58.10, showing that evidence-level modeling provides a finer granularity for reliability control than modality-level modeling; soft target-independent weights reach only 2.85, since down-weighting is less decisive than removal at this scale.

Target-relative reasoning is critical for evidence selection. Evidence-level reliability modeling without target awareness remains limited: target-independent evidence weights may identify generally informative evidence but cannot determine whether a specific evidence unit explains the current target. By introducing target-relative verification, BETC achieves the best results, raising Bi-C cIoU from 58.10 to 59.80 and reducing Δ from 2.50 to 1.60. This result supports the hypothesis that evidence responsibility emerges from the relationship between evidence and candidate targets.

4.5 Analyzing Target-Relative Evidence Responsibility

We analyze whether evidence responsibility emerges from the relationship between evidence and target context, and how such target-relative allocation changes under different prompt mismatch conditions.

Qualitative evidence adjustment. Figure 3 compares COSINE and BETC under Text-R, Visual-R, and Bi-R. The traces illustrate reduced responsibility for conflicting evidence and retained weight for target-compatible units, alongside the resulting masks. These examples suggest that evidence contribution is adjusted according to the grounding context, motivating a direct test of target dependence.

Target dependence under fixed multimodal evidence. A central hypothesis of BETC is that evidence responsibility should not be determined solely by the prompt evidence itself, but by how well that evidence explains the current target. To test this property, we construct paired evaluations with identical multimodal evidence and compatible or incompatible target contexts. If responsibility were determined only from the evidence itself, the same evidence would receive similar responsibility in both contexts. BETC instead should assign different responsibilities according to target context. For each pair, let $\bar{\gamma} = \frac{1}{8}\sum_{i=1}^{8}\gamma_i$ denote the mean responsibility across all evidence factors, and define $\Delta\bar{\gamma} = \bar{\gamma}_{\text{comp}} - \bar{\gamma}_{\text{incomp}}$. The distribution shifts toward positive values, indicating that BETC assigns higher responsibility to the same evidence when it becomes target-compatible. The remaining

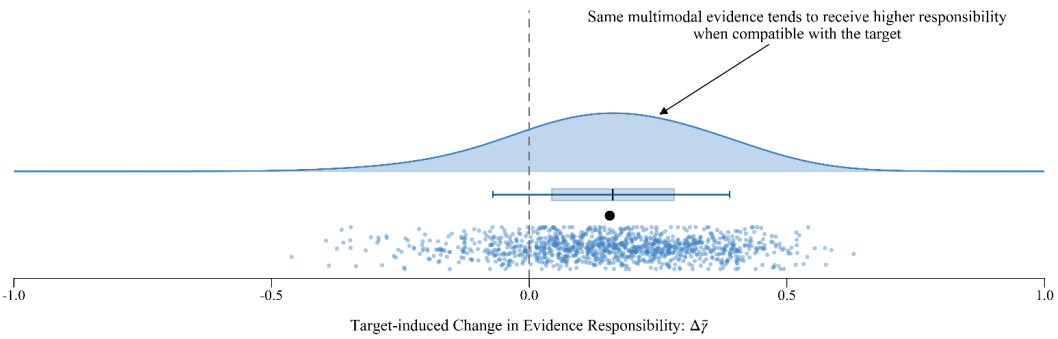

Figure 4: **Target-relative evidence responsibility under fixed multimodal evidence.** Each point represents one paired evaluation under fixed multimodal evidence. The x-axis shows the paired change in mean evidence responsibility. The half-violin shows the distribution of paired differences, points denote individual comparisons, and the black marker indicates the mean paired difference.

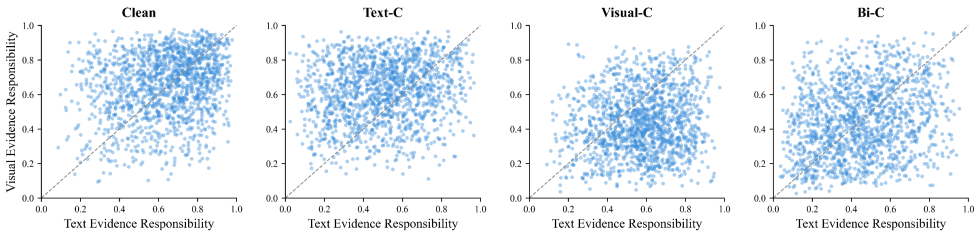

Figure 5: **Condition-wise evidence allocation under prompt mismatch.** Each point is a test episode, positioned by its mean text and visual responsibility. The distributions show how evidence allocation changes under clean prompts, text corruption, visual corruption, and bilateral corruption.

variation indicates that responsibility adjustment depends on the evidence-target relationship rather than following a fixed rule based only on target compatibility. This result provides direct diagnostic evidence that BETC models evidence responsibility relative to the target instead of assigning fixed responsibilities based only on the prompt.

Condition-wise evidence allocation. Figure 5 plots each test episode by the mean responsibility assigned to textual and visual evidence. These distributions show that BETC dynamically reallocates evidence responsibility according to prompt reliability and target relevance.

5 Conclusion

We presented BETC for multimodal grounding with partially reliable prompts. Its central idea is to assess evidence according to its relationship with candidate targets: evidence discovers target hypotheses, while target-conditioned verification determines which evidence contributes to the final prediction. Across controlled prompt perturbations on PACO and retrieval-induced mismatches on PhraseCut, BETC improves grounding robustness while maintaining competitive performance on clean prompts. Ablation studies further validate the advantage of target-relative evidence reasoning over target-independent evidence weighting and the effectiveness of jointly modeling discovery and verification.

AI Use Statement

Generative AI tools were used as auxiliary tools during the research process, including language refinement, proofreading, code assistance, debugging support, and improving the presentation of experimental analyses. The authors are fully responsible for the research ideas, methodology, experimental design, implementation decisions, results, interpretations, and conclusions presented in this paper.

REFERENCES

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Yuchen Guan, Chong Sun, Canmiao Fu, Zhipeng Huang, Chun Yuan, and Chen Li. Text-guided visual prompt dino for generic segmentation. *arXiv preprint arXiv:2508.06146*, 2025.

Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal instruction tuned llms with fine-grained visual perception. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13980–13990. IEEE, 2024.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2989–2998, 2023.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy, 2024. *arXiv preprint arXiv:2403.14610*, 2024.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25472–25483, 2026.

Amin Karimi and Charalambos Poullis. Dsv-lfs: Unifying llm-driven semantic cues with visual features for robust few-shot segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4584–4594. IEEE, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022b.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7070. IEEE, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023.

Yang Liu, Yufei Yin, Chenchen Jing, Muzhi Zhu, Hao Chen, Yuling Xi, Bo Feng, Hao Wang, Shiyu Li, and Chunhua Shen. Unified open-world segmentation with multi-modal prompts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21557–21567. IEEE, 2025.

Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M Alvarez, Zuxuan Wu, and Yu-Gang Jiang. Segic: Unleashing the emergent correspondence for in-context segmentation. In *European Conference on Computer Vision*, pp. 203–220. Springer, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7141–7151. IEEE, 2023.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18061–18070. IEEE, 2022.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anythi. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24591–24602. IEEE, 2025.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11676–11685. IEEE, 2022.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10213–10222. IEEE, 2020.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4964–4974. IEEE, 2022.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18113–18123. IEEE, 2022.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2955–2966. IEEE, 2023a.

Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. Multi-modal queried object detection in the wild. *Advances in Neural Information Processing Systems*, 36:4452–4469, 2023b.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *International Conference on Learning Representations*, volume 2024, pp. 55036–55055, 2024.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18134–18144. IEEE, 2022.

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

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15116–15127. IEEE, 2023a.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023b.

SUPPLEMENTARY MATERIAL

This supplement provides additional analyses, implementation details, and protocol clarifications supporting the central claim of BETC: reliable multimodal grounding requires reasoning about evidence relative to the target context rather than directly aggregating prompt information. All diagnostic analyses are designed to characterize the proposed mechanism beyond the main benchmark evaluation. We first describe baseline controls and implementation details, then provide the mathematical interpretation of target-relative evidence responsibility, followed by evaluation protocols and additional analyses.

We distinguish official dataset splits from constructed diagnostic populations. The latter are used only for controlled analyses and do not define alternative training or evaluation splits.

A TARGET-INDEPENDENT RELIABILITY CONTROLS

The custom controls use the BETC adapters, evidence tokenizer, output interface, and native COSINE decoder unless stated otherwise. They vary the evidence-use or inference rule. These controls isolate target-relative reasoning while keeping the evidence representation, segmentation interface, and optimization protocol consistent.

A.1 WEIGHTING SOURCES FOR TABLE 3

COSINE has no explicit reliability operation. The modality- and evidence-level controls use target-independent weights or filtering, whereas BETC uses the target-relative responsibility $\gamma_i = \sum_j q_j \sigma(G_{ji})$. Detailed target-independent unit weighting is given in Eq. (9); only BETC conditions evidence use on candidate targets.

A.2 TARGET-INDEPENDENT EVIDENCE WEIGHTING CONTROLS

Target-independent Evidence Weighting. Each evidence unit is assigned a weight from the unit itself,

$$\omega_i = \sigma(\mathrm{MLP}_{\mathrm{ind}}(e_i)), \qquad \widehat{P}_i = \omega_i P_i, \tag{9}$$

where $\mathrm{MLP}_{\mathrm{ind}}$ has no target feature, candidate feature, response map, or cross-unit normalization as input. The weighted units use the same fixed readout as Eq. (8) and are trained with the retained segmentation loss. This control evaluates target-independent evidence weighting rather than target-relative responsibility estimation.

B TARGET-RELATIVE EVIDENCE RESPONSIBILITY FORMULATION

B.1 TARGET-RELATIVE EXPLANATION MODEL

This formulation provides a probabilistic interpretation of why evidence responsibility should depend on target context. Instead of assigning fixed confidence to each prompt component, BETC models evidence participation conditioned on candidate targets. In Eq. (6), $C_{ji} \in \{0, 1\}$ indicates participation of evidence unit i in the explanation of anchor hypothesis j. The finite conditional model uses $\bar{\pi}_j = 1/K$ and Bernoulli base factors with $w = \sigma(\beta)$; the reference fixes $\beta = 0$.

For fixed j, summing over C_{ji} yields $1 - w + we^{G_{ji}}$. Define

$$\Psi_\beta(g) = \log(1 - w + we^g) \tag{10}$$

$$= \mathrm{Softplus}(g + \beta) - \mathrm{Softplus}(\beta). \tag{11}$$

Consequently $q_j \propto \bar{\pi}_j \exp[\sum_i \Psi_\beta(G_{ji})]$, which gives Eq. (7) with the uniform prior and $\beta = 0$ ($w = 1/2$). Conditional participation and its marginal are

$$p(C_{ji} = 1 \mid j, S, \mathcal{H}) = \sigma(G_{ji} + \beta), \tag{12}$$

$$p(C_{ji} = 1 \mid S, \mathcal{H}) = \sum_j q_j \sigma(G_{ji} + \beta). \tag{13}$$

These marginals give BETC its coupled interpretation: evidence contributions determine candidate belief, which in turn determines candidate-conditioned evidence participation. Unlike target-independent weighting, the estimated participation depends on the posterior distribution over candidate targets, making evidence contribution conditional on the grounded target context. Both are evaluated in closed form within one forward pass.

B.2 Analysis of Evidence Responsibility Estimation

The normalizer in each gain column is independent of j, so

$$\mathrm{softmax}_j \sum_i G_{ji} = \mathrm{softmax}_j \sum_i S_{ji}. \tag{14}$$

Thus relative-gain normalization alone does not change candidate ranking; the contribution of BETC comes from applying the participation marginalization before aggregating evidence and coupling the resulting belief to evidence responsibility. An all-zero gain column is neutral evidence, yielding $\gamma_{ji} = w$ and $\gamma_i = w$. Responsibilities remain prompt-side evidence weights, while q selects among image-space hypotheses rather than directly masking image features.

C Implementation Details

This section fixes the COSINE reference specification: feature taps, geometry, operators, native prompt readout, training, and output formation. All COSINE-based comparisons use the same backbone, optimizer, training budget, three random seeds, and paired evaluation protocol unless a row explicitly changes one of these factors. Backbone-specific interfaces are part of the reference contract rather than interchangeable tensor adapters.

C.1 Reference Backbone and Feature Taps

The multi-scale COSINE reference uses DINOv2 ViT-S/14 and OpenCLIP ConvNeXt-Base-W (Caron et al., 2021; Oquab et al., 2023). The vision-language backbones are initialized from their respective pretrained weights. Since COSINE provides no unified checkpoint for this backbone configuration, the COSINE framework and BETC components are trained under the optimization protocol described below. We adopt this smaller backbone configuration to keep all COSINE/BETC comparisons computationally matched. The following is a reference configuration, with dataset manifests bound separately to each experiment.

Input interfaces. Image preprocessing follows COSINE's original image interface. Text prompts and visual exemplars use the same prompt preprocessing pipeline as COSINE, while BETC further introduces modality-specific evidence decomposition and NULL evidence modeling.

Native DINO and CLIP neck outputs are concatenated channelwise at each scale and fused by COSINE's original merging modules. The resulting $p2, p3, p4, p5$ grids are $256^2, 128^2, 64^2, 32^2$, each with 256 channels. The complete pyramid goes to the original pixel decoder, which generates its multi-scale and mask features. Only fused $p4$ is projected for BETC reasoning:

$$F_I = \mathrm{LN}(F_{p4} W_I + b_I). \tag{15}$$

The valid spatial domain Ω contains the non-padded feature cells. The same domain is used for response normalization, candidate construction, pooling, and auxiliary supervision so that training and inference share one candidate geometry.

C.2 Evidence Construction and Operators

Writing $U_t^0 = T^0$, $U_v^0 = V^0$, input adaptation is always learned:

$$X_m = \mathrm{LN}(U_m^0 W_m^{\mathrm{in}} + b_m^{\mathrm{in}}), \tag{16}$$

including when the input is already 256-dimensional. One independent, never-masked null token per modality is appended after input normalization. The shared queries and MHA projections implement

$$\bar{X}_m = [X_m; n_m], \tag{17}$$

$$E_m = \mathrm{LN}\left[Q^{\mathrm{slot}} + \mathrm{Dropout}(\mathrm{MHA}(Q^{\mathrm{slot}}, \bar{X}_m, \bar{X}_m))\right]. \tag{18}$$

The tokenizer uses four shared slots per modality and the same evidence-space interface at training and inference.

Discovery and verification each use independent bias-free 256-dimensional projections, no dropout, and $1/\sqrt{256}$ scaling. Discovery normalizes over valid spatial cells:

$$R_i(x) = \frac{\exp z_i(x)}{\sum_{y\in\Omega} \exp z_i(y)}, \quad x \in \Omega; \qquad R_i(x) = 0 \ (x \notin \Omega). \tag{19}$$

For $x \in \Omega$, a peak must satisfy

$$R_i(x) = \max_{y\in\mathcal{N}(x)} R_i(y), \tag{20}$$

$$R_i(x) \geq .5 \max_{y\in\Omega} R_i(y), \quad R_i(x) \geq 2/|\Omega|, \tag{21}$$

where $\mathcal{N}(x)$ is its valid centered 3×3 neighborhood. Peaks are merged by non-maximum suppression and at most $K_{\max} = 20$ candidate anchors are retained; if no peak is found, the global response maximizer supplies one anchor. The prior is uniform over the retained hypotheses.

Pool $h_j = \mathrm{LN}(|\mathcal{N}(c_j)|^{-1} \sum_{x\in\mathcal{N}(c_j)} F_I(x))$. Verification uses Eq. (4) with no evidence-axis softmax. Stable gain computation is

$$G_{ji} = S_{ji} - \mathrm{logsumexp}_k(S_{ki}) + \log K. \tag{22}$$

Candidate coordinates and NMS are stop-gradient, while selected features, verification, q, γ, and readout retain gradients. Candidate anchors are generated from response maps; ground-truth masks are used for supervision and evaluation.

C.3 Native Prompt Readout and Mask Output

Responsibility-conditioned evidence representations are passed through COSINE's original prompt projection, normalization, slot aggregation, and mask-generation pipeline. The decoder and image feature pathway are unchanged; reduced-input cases expose the learned null token for the unavailable modality. We use the native COSINE final-mask rule for all reported evaluations.

C.4 Loss Functions

Auxiliary supervision. Resize/pad the binary target mask with the target geometry using nearest-neighbor, then nearest-neighbor downsample to 64^2. Let $a_i = \sum_{x\in\Omega} M_I^*(x) R_i(x)$. The per-episode discovery loss is

$$a^+ = \tau_d \left[\mathrm{logsumexp}_i(a_i/\tau_d) - \log N_e\right], \tag{23}$$

$$\mathcal{L}_{\mathrm{disc}} = -\log(a^+ + \epsilon). \tag{24}$$

This smooth coverage objective favors well-grounded evidence but is not strictly single-slot supervision and does not guarantee disentanglement. For the same valid 3×3 windows, set $u_j = |\mathcal{N}(c_j)|^{-1} \sum_{x\in\mathcal{N}(c_j)} M_I^*(x)$ and $q_j^* = \mathrm{softmax}_j(u_j/.1)$. All-zero occupancy yields uniform q^*. Assignment loss is

$$\mathcal{L}_{\mathrm{target}} = \sum_{j=1}^{K} q_j^* \log(q_j^*/q_j). \tag{25}$$

Use actual N_e and K, sum over valid candidates, and average both auxiliary losses over the batch. If downsampling eliminates the target, assign zero auxiliary loss to that episode and retain native segmentation supervision.

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{disc}}\mathcal{L}_{\mathrm{disc}} + \lambda_t \mathcal{L}_{\mathrm{target}}. \tag{26}$$

Native segmentation criterion. We follow COSINE's native segmentation criterion, matching, point-sampling, Dice terms, and auxiliary-output supervision; BETC adds only the objectives in Eq. (26). The reference uses the fixed COSINE weights and $\lambda_{\mathrm{disc}} = \lambda_t = .5$.

Table 4: BETC Training and Inference Configuration

Item	Setting
Backbones	DINOv2 ViT-S/14; OpenCLIP ConvNeXt-Base-W
Evidence / candidates	4 slots per modality, 256 dimensions; $K_{\max} = 20$
Tokenizer	4 heads; attention/output dropout .1; no FFN
Optimizer	AdamW; new 10^{-4}, pretrained trainable 10^{-5}
Schedule	50k optimizer updates; 5% linear warmup; cosine decay to zero
Batch / precision	effective batch 16; FP16 AMP, FP32 inference reductions
Clipping / seed	global gradient norm 1; 3 seeds
Prompt / output	unnormalized weighted mean; semantic score threshold .7

Table 5: Dataset split statistics of PACO-LVIS and PhraseCut. Episodes are constructed from the corresponding official source partitions and define the full-split optimization and evaluation populations.

Dataset	Official split	Episodes
PACO-LVIS	train	89,243
PACO-LVIS	validation	4,793
PACO-LVIS	test	18,410
PhraseCut	train	82,357
PhraseCut	validation	5,517
PhraseCut	test	3,489

C.5 Optimization Details

Full-adaptation reference parameter groups. DINO and CLIP foundation encoders, including the text transformer, are frozen and remain in evaluation mode. Pretrained necks, feature blenders, pixel decoder, and MFormer are trainable at 10^{-5}. New projections, queries, null tokens, and BETC parameters train at 10^{-4}. AdamW uses $(\beta_1, \beta_2) = (.9, .999)$ and $\epsilon_{\text{opt}} = 10^{-8}$; weight decay .05 excludes biases, normalization parameters, and embeddings.

The learning rate warms linearly from zero to each group's rate over the first 5% of optimizer updates, then follows cosine decay to zero. Use FP16 autocast and gradient scaling, unscale before global-norm clipping at 1, and accumulate to effective batch 16. Three seeds per configuration initialize the model and data sampling, including workers; the seed values are part of the run configuration, and results are reported as mean$\pm$std.

Clean training always uses the clean arm; mismatch-aware training instead draws one of the four arms uniformly. All matched controls are trained for the same 50K optimizer updates. Parameter fitting uses training manifests only. Checkpoint selection maximizes the mean corrupt-condition cIoU on a disjoint development manifest, breaking ties by the earlier update; PhraseCut uses val for this role and reserves test for final evaluation. The evaluation-conditioned exemplar bank is kept separate from training.

D Dataset Split Statistics and Protocol Overview

We distinguish official dataset splits from constructed prompt reliability protocol populations. Official dataset splits define the source partitions for optimization and evaluation, whereas protocol populations describe additional diagnostic or exemplar-construction subsets.

E Evaluation Protocol Details for Partially Reliable Prompts

E.1 Protocol Overview

We design two complementary protocols to evaluate multimodal grounding under partially reliable prompts. PACO provides a controlled evaluation where textual and visual evidence conflicts are explicitly introduced while preserving target identity, allowing us to isolate whether models can

identify misleading evidence. PhraseCut evaluates retrieval-induced mismatches, where retrieved prompts may not correspond to the target instance, providing a realistic assessment of evidence reliability. Together, these protocols evaluate both controlled robustness and natural prompt reliability.

We construct two complementary prompt reliability protocols based on PACO and PhraseCut. Their full-split optimization and evaluation populations are summarized in Table 5. An additional PACO controlled-conflict diagnostic subset supports paired four-condition analysis, while the PhraseCut protocol evaluates retrieval-induced mismatches in natural referring expressions.

The main paired benchmarks use immutable construction records. Target masks come from source annotations, and prompt construction does not use model predictions. The larger target-context diagnostics and full-split fine-tuning populations are specified separately in Section G; their counts are not interchangeable with the main benchmark populations.

E.2 Controlled Attribute Corruption on PACO

The additional PACO controlled-conflict diagnostic subset is designed to study evidence reliability under controlled attribute conflicts.

The PACO controlled-conflict diagnostic subset is constructed from a filtered 3,140-object pool derived from the official PACO validation partition. It is designed for controlled paired four-condition analysis, rather than to define the full PACO-LVIS protocol scale reported in Table 5. This partition specifies the provenance of the fixed diagnostic population, rather than an additional model-validation split. Because the diagnostic protocol requires controllable attribute conflicts across multiple prompt conditions, only objects satisfying all construction constraints are retained. Therefore, this population is a protocol subset rather than the complete PACO validation partition. Both targets and exemplar donors in the four-way diagnostic protocol resolve to this pool, which is used only for diagnostic episode construction; model optimization uses separate training data.

`clean_single_attr` means one selected attribute in a prompt, not one attribute on the object. For color, the target must have one dominant color matching the selected prompt attribute; for material, one groundable positive material must match. The corresponding unknown flag must be zero. The frozen prompt strings are consumed verbatim. A clean exemplar comes from a different image and matches category and the selected clean attribute. Corrupted visual exemplars retain category but change that attribute. Masks use the source PACO segmentation through COCO-compatible polygon/RLE decoding at the source image size.

The four arms share the same ordered 437 pair IDs: 323 color and 114 material pairs, covering 424 target annotations in 398 images. A text replacement belongs to the same attribute group, is outside the target's positive set, differs from the clean attribute, and is not an "other" bucket; unknown groups are excluded. Replacement and donor selection are deterministic and independent of model predictions. Bi-C uses a different wrong attribute from the text corruption for its visual donor. The four arms thus form a paired stress protocol, not a strict factorial reuse of the same two interventions.

Severity. The four text-severity records each contain the same 1,884 targets, with zero, color-only, material-only, or both attributes corrupted. The middle severity aggregates the two single-attribute interventions per target. These records specify text changes and have no visual-exemplar provenance. They define the text intervention axis only; a paired-input evaluation additionally requires an explicit fixed donor manifest.

E.3 Natural Retrieval Mismatch on PhraseCut

The PhraseCut protocol evaluates whether target-relative evidence reasoning remains effective for natural referring expressions. Unlike PACO, mismatches are introduced through retrieval-based prompt construction rather than manually specified attribute replacement.

The target subset retains human expressions with a strict single whitelist attribute and no relation annotation. Normalization lowercases, strips ends, and collapses whitespace; matching is exact after normalization, with no category or attribute synonym merging. The color vocabulary is {black, white, red, blue, green, brown, yellow, orange, pink, purple, gray, grey, beige, tan}. The material vocabulary is {wood, wooden, metal, plastic, glass, concrete, brick, stone, leather, steel, iron, ce-

Table 6: Reference PhraseCut Loader Join by Shared Pair ID

Arm	Text	Visual region
Clean	target phrase	Clean donor
Text-R	retrieved phrase	same Clean donor
Visual-R	target phrase	Visual-R donor
Bi-R	retrieved phrase	Visual-R donor

Table 7: Constructed prompt reliability protocol populations. These populations support controlled diagnostics or visual exemplar construction and are distinct from the full-split statistics in Table 5.

Protocol population	Purpose	Statistics
PACO controlled-conflict diagnostic subset	Diagnostic evaluation	437 paired cases
PACO text severity subset	Diagnostic analysis	1,884 targets
PhraseCut donor bank	Visual exemplar retrieval	87,605 records

ramic, paper, fabric, cloth, rubber, wicker, cardboard, marble, wool, cotton, denim}. This restricted scope does not cover general relation reasoning.

We use a closed training donor bank and construct paired episodes from the source train, validation, and test partitions under the protocol summarized in Table 7. The donor bank and paired target episodes describe different protocol components: the donor bank provides candidate visual exemplars, whereas paired target episodes define grounding samples. All visual donors are sampled from the training donor bank, while target episodes follow their corresponding official partitions. Their sizes therefore should not be interpreted as alternative dataset splits. Source train episodes are used for optimization, while frozen validation and test episodes define reproducible evaluation settings. Validation combines `refer_val` and `refer_miniv`.

Candidate donors must come from a different image, have the same exact normalized object name, and belong to the training bank. Clean visual retrieval matches the selected attribute. Mismatched donors have a different attribute in the same group. Semantic-equivalent attribute pairs are excluded (e.g., gray/grey and wood/wooden) to avoid direct semantic overlap between retrieved and target attributes; the full exclusion list is fixed in the evaluation protocol.

Donor selection is deterministic and independent of model predictions. Visual mismatch prefers a donor different from the selected text donor when one is available; a documented fallback handles cases without an alternative.

The Text-R JSON omits a visual field. The reference loader joins the Clean visual donor by `pair_id`, as in Table 6; the target region is used only for supervision and evaluation, not as an inference input. The same-attribute retrieved-text control also fixes this Clean donor. Text replacement preserves the entire retrieved human phrase, so words beyond the designated attribute may differ. The same-attribute source control tests that broader phrase-source change.

Mask convention. We rasterize all valid source-resolution PhraseCut polygons and union them into a binary target mask; invalid polygons are excluded during manifest auditing. Bounding boxes use $[x, y, w, h]$. Exemplar crop/mask transforms follow Section C.1.

E.4 Evaluation Definitions and Controls

Evaluate binary masks at original image resolution. For episode n, let I_n and U_n be intersection and union counts on pixels valid for evaluation. Source ignore pixels, when supplied, are excluded from both; padding never enters evaluation. Source-valid ground truth must be nonempty; malformed or empty masks are excluded through the manifest audit, not model-dependent filtering. The reference metrics are

$$\mathrm{cIoU} = 100\frac{\sum_n I_n}{\sum_n U_n}, \qquad \mathrm{mIoU} = \frac{100}{N}\sum_n \frac{I_n}{U_n}. \tag{27}$$

PACO primary aggregation uses all 437 pairs. A target-macro IoU first averages episode IoUs within each unique target and then averages across 424 targets; it is distinct from cumulative intersection-over-union. Counts are fixed metadata, not seed-wise measurements.

Table 8: **Prompt contribution analysis under partial modality availability on PACO and PhraseCut** (cIoU, %, mean±std over 3 seeds). No Prompt denotes grounding without auxiliary prompts, while Text-only and Visual-only use only the clean textual prompt or visual exemplar, respectively. All results use clean-trained checkpoints to isolate the contribution of correctly aligned textual and visual evidence under clean conditions.

	PACO			PhraseCut		
Method	No Prompt↑	Text-only↑	Visual-only↑	No Prompt↑	Text-only↑	Visual-only↑
COSINE	$\mathbf{53.80 \pm 0.34}$	$\mathbf{56.10 \pm 0.28}$	$\mathbf{56.20 \pm 0.30}$	$\mathbf{40.51 \pm 0.38}$	$\mathbf{43.60 \pm 0.33}$	$\mathbf{43.10 \pm 0.35}$
BETC	$\mathbf{54.10 \pm 0.32}$	$\mathbf{56.50 \pm 0.26}$	$\mathbf{56.53 \pm 0.28}$	$\mathbf{43.56 \pm 0.36}$	$\mathbf{44.00 \pm 0.31}$	$\mathbf{44.60 \pm 0.33}$

Each configuration is evaluated using 3 seeds; we report the mean and standard deviation. The clean score minus the mean of the three corrupt-arm scores is reported as Δ, also as mean±std over seeds.

F Additional Diagnostics of Target-Relative Reasoning

The additional results examine internal controls, supervision, prompt semantics, evidence representations, inference diagnostics, and system cost. Each table retains its stated evaluation population. Section G separately specifies the target-context diagnostic protocol.

F.1 Evidence Helpfulness Under Reliable Conditions

Motivation. Multimodal prompts provide additional evidence beyond the target image. However, whether such evidence improves grounding depends on its consistency with the target context. We first verify that correctly aligned textual and visual prompts provide useful evidence under clean conditions. This establishes that BETC does not aim to discard auxiliary prompts, but to determine when such evidence should contribute to grounding.

Table 8 shows that correctly aligned textual and visual prompts provide useful evidence under clean conditions. Text-only and Visual-only inputs improve grounding relative to the No Prompt setting on both PACO and PhraseCut, confirming that BETC does not aim to discard auxiliary prompts, but to determine when such evidence should contribute to grounding.

However, the benefit of additional evidence depends on its reliability. In the mismatch evaluations of Tables 1–2, textual and visual prompts inconsistent with the target can become misleading evidence. In the natural-language evaluation, the bilateral mismatch condition performs below the No Prompt setting, indicating that inconsistent multimodal prompts can actively degrade grounding performance rather than merely fail to provide useful information.

These observations motivate target-relative evidence reasoning. The key challenge is to determine whether each evidence source explains the current target, allowing informative evidence to be preserved while misleading evidence receives lower responsibility.

F.2 Evidence Capacity and NULL Evidence Modeling

Motivation. BETC introduces several components to perform target-relative evidence reasoning, including latent evidence decomposition, candidate target discovery, and auxiliary supervision for guiding evidence–target inference. We analyze the sensitivity of three key hyperparameters: the number of evidence slots per modality, the maximum number of candidate target hypotheses, and the weights of auxiliary objectives.

These parameters control different stages of BETC reasoning. Evidence slots determine the granularity of evidence decomposition, $K_{\max}$ controls the diversity of candidate targets considered during verification, and auxiliary loss weights regulate the strength of intermediate reasoning supervision. This analysis examines whether the observed robustness improvements depend on carefully tuned configurations.

The results show that BETC remains stable across different parameter settings, while the default configuration achieves a strong overall trade-off.

Table 9: Sensitivity to core BETC hyperparameters on PACO. Each sweep varies one factor while retaining the reference training and evaluation protocol. Values are cIoU (%, mean±std over 3 seeds).

Parameter	Setting	Clean↑	Text-C↑	Visual-C↑	Bi-C↑
Slots / modality	1	61.21 ± 0.23	59.31 ± 0.27	58.43 ± 0.31	57.91 ± 0.38
	2	61.90 ± 0.24	60.30 ± 0.27	59.60 ± 0.29	59.10 ± 0.32
	4	$\mathbf{62.00 \pm 0.22}$	$\mathbf{61.00 \pm 0.25}$	$\mathbf{60.40 \pm 0.27}$	$\mathbf{59.80 \pm 0.29}$
	8	61.95 ± 0.23	60.90 ± 0.26	60.30 ± 0.28	59.60 ± 0.30
$K_{\max}$	5	61.80 ± 0.25	60.20 ± 0.27	59.40 ± 0.30	58.90 ± 0.33
	10	61.95 ± 0.23	60.85 ± 0.26	60.20 ± 0.28	59.60 ± 0.30
	20	$\mathbf{62.00 \pm 0.22}$	$\mathbf{61.00 \pm 0.25}$	$\mathbf{60.40 \pm 0.27}$	59.80 ± 0.29
	40	62.02 ± 0.22	61.05 ± 0.25	60.45 ± 0.27	$\mathbf{59.90 \pm 0.29}$
$\lambda_{\text{disc}} = \lambda_t$	0.25	61.85 ± 0.24	60.45 ± 0.27	59.80 ± 0.29	59.10 ± 0.32
	0.5	$\mathbf{62.00 \pm 0.22}$	$\mathbf{61.00 \pm 0.25}$	$\mathbf{60.40 \pm 0.27}$	$\mathbf{59.80 \pm 0.29}$
	1.0	61.95 ± 0.23	60.90 ± 0.26	60.25 ± 0.28	59.70 ± 0.30

Table 10: Effect of the learned NULL evidence token on PACO. *w/o NULL* removes the learned NULL evidence token. Both rows use matched training and evaluation settings; lower $\Delta\downarrow$ is better. cIoU (%, mean±std over 3 seeds).

Variant	Clean	Text-C	Visual-C	Bi-C	$\Delta\downarrow$
BETC w/o NULL	61.90 ± 0.24	60.00 ± 0.27	59.30 ± 0.28	58.60 ± 0.31	2.60 ± 0.29
BETC	$\mathbf{62.00 \pm 0.22}$	$\mathbf{61.00 \pm 0.25}$	$\mathbf{60.40 \pm 0.27}$	$\mathbf{59.80 \pm 0.29}$	$\mathbf{1.60 \pm 0.27}$

For evidence slots, increasing the number from 1 to 4 improves robustness under corrupted prompts, indicating that decomposing prompts into multiple evidence factors helps separate informative and misleading cues. Increasing the number to 8 provides limited additional gains, suggesting that excessive decomposition introduces redundant evidence factors without improving target-relative reasoning.

For candidate hypotheses, performance improves as $K_{\max}$ increases from 5 to 20, indicating that maintaining sufficient candidate diversity helps BETC discover plausible target hypotheses before evidence verification. Further increasing $K_{\max}$ to 40 yields marginal improvement, showing that a moderate candidate set is sufficient for effective verification while avoiding unnecessary candidate evaluation overhead.

For auxiliary supervision weights, $\lambda_{\text{disc}} = \lambda_t = 0.5$ achieves the best overall results. Lower weights provide less guidance for evidence–target inference, while larger weights do not further improve performance, indicating that intermediate reasoning supervision should complement rather than dominate the segmentation objective.

Overall, the sensitivity analysis indicates that BETC is not dependent on a narrow parameter choice and that its robustness is consistently associated with the proposed evidence–target reasoning framework. **Motivation.** BETC assumes that multimodal prompts may contain both target-related and irrelevant evidence. Therefore, the model should not be forced to utilize every extracted evidence factor during grounding. To enable this behavior, BETC introduces a learnable NULL evidence token, which provides an explicit option for evidence factors that cannot explain the current target.

We compare BETC with and without NULL evidence modeling to evaluate whether this non-explanatory state is necessary for robust grounding under partially reliable prompts.

The results indicate the importance of NULL evidence modeling for reliable multimodal grounding. Removing the NULL token consistently degrades performance under prompt mismatch conditions, indicating that forcing all evidence factors to participate in grounding can introduce unreliable information.

With NULL evidence modeling, BETC can assign low responsibility to evidence factors that fail to explain the current target, while preserving informative evidence factors. This behavior is particularly important under partially reliable prompts, where some textual descriptions or visual exemplars may be inconsistent with the target.

Table 11: Computational overhead of BETC compared with COSINE under the same backbone and evaluation setting. Total and trainable parameter counts are reported in millions; latency is reported as mean±std over three runs, while peak memory is reported as the mean.

Method	Total Params (M)	Trainable Params (M)	Latency / img (ms)	Peak Memory (GB)
COSINE	231.46	30.02	259.42 ± 1.68	2.81
BETC	232.42	30.97	263.23 ± 6.45	3.33
BETC − COSINE	+0.96	+0.95	+3.81	+0.52

These results support the design principle of BETC: reliable grounding requires not only identifying useful evidence, but also recognizing when a piece of evidence should not contribute to the prediction.

F.3 Computational Cost

Motivation. BETC introduces evidence decomposition and target-relative verification modules, which may introduce additional overhead compared with direct multimodal fusion. We therefore analyze the computational cost of BETC to examine whether its robustness improvements are achieved through excessive model complexity. We report total and trainable parameters, inference latency, and peak memory relative to COSINE under the same backbone and evaluation setting.

The computational analysis shows that BETC achieves improved robustness with only moderate additional overhead. Compared with COSINE, BETC introduces additional parameters mainly from evidence decomposition and target-relative reasoning modules, while keeping the backbone, image feature pathway, and segmentation decoder unchanged.

Although BETC performs additional evidence–target interactions during inference, the overall increase remains limited because the reasoning process operates on compact evidence factors and a small number of candidate target hypotheses rather than dense image-level features.

These results suggest that the performance gains of BETC are not simply caused by increased model capacity or computational resources. Instead, the improvements arise from explicitly modeling target-relative evidence responsibility while maintaining an efficient inference pipeline.

G Additional Analysis of Evidence Responsibility

We provide additional construction details and controls for the target-relative evidence responsibility analysis presented in the main paper. These details supplement the main paired benchmarks without introducing a separate experimental claim.

G.1 Paired Diagnostic Construction

Paired inputs and source integrity. Each pair fixes the target image and ground-truth mask across Clean, text-mismatch, visual-mismatch, and bilateral-mismatch conditions. The text-mismatch arm retains the Clean visual exemplar, and the bilateral arm reuses exactly the text and visual interventions of the two single-mismatch arms. Target masks are used for supervision and evaluation; visual prompts come from exemplar images.

PhraseCut construction. The relation-free, single-whitelisted-attribute selection retains 97,164 quality-checked tasks. We construct a deterministic held-out exemplar partition from official training images and exclude it from both fine-tuning targets and prompt inputs, preventing overlap between training and evaluation prompt sources. Main train, validation, and test manifests contain 73,914, 5,485, and 3,474 pairs, respectively. Validation merges the official val and miniv annotations. Exact normalized object names, attribute groups, compatibility exclusions, and different-image constraints define donor eligibility. A deterministic target-dependent sampler prefers less reused candidates; it does not assign an image a fixed reliable or unreliable role. Three validation/test donor seeds share the same targets; the training manifest uses donor seed zero. These are donor-construction seeds, distinct from model-training seeds. Retrieved text preserves the complete source phrase. This selection tests natural attribute-bearing phrases, not unrestricted relational language.

PACO diagnostic populations. The PACO target-context protocol uses raw PACO validation object annotations and contains 4,170 pairs over 2,755 targets in 1,906 images. Clean and incorrect exemplars retain the object category and differ in the selected color or material under positive-attribute exclusions. Full-source same-image, same-category annotations are consulted for ambiguity filtering under the source annotation constraints. A separate 3,769-record stress subset uses distinct incorrect attribute values. Fixed-visual severity includes 1,472 targets: a single exemplar matching both color and material is held constant while text corrupts neither attribute, color alone, material alone, or both. The severity levels are $0, 1/2, 1/2, 1$; the two single-attribute conditions are averaged within each target before aggregation. This fixes a partial-text-corruption axis with valid visual evidence; its evidence-utilization controls are specified in Section G.4. It is a different population from the original 1,884-target text-only intervention records. This validation-derived PACO population is an evaluation protocol, distinct from the full official-split fine-tuning release; PhraseCut-trained checkpoints can be evaluated across datasets without selecting hyperparameters on PACO targets.

G.2 Evidence Responsibility Controls

Method versus training exposure. For a controlled comparison, the reference design crosses COSINE/BETC with Clean-only/mismatch-aware fine-tuning, separately on each dataset. It fixes the DINOv2-S/14 and ConvNeXt-Base-W pool, training targets, initialization policy, update budget, trainable native-decoder subset, and checkpoint-selection rule. The model pool remains frozen. Clean-only uses the Clean arm; mismatch-aware chooses an arm by a deterministic hash of training seed, epoch, and pair ID. Every checkpoint is evaluated on all four arms. The method contrast is BETC versus COSINE at the same training exposure; the augmentation contrast is mismatch-aware versus Clean-only within a method.

Representation-matched training controls. Representation-matched training controls ensure that BETC comparisons isolate responsibility estimation.

Candidate and operator checks. Candidate coverage asks whether a target-overlapping anchor exists; candidate belief asks how mass is assigned within that set; segmentation evaluates the decoder output. They are separate outcomes because q is not a spatial mask. At fixed candidates, $\mathrm{Softmax}_j(\sum_i G_{ji}) = \mathrm{Softmax}_j(\sum_i S_{ji})$: the normalization terms removed from each evidence column are constant across candidates. These two linear sums therefore provide an operator-level equivalence analysis.

G.3 Target-context Analysis

Fixed multimodal evidence, changing target context. The diagnostic evaluates whether BETC's evidence responsibility changes with the target context while holding the multimodal evidence fixed. Each paired comparison uses the same textual and visual evidence in two episodes, one with a target compatible with the evidence and the other with an incompatible target.

For each episode, BETC produces eight evidence responsibilities $\{\gamma_i\}_{i=1}^{8}$, corresponding to the textual and visual evidence factors. We summarize them as

$$\bar{\gamma} = \frac{1}{8}\sum_{i=1}^{8}\gamma_i,$$

and report the paired difference

$$\Delta\bar{\gamma} = \bar{\gamma}_{\mathrm{comp}} - \bar{\gamma}_{\mathrm{incomp}}.$$

The paired diagnostic holds the multimodal evidence fixed across target contexts. Target masks are used only for supervision and evaluation, and are not provided during inference. Paired statistics are grouped by target image.

G.4 Additional Protocol Clarification

The fixed-visual PACO diagnostic population contains 1,472 targets with one exemplar matching both color and material. It defines fixed visual evidence with four text conditions: no attribute corruption, color-only corruption, material-only corruption, and corruption of both attributes. The

donor and target remain fixed across these conditions. These records are distinct from the 1,884-target text-intervention profile.

Aggregation and reproducibility. Evaluation uses cIoU, computed as total intersection divided by total union. Final full-split performance uses the corresponding official test manifests.