# Robust Egocentric Referring Video Object Segmentation via Dual-Modal Causal Intervention

**Haijing Liu,** **Zhiyuan Song,** **Hefeng Wu**[*] **Tao Pu,** **Keze Wang,** **Liang Lin**[*]

Sun Yat-sen University, Guangzhou 510006, China

liuhj66@mail2.sysu.edu.cn, songzhy29@mail2.sysu.edu.cn, wuhefeng@gmail.com,
putao3@mail2.sysu.edu.cn, kezewang@gmail.com, linliang@ieee.org

## Abstract

Egocentric Referring Video Object Segmentation (Ego-RVOS) aims to segment the specific object actively involved in a human action, as described by a language query, within first-person videos. This task is critical for understanding egocentric human behavior. However, achieving such segmentation robustly is challenging due to ambiguities inherent in egocentric videos and biases present in training data. Consequently, existing methods often struggle, learning spurious correlations from skewed object-action pairings in datasets and fundamental visual confounding factors of the egocentric perspective, such as rapid motion and frequent occlusions. To address these limitations, we introduce **C**ausal **E**go-**RE**ferring **S**egmentation (**CERES**), a plug-in causal framework that adapts strong, pre-trained RVOS backbones to the egocentric domain. CERES implements dual-modal causal intervention: applying backdoor adjustment principles to counteract language representation biases learned from dataset statistics, and leveraging front-door adjustment concepts to address visual confounding by intelligently integrating semantic visual features with geometric depth information guided by causal principles, creating representations more robust to egocentric distortions. Extensive experiments demonstrate that CERES achieves state-of-the-art performance on Ego-RVOS benchmarks, highlighting the potential of applying causal reasoning to build more reliable models for broader egocentric video understanding.

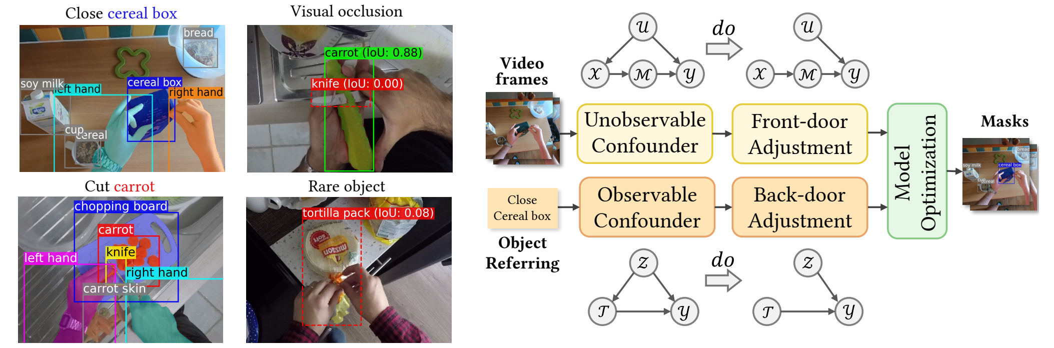
## 1 Introduction

Egocentric vision, capturing the world from a first-person perspective, offers invaluable data for understanding human interaction and behavior. Within this domain, Egocentric Referring Video Object Segmentation (Ego-RVOS) [36] presents a key task: segmenting the specific object actively involved in a human action, as identified by a natural language query combining object and action descriptions (Figure 1(a)). Successfully addressing Ego-RVOS paves the way for machines to develop a deeper comprehension of dynamic scenes, integrating visual perception, language understanding, and temporal reasoning. Prior work, such as ActionVOS [36], established settings for this task, notably utilizing action descriptions alongside object names and sometimes adapting pre-trained models with specialized loss functions to focus on active objects.

However, developing robust Ego-RVOS models faces significant hurdles. Current approaches often struggle because they learn spurious correlations rather than genuine cause-and-effect relationships [18, 60, 50, 6, 27]. These spurious correlations stem from two main sources, as illustrated by typical failure cases in Figure 1(b). First, dataset biases often exist where certain object categories frequently co-occur with specific actions, leading models to rely on these statistical shortcuts instead

[*] Corresponding authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Example of Ego-RVOS (b) Failures due to bias (c) Overview of the CERES causal framework

Figure 1: Motivation and overview of the CERES for addressing biases. (a) Ego-RVOS needs to segment the objects related to action (positive objects, colored) instead of objects unrelated to action (negative objects, gray). (b) Example failure cases of baseline [36] because of visual occlusion and rare objects outside the training set. (c) Our CERES from text and visual modal performs causal intervention to achieve robust Ego-RVOS.

of truly grounding the language query [5, 40, 55]. Second, the inherent nature of egocentric video introduces fundamental visual confounding factors: rapid camera movements, frequent hand-object occlusions, and perspective distortions create complex visual patterns that can mislead models, particularly given the domain shift from typical third-person pre-training data [46, 19, 44, 12, 7, 56]. This reliance on superficial cues renders models brittle and unreliable.

To overcome these challenges and foster robust segmentation, we propose **C**ausal **E**go-**RE**ferring **S**egmentation (CERES), a plug-in causal framework that adapts strong, pre-trained RVOS backbones to the egocentric domain by employing dual-modal causal intervention.

Instead of merely learning correlations, CERES aims to identify and model the underlying causal pathways from the dual-modal inputs (vision $\mathcal{X}$, text $\mathcal{T}$) to the segmentation output ($\mathcal{Y}$), intervening to remove confounding influences. As illustrated in Figure 1(c), we conceptualize the Ego-RVOS process using a causal graph and outline the CERES framework. We identify two primary confounding issues: (1) For the **observable language bias**, stemming from dataset statistics (confounder $\mathcal{Z}$), CERES applies principles inspired by backdoor adjustment [38, 39]. This aims to block the spurious path $\mathcal{T} \leftarrow \mathcal{Z} \rightarrow \mathcal{Y}$ and estimate the direct causal effect of the text query $\mathcal{T}$ on the output $\mathcal{Y}$, $P(\mathcal{Y} \mid \mathrm{do}(\mathcal{T}))$. (2) For the **unobservable visual confounding**, originating from inherent egocentric factors ($\mathcal{U}$), CERES utilizes principles based on **front-door adjustment** [39]. This requires identifying a mediator variable $\mathcal{M}$ that captures the causal effect flowing from vision $\mathcal{X}$ to the output $\mathcal{Y}$. The goal is to estimate the causal effect of the visual input $\mathcal{X}$ on $\mathcal{Y}$, $P(\mathcal{Y} \mid \mathrm{do}(\mathcal{X}))$, by adjusting for the effect mediated through $\mathcal{M}$ while blocking the confounding path $\mathcal{X} \leftarrow \mathcal{U} \rightarrow \mathcal{Y}$.

Implementing the front-door adjustment necessitates a carefully chosen mediator $\mathcal{M}$. Egocentric visual features $\mathcal{X}$ are susceptible to confounding ($\mathcal{U}$) through factors like rapid motion and occlusion. A purely visual mediator risks inheriting this confounding, potentially violating front-door requirements. We hypothesize that incorporating geometric structure can yield a more robust mediator, less sensitive to $\mathcal{U}$. Therefore, we propose a mediator $\mathcal{M}$ integrating semantic visual knowledge ($\mathcal{M}_v$) with geometric depth information ($\mathcal{M}_d$), both derived from $\mathcal{X}$. Leveraging depth cues provides robustness against visual distortions inherent in $\mathcal{U}$, aiming to better isolate the back-door path $\mathcal{X} \rightarrow \mathcal{M} \rightarrow \mathcal{Y}$. CERES employs attention [60] to realize this vision-depth fusion and implement the necessary causal adjustments within an end-to-end framework. The main contributions of this work are:

- We propose CERES, a novel framework applying causal inference principles to tackle key robustness challenges in Ego-RVOS.
- We employ backdoor adjustment concepts to mitigate language biases arising from spurious correlations in object-action dataset statistics.
- We utilize front-door adjustment concepts, implemented via a novel vision-depth mediator, to address fundamental visual confounding inherent in the egocentric perspective.
- Extensive experiments across diverse backbones demonstrate that CERES achieves state-of-the-art performance on VISOR, VOST and VSCOS datasets, significantly improving robustness against both linguistic and visual biases.

## 2 Related Work

- **Referring Video Object Segmentation (RVOS)** Referring Video Object Segmentation (RVOS) aims to segment the object referred to by a natural language expression throughout a video. Existing RVOS tasks [17, 22, 34, 43, 58] are typically constructed by annotating referring expressions on existing video segmentation benchmarks. These expressions often describe static attributes of a single target object. The recent MeViS dataset [13] introduces motion-based language descriptions for video object segmentation. Various methods have been proposed for RVOS [3, 14, 8, 54, 52, 24]. For instance, SLVP [32] extends RVOS to the VISOR dataset [12], while ActionVOS [36] incorporates action narrations to segment active objects in egocentric videos. Despite these advances, most approaches overlook critical challenges such as occlusion and label bias in referring expressions. In this work, we propose a causal inference-based framework to address these issues, enabling more robust and generalizable RVOS models.

- **Causal Inference** Causal inference has become an increasingly popular tool for uncovering task causality [39], and has been widely integrated into deep learning systems, especially in vision-language tasks such as image recognition [51, 53, 66, 65, 33], image captioning [61, 25], and visual question answering [23, 60]. A common approach is to apply adjustment techniques to mitigate the influence of confounding variables, with some studies exploring counterfactual reasoning [37, 1, 35]. In this work, we focus on intervention-based methods due to their practicality. However, most existing causal learning frameworks are limited to relatively simple tasks and rarely consider complex embodied settings like RVOS. Moreover, current approaches typically apply either back-door [25, 50, 64, 68] or front-door [28, 60, 61] adjustments independently across modalities, failing to account for both observable and unobservable confounders in a unified manner. Unlike prior works such as GOAT [49], which tackles confounders in vision, language, and action history, we propose the first causal framework tailored for RVOS. Specifically, we introduce a novel front-door adjustment that integrates depth information and adjacent frames features to refine segmentation decisions under occlusion. Meanwhile, we design a back-door blocking strategy to statistically correct biases in referring expressions and action labels. Our approach effectively addresses both visual and linguistic confounding effects, leading to more robust and generalizable Ego-RVOS models.

## 3 Preliminary

**Task Formulation.** The Egocentric Referring Video Object Segmentation (Ego-RVOS) task [36] requires segmenting a specific object instance involved in an action within a first-person video. The input consists of an egocentric video clip, represented as a sequence of frames $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{t}$, where $x_t \in \mathbb{R}^{H\times W\times 3}$ is the RGB frame realization at time $t$. Alongside the video, a language query $\mathbf{T}$, realized as $\mathbf{t}_{\text{txt}}$, provides both the object category name and a description of the action (e.g., "knife used to cut carrot").

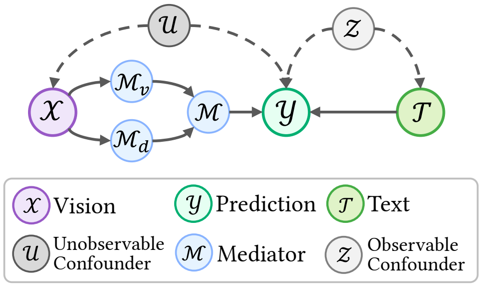


Figure 2: The proposed SCM for Ego-RVOS. (Dashed lines indicate confounding paths; solid lines indicate causal paths.)

The objective is to predict a sequence of binary segmentation masks $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^{t}$, where each realization $y_t \in \{0,1\}^{H\times W}$ precisely delineates the pixels belonging to the object instance specified by the query $t$ and actively participating in the described action within frame $x_t$. This task demands robust integration of visual perception ($\mathcal{X}$), language understanding ($\mathcal{T}$), and reasoning about object-action relationships over time.

**Structural Causal Model of Ego-RVOS.** To systematically address the biases inherent in Ego-RVOS, we formulate the task using a Structural Causal Model (SCM) [38], as depicted in Figure 2. This model posits that the visual input $\mathcal{X}$ and text query $\mathcal{T}$ are the direct causes of the segmentation output $\mathcal{Y}$. However, this ideal relationship is often confounded in practice.

We identify two primary confounders. First, an unobserved confounder $\mathcal{U}$ encapsulates intrinsic egocentric visual characteristics (e.g., rapid motion, occlusions) [44]. $\mathcal{U}$ affects both the visual input $\mathcal{X}$ and the output $\mathcal{Y}$, creating a spurious backdoor path $\mathcal{X} \leftarrow \mathcal{U} \rightarrow \mathcal{Y}$. Second, an observable confounder $\mathcal{Z}$ represents dataset-level statistical biases, such as skewed object-action co-occurrences [40]. $\mathcal{Z}$

influences both the text queries $\mathcal{T}$ and the labels $\mathcal{Y}$, forming another backdoor path $\mathcal{T} \leftarrow \mathcal{Z} \rightarrow \mathcal{Y}$. These backdoor paths lead models to learn superficial correlations rather than true causal relationships.

To mitigate the visual confounding from $\mathcal{U}$, we employ the front-door criterion. This involves an intermediate mediator $\mathcal{M}$ that captures the causal effect from $\mathcal{X}$ to $\mathcal{Y}$. Recognizing that purely visual knowledge might still be tainted by $\mathcal{U}$, we propose a more robust mediation strategy. As shown in Figure 2, we conceptualize the visual information $\mathcal{X}$ as giving rise to distinct semantic knowledge $\mathcal{M}_v$ (what objects are present) and geometric depth knowledge $\mathcal{M}_d$ (their spatial layout and structure). We hypothesize that $\mathcal{M}_v$ forms the primary pathway to an intermediate representation $\mathcal{M}$ ($\mathcal{X} \rightarrow \mathcal{M}_v \rightarrow \mathcal{M} \rightarrow \mathcal{Y}$), while $\mathcal{M}_d$ offers a complementary, potentially more robust, pathway influencing $\mathcal{Y}$ ($\mathcal{X} \rightarrow \mathcal{M}_d \rightarrow \mathcal{Y}$). This decomposition aims to leverage the stability of geometric cues ($\mathcal{M}_d$) to buttress the semantic interpretation ($\mathcal{M}_v$), leading to a mediator $\mathcal{M}$ (or a combined influence on $\mathcal{Y}$) that is less susceptible to the distortions introduced by $\mathcal{U}$. The specific mechanisms for realizing these causal adjustments will be detailed in our method section.

# 4 Methodology

The CERES (Causal Egocentric Referring-Segmentation) framework implements distinct causal adjustment strategies to address the language and visual biases inherent in Ego-RVOS, as outlined in our Structural Causal Model (SCM, Figure 2). Specifically, to counteract:

• **Language Bias**: Stemming from the observable confounder $\mathcal{Z}$ in the $\mathcal{T} \leftarrow \mathcal{Z} \rightarrow \mathcal{Y}$ pathway, CERES applies back-door adjustment.

• **Visual Bias**: Originating from the unobserved confounder $\mathcal{U}$ affecting the $\mathcal{X} \leftarrow \mathcal{U} \rightarrow \mathcal{Y}$ pathway, CERES employs front-door adjustment. This is operationalized using a mediator $\mathcal{M}$ which is carefully constructed from semantic visual features ($\mathcal{M}_v$) and geometric depth features ($\mathcal{M}_d$) derived from the visual input $\mathcal{X}$.

The subsequent sections detail the specific formulations for these back-door and front-door adjustments and describe their neural network implementations within CERES.

## 4.1 Language De-biasing via Back-Door Adjustment

Dataset statistics often correlate a textual query $\mathcal{T}$ with its target mask $\mathcal{Y}$ through a visible confounder $\mathcal{Z}$ (e.g., frequent “knife–cut” pairs). Following Pearl’s back-door criterion [39], the interventional distribution is

$$P(\mathcal{Y} \mid \mathrm{do}(\mathcal{T} = t)) = \sum_{z} P(\mathcal{Y} \mid \mathcal{T} = t, \mathcal{Z} = z)\; P(\mathcal{Z} = z) = \mathbb{E}_{\mathcal{Z}}\big[P(\mathcal{Y} \mid t, z)\big]. \tag{1}$$

**Normalized–exponential approximation.** A modern segmenter first maps inputs to pre–activation scores (logits) $s_{\mathcal{Y}}(t, z)$ and then applies the $\mathrm{Softmax}$ function to obtain probabilities. For many practical score distributions, the expectation of Softmax outputs can be closely approximated by applying the Softmax function to the expected scores. This is because for any function of the form $f(z) = \exp(g(z))$, the weighted geometric mean $\prod_z f(z)^{P(z)}$ is equal to $\exp\big(\mathbb{E}_{\mathcal{Z}}[g(z)]\big)$. Given the exponential nature of the Softmax function, this leads to the following approximation (often referred to as the Normalized Weighted Geometric Mean or NWGM approximation [2, 25]):

$$\mathbb{E}_{\mathcal{Z}}\big[\,\mathrm{Softmax}\big(s_{\mathcal{Y}}(t, z)\big)\big] \;\approx\; \mathrm{Softmax}\big(\mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Y}}(t, z)]\big). \tag{2}$$

**Additive score assumption.** We further assume the pre-activation scores decompose as $s_{\mathcal{Y}}(t, z) \simeq s_{\mathcal{T}}(t) + s_{\mathcal{Z}}(z)$, a standard design when text and bias features are fused by summation before the final classifier. Substituting into Equation (2) gives the *de-confounded score*:

$$s'_{\mathcal{Y}}(t) = s_{\mathcal{T}}(t) + \mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Z}}(z)]. \tag{3}$$

**Practical estimator.** We instantiate $s_{\mathcal{T}}(t) = \mathbf{w}^{\top}\mathbf{f}_{\mathcal{T}}(t)$ with a text encoder $\mathbf{f}_{\mathcal{T}}$. A dictionary $\{\mathbf{f}_{\mathcal{Z}}(z_i)\}_{i=1}^{K}$ of confounder embeddings is built once from the training set (each $z_i$ is a unique object–action pair); the empirical frequency of each $z_i$ serves as $P(z_i)$. The expectation becomes the fixed vector $\bar{\mathbf{f}}_{\mathcal{Z}} = \sum_{i=1}^{K} P(z_i)\, \mathbf{f}_{\mathcal{Z}}(z_i)$. Finally, the de-biased text representation is

$$\mathbf{f}'_{\mathcal{T}}(t) = \mathbf{f}_{\mathcal{T}}(t) + \bar{\mathbf{f}}_{\mathcal{Z}}, \qquad \text{and } s'_{\mathcal{Y}}(t) = \mathbf{w}^{\top}\mathbf{f}'_{\mathcal{T}}(t). \tag{4}$$

This implements Equation (3), providing an approximation of $P(\mathcal{Y} \mid \mathrm{do}(\mathcal{T}))$ that is provably back-door adjusted under the stated assumptions.

### 4.2 Visual De-biasing via Front-Door Adjustment

The visual pathway is confounded by an *unobserved* confounder $\mathcal{U}$ (e.g., rapid camera motion, occlusions), rendering a back-door adjustment strategy impossible. Instead, we exploit the two-step mediator process $\mathcal{X} \to (\mathcal{M}_v, \mathcal{M}_d) \to \mathcal{M} \to \mathcal{Y}$ to apply front-door identification, where $\mathcal{M}_v$ represents semantic visual features and $\mathcal{M}_d$ represents geometric features.

**Front-door estimand.** For the causal chain $\mathcal{X} \to (\mathcal{M}_v, \mathcal{M}_d) \to \mathcal{M} \to \mathcal{Y}$, Pearl's front-door theorem yields:

$$P(\mathcal{Y} \mid \mathrm{do}(\mathcal{X} = x)) = \sum_{m} \sum_{x'} P(\mathcal{Y} \mid \mathcal{M} = m, \mathcal{X} = x')\, P(\mathcal{M} = m \mid \mathcal{X} = x)\, P(\mathcal{X} = x'). \tag{5}$$

To implement this, two key expectations need to be approximated using their feature embeddings (denoted by bold capitals): (i) the effect of general visual context $P(x')$, approximated via $\hat{\mathbf{X}} \approx \mathbb{E}_{\mathcal{X}'}[\mathbf{X}']$, and (ii) the effect of the current visual input $x$ on the mediator $P(m \mid x)$, approximated via $\hat{\mathbf{M}} \approx \mathbb{E}_{\mathcal{M}|\mathcal{X}=x}[\mathbf{M}]$.

**Mediator design.** $\mathcal{M}_v$ encodes high-level semantics from RGB features but can be sensitive to the confounder $\mathcal{U}$. $\mathcal{M}_d$ encodes geometric information (e.g., from pretrained monocular estimation model) and is empirically more robust to $\mathcal{U}$. These are extracted by modality-specific encoders, producing sets of token vectors: $\mathbf{M}_v(x) = \{\mathbf{m}_{v,j}\}_{j=1}^{L_v}$ from the current visual input $x$, and similarly $\mathbf{M}_d(x) = \{\mathbf{m}_{d,k}\}_{k=1}^{L_d}$.

**Self-normalizing token aggregation for mediator components.** To approximate the conditional expectations $\mathbb{E}[\mathbf{M}_v \mid \mathcal{X} = x]$ and $\mathbb{E}[\mathbf{M}_d | \mathcal{X} = x]$ (which contribute to forming $\hat{\mathbf{M}}$), we employ a weighted aggregation of tokens. The weights are derived from a normalized exponential of dot-product similarities between query projections (derived from $x$) and key projections of the tokens:

$$\hat{\mathbf{M}}_v = \sum_{j=1}^{L_v} \frac{\exp(\langle \mathbf{q}_v(x), \mathbf{k}_{v,j}(\mathbf{m}_{v,j}) \rangle)}{\sum_{p=1}^{L_v} \exp(\langle \mathbf{q}_v(x), \mathbf{k}_{v,p}(\mathbf{m}_{v,p}) \rangle)}\, \mathbf{m}_{v,j}, \;\; \hat{\mathbf{M}}_d = \sum_{k=1}^{L_d} \frac{\exp(\langle \mathbf{q}_d(x), \mathbf{k}_{d,k}(\mathbf{m}_{d,k}) \rangle)}{\sum_{q=1}^{L_d} \exp(\langle \mathbf{q}_d(x), \mathbf{k}_{d,q}(\mathbf{m}_{d,q}) \rangle)}\, \mathbf{m}_{d,k}, \tag{6}$$

where $\mathbf{q}_*(\cdot)$ and $\mathbf{k}_{*,\cdot}(\cdot)$ represent learned query and key projection functions.

**Cross-modal aggregation for the final mediator $\hat{\mathbf{M}}$.** We assume that the latent mediator $\mathbf{M}$ is a *convex combination* of the visual tokens $\mathbf{M}_v = \{\mathbf{m}_{v,1}, \ldots, \mathbf{m}_{v,L_v}\}$ whose coefficients depend only on the geometric cue $\mathbf{M}_d$. Restricting the predictor to this **attention-linear family (ALF)** yields the set:

$$\mathcal{F}_{\text{ALF}} = \Big\{ f_\alpha(\mathbf{M}_v) = \sum_j \alpha_j \mathbf{m}_{v,j} \;\Big|\; \alpha = \mathrm{Softmax}\big(s(\mathbf{M}_d, \mathbf{M}_v)\big) \Big\}. \tag{7}$$

With squared error as the loss, the minimum–mean–square–error (MMSE) estimator inside $\mathcal{F}_{\text{ALF}}$ is obtained by choosing the scaled dot-product score $s(\mathbf{M}_d, \mathbf{m}_{v,j}) = \langle W_Q \mathbf{M}_d,\, W_K \mathbf{m}_{v,j} \rangle / \sqrt{d}$.

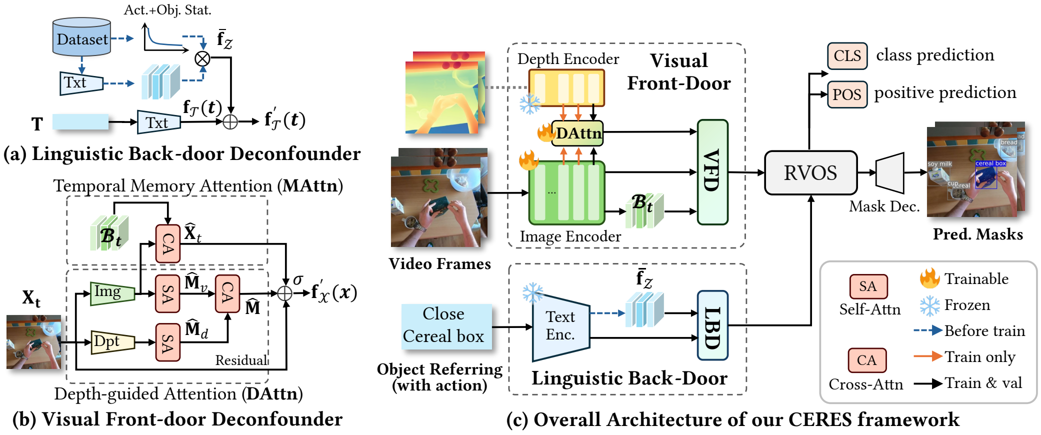
Appendix proves this result via a Lagrange-multiplier solution of the simplex-constrained quadratic program [4], following the arguments of Perez et al. [41].

Hence the conditional expectation takes the familiar cross-attention form [48]:

$$\mathbb{E}[\mathbf{M} \,|\, \mathbf{M}_v(x), \mathbf{M}_d(x)] = \sum_{j=1}^{L_v} \frac{\exp\big(\langle W_Q \mathbf{M}_d(x), W_K \mathbf{m}_{v,j} \rangle\big)}{\sum_{p=1}^{L_v} \exp\big(\langle W_Q \mathbf{M}_d(x), W_K \mathbf{m}_{v,p} \rangle\big)}\, \mathbf{m}_{v,j}. \tag{8}$$

Because $\hat{\mathbf{M}}_d(x) \approx \mathbb{E}[\mathbf{M}_d \,|\, \mathcal{X} = x]$ is purposely designed to be insensitive to the egocentric bias $\mathcal{U}$, we obtain a bias-robust estimate of the overall mediator by plugging $\hat{\mathbf{M}}_d(x)$ into Equation (8): $\hat{\mathbf{M}}(x) = \mathrm{Attn}(Q = \hat{\mathbf{M}}_d,\; K = V = \mathbf{M}_v)$. Equation (8) concretely instantiates our claim that "$\mathbf{M}_d$ guides the aggregation of $\mathbf{M}_v$", while ensuring that the learned weights are less exposed to the confounder $\mathcal{U}$. A complete derivation and additional empirical justification are provided in Appendix.

**Memory-bank estimator for $\hat{\mathbf{X}}$.** Most front-door implementations [60, 49] pre-compute a global dictionary of training frames and apply cross-attention against that dictionary to approximate the expectation $\mathbb{E}_{\mathbf{X}'}[\mathbf{X}']$.



**(a) Linguistic Back-door Deconfounder**

**(b) Visual Front-door Deconfounder**

**(c) Overall Architecture of our CERES framework**

Figure 3: Overview of the CERES framework. The Linguistic Back-door Deconfounder (LBD) de-biases the input text query $\mathbf{T}$ into features $\mathbf{f}'_{\mathcal{T}}(t)$. Concurrently, the Visual Front-door Deconfounder (VFD) processes the video frame $\mathbf{X}_t$; it forms a vision-depth mediator $\hat{\mathbf{M}}(x_t)$ using Depth-guided Attention (DAttn) and estimates temporal visual context $\hat{\mathbf{X}}_t$ via Memory Attention (MAttn), yielding de-biased visual features $\mathbf{f}'_{\mathcal{X}}(x_t)$. These de-biased multimodal features are then used by the RVOS model to predict the segmentation mask $\hat{\mathbf{Y}}_t$.

Such a static lookup is impractical for long, dynamic egocentric streams. We instead assume (A1) short-range stationarity: within a window of $W$ frames, the marginal distribution of the raw frames $x_t$ is approximately unchanged. Under (A1), the last $W$ frames form an i.i.d. Monte-Carlo sample of the stationary distribution, so their empirical average is an unbiased estimate of $\mathbb{E}_{\mathbf{X}'}[\mathbf{X}']$. Concretely, we keep a sliding memory bank $\mathcal{B}_t = \{x_{t-\tau}\}_{\tau=1}^{W}$ and define

$$\hat{\mathbf{X}}_t = \sum_{\tau=1}^{W} \frac{\exp\big(\mathrm{sim}(\mathbf{x}_t, \mathbf{x}_{t-\tau})\big)}{\sum_{\sigma=1}^{W} \exp\big(\mathrm{sim}(\mathbf{x}_t, \mathbf{x}_{t-\sigma})\big)} \, \mathbf{x}_{t-\tau}, \tag{9}$$

where $\mathbf{x}_t$ (and $\mathbf{x}_{t-\tau}$) is the frame-level embedding of $x_t$, and $\mathrm{sim}(\cdot,\cdot)$ is a learnable dot-product score as in standard attention [48]. Equation (9) is nothing but a soft weighting of Monte Carlo samples, similar to temperature-scaled importance sampling [63].

**Theoretical guarantee.** By the law of large numbers, the empirical mean over $\mathcal{B}_t$ converges to $\mathbb{E}_{\mathbf{X}'}[\mathbf{X}']$ as $W \to \infty$. Because the softmax weights in Equation (9) satisfy $\sum_\tau w_\tau = 1$ and are bounded, the same convergence holds for $\hat{\mathbf{X}}_t$. Combined with the Attention-Linear-Family (ALF) model used for $\hat{\mathbf{M}}$, the overall network is a consistent estimator of the front-door integral.

### 4.3 Overall Architecture

The CERES framework applies our proposed causal adjustment modules, the Linguistic Back-door Deconfounder (LBD) and the Visual Front-door Deconfounder (VFD), to a pre-trained Referring Video Object Segmentation (RVOS) model, typically trained on third-person datasets. An overview of the architecture is depicted in Figure 3.

**Linguistic Back-door Deconfounder (LBD).** To mitigate language bias, the LBD module (Section 4.1) first constructs a confounder dictionary. This dictionary comprises embeddings of unique object-action pairs ($z_i$) and their empirical frequencies $P(z_i)$ derived from the training dataset statistics. The text encoder from the pre-trained RVOS model is used to obtain the initial text query embedding $\mathbf{f}_{\mathcal{T}}(t)$ and the confounder embeddings $\mathbf{f}_{\mathcal{Z}}(z_i)$. During both training and inference, the de-biased text representation $\mathbf{f}'_{\mathcal{T}}(t)$ is computed using Eq. (4), effectively adjusting for spurious correlations learned from dataset statistics.

**Visual Front-door Deconfounder (VFD).** The VFD module (Section 4.2) addresses visual confounding. For each video frame $x_t$, visual features $\mathbf{X}_t^{\text{rgb}}$ are extracted using the image encoder of the pre-trained RVOS model. Concurrently, geometric depth features $\mathbf{X}_t^{\text{depth}}$ are obtained from a pre-trained monocular depth estimation model's encoder.

To construct the mediator component $\hat{\mathbf{M}}(x_t)$, features from the last $n$ layers of both the RGB encoder ($\mathbf{M}_{v,l}(x_t)$ for layer $l$) and depth encoder ($\mathbf{M}_{d,l}(x_t)$ for layer $l$) are utilized. For each of these $n$ layers, **Depth-guided mediator Attention (DAttn)** combines these modalities:

$$\hat{\mathbf{m}}_l(x_t) = \mathrm{DAttn}(Q = \hat{\mathbf{M}}_{d,l}(x_t), K = V = \mathbf{M}_{v,l}(x_t)), \tag{10}$$

Table 1: Comparison (%) with state-of-the-art methods on VISOR. Best results are highlighted in **bold**, second-best in underline within the same backbone. ↑ indicates higher is better, ↓ indicates lower is better.

| **Method** | **Backbone** | **mIoU**$^{\oplus}$**↑** | **cIoU**$^{\oplus}$**↑** | **mIoU**$^{\ominus}$**↓** | **cIoU**$^{\ominus}$**↓** | **gIoU↑** | **Acc↑** | **F1↑** |
|---|---|---|---|---|---|---|---|---|
| ReferFormer | R101 | 59.9 | 66.4 | 30.5 | 52.1 | 55.3 | 58.6 | 64.2 |
| ReferFormer+ | R101 | 58.2 | 64.8 | **14.3** | **18.9** | 63.1 | 67.6 | 68.9 |
| ActionVOS | R101 | 59.9 | 67.2 | 16.3 | 28.5 | 69.9 | 73.4 | 73.7 |
| Ours | R101 | **64.0** | **72.8** | 15.3 | 25.6 | **72.4** | **76.3** | **77.1** |
| ReferFormer+ | VSwinB | 61.1 | 68.9 | 19.2 | 36.8 | 68.4 | 73.2 | 74.0 |
| ActionVOS | VSwinB | 62.9 | 70.9 | 20.0 | 38.8 | 69.5 | 70.7 | 74.3 |
| Ours | VSwinB | **65.4** | **72.5** | **19.1** | **35.1** | **72.1** | **74.7** | **75.9** |
| HOS | SwinL | 55.1 | 59.2 | **13.5** | **17.3** | 66.5 | 70.3 | 69.7 |
| ActionVOS | SwinL | 66.3 | 71.9 | 22.8 | 42.5 | 68.7 | 73.4 | 75.5 |
| Ours | SwinL | **67.0** | **73.6** | 16.9 | 28.6 | **71.8** | **75.2** | **76.2** |

Table 2: Comparison (%) with state-of-the-art methods on the novel subset of VISOR. Best results are in **bold**, second-best in underline. ↑ indicates higher is better, ↓ indicates lower is better.

| **Method** | **mIoU**$^{\oplus}$**↑** | **cIoU**$^{\oplus}$**↑** | **mIoU**$^{\ominus}$**↓** | **cIoU**$^{\ominus}$**↓** | **gIoU↑** | **Acc↑** | **F1↑** |
|---|---|---|---|---|---|---|---|
| ReferFormer+ | 47.2 | 54.1 | 13.5 | 21.8 | 50.1 | 51.9 | 57.7 |
| HOS | 45.8 | 49.7 | **7.8** | **11.1** | 61.9 | 64.5 | 66.0 |
| ActionVOS | 55.3 | 62.8 | 14.5 | 25.4 | 65.8 | 69.4 | 71.9 |
| Ours | **60.0** | **69.9** | 14.4 | 20.4 | **67.9** | **72.2** | **75.8** |

where $\hat{\mathbf{M}}_{d,l}(x_t)$ is the aggregated depth feature representation for layer $l$ (analogous to Eq. (6) applied layer-wise). This yields layer-specific mediator representations $\{\hat{\mathbf{m}}_l(x_t)\}_{l=1}^{n}$. For the final mask prediction, we typically use the mediator from the last processed layer, $\hat{\mathbf{M}}(x_t) = \hat{\mathbf{m}}_n(x_t)$.

To estimate the general visual context $\hat{\mathbf{X}}_t$, a memory bank stores recent frame features (from $\mathbf{X}_{\text{rgb}}$), augmented with temporal positional encodings. **Temporal Memory Attention (MAttn)** then computes $\hat{\mathbf{X}}_t$ as per Eq. (9). The de-biased visual feature $\mathbf{f}'_{\mathcal{X}}(x_t)$ is then formed by integrating the mediator and context information. Specifically, the final mediator $\hat{\mathbf{M}}(x_t)$ and the context $\hat{\mathbf{X}}_t$ are concatenated and processed through an MLP followed by a gated residual connection:

$$\mathbf{f}'_{\mathcal{X}}(x_t) = \sigma_{\mathbf{M},\mathbf{X}} \cdot \text{MLP}([\hat{\mathbf{M}}(x_t); \hat{\mathbf{X}}_t]) + (1 - \sigma_{\mathbf{M},\mathbf{X}}) \cdot \mathbf{X}_t. \tag{11}$$

This $\mathbf{f}'_{\mathcal{X}}(x_t)$ represents the visual input adjusted for egocentric confounders, $\sigma_{\mathbf{M},\mathbf{X}}$ is the gate of residual path for easier to train [45, 20]. For auxiliary losses during training, similar de-biased features $\mathbf{f}'_{\mathcal{X},l}(x_t)$ are computed using the intermediate layer-specific mediators $\hat{\mathbf{m}}_l(x_t)$.

**Output Generation and Training.** The de-biased text features $\mathbf{f}'_{\mathcal{T}}(t)$ and visual features $\mathbf{f}'_{\mathcal{X}}(x_t)$ (or $\mathbf{f}'_{\mathcal{X},l}(x_t)$ for intermediate layers) are then utilized by the subsequent components of the RVOS model. These typically include a classification head to predict the object category, a "positive" head to identify if the object is actively involved in the action, and a mask decoder to generate the final segmentation mask $\hat{y}_t$. The model is trained using a standard segmentation loss. Following prior work, auxiliary segmentation losses are applied to the outputs derived from the de-biased visual features of the $n$ intermediate layers during training. During inference, only the de-biased visual feature from the final considered layer is used for prediction.

# 5 Experiment

## 5.1 Experiment Settings

• **Datasets.** Following previous work [36], we evaluate our method on three public egocentric video datasets: VISOR [12], VOST [46], and VSCOS [62]. VISOR, derived from EPIC-KITCHENS [10, 11], provides annotations for hands and active object interactions; we utilize its training and validation splits. After pre-processing, this yields 13,205 videos (76,873 objects) for training and 467 videos (1,841 objects) for validation, where validation objects are manually annotated as positive or negative. VOST and VSCOS are used for validation only. VOST assesses performance on objects undergoing transformations. VSCOS focuses on state-changing objects; its validation data is filtered to prevent overlap with the VISOR training set.

- **Metrics.** We follow established evaluation protocols [26, 36]. Key metrics include mean Intersection over Union (mIoU) and cumulative IoU (cIoU), reported separately for positive (mIoU$^{\oplus}$, cIoU$^{\oplus}$) and negative (mIoU$^{\ominus}$, mIoU$^{\ominus}$) objects to assess segmentation of interacted and non-interacted instances, respectively. We also report generalized IoU (gIoU) [26] for a combined assessment of segmentation and target classification. Additionally, Precision (P), Recall (R), and Accuracy (Acc) are used for the binary classification of object activity. For these classification-related evaluations (gIoU, P, R, Acc), a prediction is considered a True Positive (TP) if its IoU with the ground truth exceeds 0.5. More metrics details are in the Appendix D.

- **Implementation Details.** CERES builds upon the ReferFormer [57] architecture, initializing with its pre-trained weights. We evaluate ResNet101 [20], Swin-Transformer-L [30], and Video Swin-Transformer-B [31] as image encoder backbones, paired with a RoBERTa [29] text encoder. Input frames are resized to 448 × 448. For the Depth-guided Attention (DAttn) of Visual Front-door Deconfounder (VFD), depth features are extracted using the encoder of a frozen pre-trained Depth Anything V2 model [59]. The Linguistic Back-door Deconfounder (LBD) defines confounders $z_i$ based on unique "verb-noun" pairs identified in the training set queries. The VFD's Temporal Memory Attention (MAttn) employs a window of $W = 5$ recent frames. The model is trained with a batch size of 4. Following the previous implement, auxiliary losses during training utilize visual features from the last three layers of the image encoder. During inference, predictions are made **online**, without access to future frames. Further optimization and hyperparameter details are provided in the Appendix.

## 5.2 Comparison with State-of-the-art Methods

We compare our CERES against several leading methods. ReferFormer [57] serves as a foundational baseline, representing a model pre-trained on third-person data and subsequently fine-tuned for Ego-RVOS. ReferFormer+ extends this by incorporating an auxiliary prediction head for "positiveness" (identifying if the object is actively involved in the action) and by including action descriptions in the referring query. EgoHOS [67] is a hand-object segmentation model; for our comparison, we train it on VISOR, treating hand-associated objects as positive targets. ActionVOS [36] is a strong recent baseline that also builds on pre-trained models and employs specialized losses for active objects.

Table 3: Comparison (%) of positive objects IoU on VSCOS and VOST.

| **Method** | **VSCOS** | | **VOST** | |
|---|---|---|---|---|
| | **mIOU↑** | **cIOU↑** | **mIOU↑** | **cIOU↑** |
| ReferFormer+ | 53.0 | 54.2 | 30.6 | 16.1 |
| HOS | 42.1 | 31.2 | 21.9 | 15.2 |
| ActionVOS | 52.5 | 57.7 | 30.2 | 17.6 |
| Ours | **55.3** | **62.5** | **32.0** | **21.7** |

On the VISOR benchmark (Table 1), CERES consistently outperforms prior methods across all backbones. With ResNet101, CERES achieves 64.0% mIoU$^{\oplus}$ (+4.1% over ActionVOS) and 72.4% gIoU (+2.5%), alongside improved accuracy and F1 scores for positive object classification. CERES also generally shows lower mIoU$^{\ominus}$ and cIoU$^{\ominus}$, indicating better discrimination against non-target objects. While HOS has low mIoU$^{\ominus}$ due to its focus on hand-proximate objects, its mIoU$^{\oplus}$ is consequently limited.

CERES's advantages are particularly evident on a VISOR subset with novel objects or actions not seen during training (Table 2). This subset serves as an open-vocabulary evaluation within VISOR, and CERES significantly surpasses previous state-of-the-art results, demonstrating strong generalization capabilities likely due to its mitigation of dataset-induced language biases.

Evaluations on VSCOS and VOST (Table 3), which feature significant object transformations, further underscore CERES's robustness. We treat these as zero-shot open-vocabulary generalization: models are trained only on VISOR and evaluated on VSCOS/VOST without any fine-tuning. For instance, on VSCOS, CERES achieves 55.3% mIoU$^{\oplus}$ and 62.5% cIoU$^{\oplus}$, exceeding ActionVOS (52.5% and 57.7%). This superior performance in challenging scenarios highlights the effectiveness of CERES's causal intervention strategies in addressing both linguistic and visual confounding factors in egocentric videos.

## 5.3 Ablation Study

We conduct ablation studies on the VISOR dataset using the ResNet101 backbone to analyze the contribution of each key component in CERES.

**Ablation of Proposed Modules.** In Table 4, the baseline model (first row) achieves 59.9% mIoU$^{\oplus}$ and 69.9% gIoU. Introducing only the LBD (second row) improves mIoU$^{\oplus}$ to 61.2% (+1.3%) and gIoU to 71.4% (+1.5%), demonstrating its effectiveness in mitigating language bias. A nonlinear MLP-based depth fusion (third row) for the mediator offers a 62.1% mIoU$^{\oplus}$ but increases mIoU$^{\ominus}$, indicating limited discriminative benefit. In contrast, our DAttn depth integration (fourth row) significantly boosts mIoU$^{\oplus}$ to 63.3% (+3.4% over baseline) and reduces mIoU$^{\ominus}$ to 15.8%, yielding substantial gains in gIoU (71.8%) and Acc (75.3%). This confirms

Table 4: Ablation study (%) of proposed modules on VISOR (ResNet101). (◇ indicates MLP-based depth fusion)

| DAttn | MAttn | LBD | $\text{mIoU}^{\oplus}\uparrow$ | $\text{mIoU}^{\ominus}\downarrow$ | gIoU↑ | Acc↑ |
|---|---|---|---|---|---|---|
| | | | 59.9 | 16.3 | 69.9 | 73.4 |
| | | ✓ | 61.2 | 16.0 | 71.4 | 74.8 |
| ◇ | | | 62.1 | 17.5 | 70.5 | 73.6 |
| ✓ | | | 63.3 | 15.8 | 71.8 | 75.3 |
| ✓ | ✓ | | 63.1 | **14.9** | 72.1 | 76.1 |
| ✓ | ✓ | ✓ | **64.0** | 15.3 | **72.4** | **76.3** |

Table 5: Performance comparison on a "hard" subset of VISOR. (RF means ReferFormer [57], ActV means ActionVOS [36])

| Method | $\text{mIoU}^{\oplus}\uparrow$ | $\text{mIoU}^{\ominus}\downarrow$ | gIoU↑ | Acc↑ |
|---|---|---|---|---|
| RF | 53.4 | 14.9 | 56.3 | 58.5 |
| RF+ | 54.2 | 14.7 | 56.7 | 58.7 |
| ActV | 58.4 | 15.1 | 69.9 | 73.1 |
| Ours | **62.3** | **14.3** | **72.2** | **75.6** |

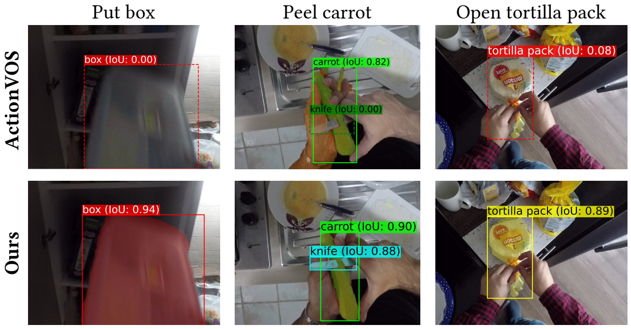


Figure 4: Qualitative analysis of ActionVOS and our CERES.

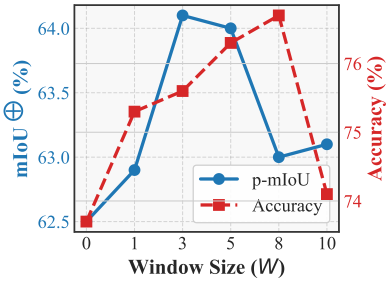


Figure 5: Effect of $W$ in MAttn.

the superiority of our causally-inspired depth mediator design within the VFD. Adding MAttn (fifth row) to complete the VFE (DAttn + MAttn) further refines performance. This configuration achieves the lowest $\text{mIoU}^{\ominus}$ (14.9%) and a strong 72.1% gIoU, highlighting TCA's role in improving discrimination by modeling broader visual context. The full CERES model (last row), integrating both the complete VFE and LBD, attains the best overall performance, with 64.0% $\text{mIoU}^{\oplus}$, 72.4% gIoU, and 76.3% Acc. While its $\text{mIoU}^{\ominus}$ (15.3%) is slightly higher than VFE-only (14.9%), the LBD's inclusion enhances the recall of positive instances. This trade-off results in superior overall identification and segmentation of target objects. These ablations validate the individual and synergistic contributions of our causal adjustment modules.

**Temporal Context Window Size for MAttn.** We analyze the impact of the temporal window size $W$ for the MAttn module in Figure 5. Setting $W = 0$ (i.e., no MAttn module) results in lower performance compared to using MAttn. As $W$ increases, $\text{mIoU}^{\oplus}$ and Acc generally improve. We found $W = 5$ provides a robust balance and consistently strong results.

**Performance on Rare Concepts.** To assess robustness against data scarcity biases, we evaluate models on a "hard" subset of VISOR, comprising 159 clips with actions or objects appearing <50 times in training. Table 5 shows that, compared to **ActionVOS**, **CERES** improves $\text{mIoU}^{\oplus}$ by +3.9% (62.3% vs 58.4%) and gIoU by +2.3% (72.2% vs 69.9%). These gains are mainly attributed to **LBD** blocking the spurious path $\mathcal{T} \leftarrow \mathcal{Z} \rightarrow \mathcal{Y}$ (back-door), while **VFD** mitigates egocentric visual confounders $\mathcal{U}$ via the vision–depth mediator (front-door). This highlights the effectiveness of CERES, particularly the LBD component, in generalizing to less frequent concepts by mitigating reliance on spurious statistical correlations.

**Qualitative Analysis.** Figure 4 qualitatively compares CERES with ActionVOS [36], showcasing CERES's superior robustness. Our method yields more accurate segmentation in challenging egocentric scenarios, including those with visual distortions like motion blur (e.g., "box", column 1) and occlusions (e.g., "knife", column 2). This highlights the VFD module's effectiveness in mitigating visual biases. Furthermore, CERES demonstrates improved handling of textual queries involving uncommon objects or actions (e.g., "tortilla pack", column 3), where ActionVOS may falter due to dataset biases. This underscores the LBD module's contribution to better language grounding. Overall, these visual examples corroborate our quantitative results, illustrating how CERES's causal interventions lead to more robust Ego-RVOS.

## 6 Conclusion

This paper introduced Causal Ego-REferring Segmentation (CERES), a novel framework that applies causal inference principles to address critical robustness challenges in Egocentric Referring Video Object Segmentation. We identified two primary sources of error: language biases stemming from dataset statistics and visual confounding inherent in the egocentric perspective. CERES tackles these by employing backdoor adjustment to mitigate spurious correlations between textual queries and segmentation outputs, and by utilizing front-door adjustment with a novel vision-depth mediator to counteract the effects of unobserved visual confounders. This dual-pronged causal intervention allows CERES to learn more robust representations, less susceptible to dataset-specific biases and egocentric visual distortions. Extensive experiments on standard Ego-RVOS benchmarks demonstrate that CERES achieves state-of-the-art performance, significantly improving segmentation accuracy and reliability, particularly in challenging scenarios with novel concepts or significant visual ambiguity. Our work underscores the potential of causal reasoning to build more generalizable and trustworthy models for complex egocentric video understanding tasks.

## Acknowledgments

This work was supported by National Key R&D Program of China under Grant No. 2021ZD0111601, National Natural Science Foundation of China (NSFC) under Grant No. 62272494, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2023A1515012845 and 2023A1515011374, and Guangdong Province Key Laboratory of Information Security Technology.

## References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Qinfeng Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10041–10051, 2020. doi: 10.1109/CVPR42600.2020.01006.

[2] Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial Intelligence*, 210:78–122, 2014. ISSN 0004-3702. doi: 10.1016/j.artint.2014.02.004.

[3] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4975–4985, 2022. doi: 10.1109/CVPR52688.2022.00493.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. Opening the vocabulary of egocentric actions. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[6] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 522–531, 2019.

[7] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, Lingbo Liu, and Liang Lin. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(12):9887–9903, 2022.

[8] Ho Kei Cheng, Seoung Wug Oh, Brian L. Price, Alexander G. Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1316–1326, 2023. doi: 10.1109/ICCV51070.2023.00127.

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. volume 26, 2013.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. doi: 10.1007/978-3-030-01225-0_44.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, pages 1–23, 2022. doi: 10.1007/s11263-021-01531-2.

[12] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 13745–13758, 2022. doi: 10.5555/3600270.3601269.

[13] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2694–2703, 2023. doi: 10.1109/ICCV51070.2023.00254.

[14] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(6):7900–7916, 2023. doi: 10.1109/TPAMI.2022.3217852.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[16] Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.

[17] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5958–5966, 2018. doi: 10.1109/CVPR.2018.00624.

[18] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/S42256-020-00257-Z. URL https://doi.org/10.1038/s42256-020-00257-z.

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[21] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165, 2020.

[22] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision (ACCV)*, pages 123–141, 2019. doi: 10.1007/978-3-030-20870-7_8.

[23] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21241–21250, 2022. doi: 10.1109/CVPR52688.2022.02059.

[24] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 22179–22188. IEEE, 2023. doi: 10.1109/ICCV51070.2023.02032.

[25] Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18020–18029. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01751.

[26] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23592–23601, 2023. doi: 10.1109/CVPR52729.2023.02259.

[27] Haijing Liu, Tao Pu, Hefeng Wu, Keze Wang, and Liang Lin. DART: dual adaptive refinement transfer for open-vocabulary multi-label recognition. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2025.

[28] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10):11624–11641, 2023. doi: 10.1109/TPAMI.2023.3284038.

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

[31] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2022. doi: 10.1109/CVPR52688.2022.00320.

[32] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2016. doi: 10.1109/CVPR.2016.9.

[33] Lei Meng, Xiangxian Li, Xiaoshuo Yan, Haokai Ma, Zhuang Qi, Wei Wu, and Xiangxu Meng. Causal inference over visual-semantic-aligned graph for image classification. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 19449–19457, 2025. doi: 10.1609/AAAI.V39I18.34141.

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, 2014. doi: 10.1109/CVPR.2014.119.

[35] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705, 2021. doi: 10.1109/CVPR46437.2021.01251.

[36] Liangyang Ouyang, Ruicong Liu, Yifei Huang, Ryosuke Furuta, and Yoichi Sato. Actionvos: Actions as prompts for video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15068, pages 216–235, 2024. doi: 10.1007/978-3-031-72684-2_13.

[37] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Javen Qinfeng Shi, and Anton van den Hengel. Counterfactual vision-and-language navigation: Unravelling the unseen. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5296–5307. Curran Associates, Inc., 2020.

[38] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[39] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

[40] Simone Alberto Peirone, Gabriele Goletto, Mirco Planamente, Andrea Bottino, Barbara Caputo, and Giuseppe Averta. Egocentric zone-aware action recognition across environments. *Pattern Recognition Letters*, 188:140–147, 2025. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2024.12.008.

[41] Guillaume Perez, Sebastian Ament, Carla Gomes, and Michel Barlaud. Efficient projection algorithms onto the weighted l1 ball. *Artificial Intelligence*, 306:103683, 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2022.103683.

[42] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(3):1623–1637, 2022. doi: 10.1109/TPAMI.2020.3019967.

[43] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. doi: 10.1109/CVPR.2019.00075.

[44] Yuhan Shen, Huiyu Wang, Xitong Yang, Matt Feiszli, Ehsan Elhamifar, Lorenzo Torresani, and Effrosyni Mavroudi. Learning to segment referred objects from narrated egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14510–14520, 2024. doi: 10.1109/CVPR52733.2024.01375.

[45] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks, 2015. URL https://arxiv.org/abs/1505.00387.

[46] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22836–22845, 2023. doi: 10.1109/CVPR52729.2023.02187.

[47] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6558–6569. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1656.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017.

[49] Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13139–13150, 2024. doi: 10.1109/CVPR52733.2024.01248.

[50] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10757–10767, 2020. doi: 10.1109/CVPR42600.2020.01077.

[51] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3071–3080, 2021. doi: 10.1109/ICCV48922.2021.00308.

[52] Wenxuan Wang, Yisi Zhang, Xingjian He, Yichen Yan, Zijia Zhao, Xinlong Wang, and Jing Liu. Beyond literal descriptions: Understanding and locating open-world objects aligned with human intentions. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics (ACL Finding)*, pages 762–776, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.43.

[53] Yuqing Wang, Xiangxian Li, Zhuang Qi, Jingyu Li, Xuelong Li, Xiangxu Meng, and Lei Meng. Meta-causal feature learning for out-of-distribution generalization. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Proceedings of the European Conference on Computer Vision Workshops (ECCV Workshops)*, volume 13806 of *Lecture Notes in Computer Science*, pages 530–545, 2022. doi: 10.1007/978-3-031-25075-0_36.

[54] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10213–10222, 2020. doi: 10.1109/CVPR42600.2020.01023.

[55] Hefeng Wu, Weifeng Chen, Zhibin Liu, Tianshui Chen, Zhiguang Chen, and Liang Lin. Contrastive transformer learning with proximity data generation for text-based person search. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7005–7016, 2024.

[56] Hefeng Wu, Yandong Chen, Lingbo Liu, Tianshui Chen, Keze Wang, and Liang Lin. SQLNet: Scale-modulated query and localization network for few-shot class-agnostic counting. *IEEE Transactions on Image Processing (TIP)*, 34:4631–4645, 2025.

[57] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4964–4974, 2022. doi: 10.1109/CVPR52688.2022.00492.

[58] Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. Segment every reference object in spatial and temporal spaces. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2538–2550, 2023. doi: 10.1109/ICCV51070.2023.00240.

[59] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 21875–21911, 06 2024. doi: 10.48550/arXiv.2406.09414.

[60] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9842–9852, 2021. doi: 10.1109/CVPR46437.2021.00972.

[61] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(11):12996–13010, 2023. doi: 10.1109/TPAMI.2021.3121705.

[62] Jiangwei Yu, Xiang Li, Xinran Zhao, Hongming Zhang, and Yu-Xiong Wang. Video state-changing object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20439–20448, 2023. doi: 10.1109/ICCV51070.2023.01869.

[63] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15922–15932, 2023. doi: 10.1109/CVPR52729.2023.01528.

[64] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. ISBN 9781713829546.

[65] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15404–15414, 2021. doi: 10.1109/CVPR46437.2021.01515.

[66] Hua Zhang, Xiaohong Li, Xiaochun Cao, and Hassan Foroosh. Multiple adverse weather conditions adaptation for object detection via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(3):1742–1756, 2024. doi: 10.1109/TPAMI.2022.3166765.

[67] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 127–145, 2022. doi: 10.1007/978-3-031-19818-2_8.

[68] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 4373–4382, 2020. doi: 10.1145/3394171.3413518.

# NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim CERES introduces a novel causal inference framework for Ego-RVOS, using backdoor adjustment for language biases and front-door adjustment with a novel vision-depth mediator for visual confounding. The paper's methodology (Section 4) details these causal mechanisms. Experimental results (Section 5), including SOTA performance on standard benchmarks (VISOR, VOST, VSCOS) and ablation studies, directly support these contributions and demonstrate the effectiveness of the proposed interventions within the defined scope of Ego-RVOS. The claims regarding improved robustness against these specific biases are also substantiated by comparative and ablation experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We dedicated a paragraph to this (Limitations') as part of the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Key assumptions are stated in the main text, and complete proofs or detailed derivations for our theoretical claims, such as the mediator construction and overall estimator consistency, are provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper details the datasets, evaluation metrics, base model, specific architectural choices for CERES components (LBD, DAttn, MAttn), key hyperparameters like batch size and window size, and explicitly states that further implementation details are available in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release our code and detailed instructions for data preparation and reproducing the main experimental results after the review period.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper outlines key experimental settings including datasets, metrics, backbone architectures, and critical module parameters in the main text, and explicitly states that further optimization and hyperparameter details are provided in the Appendix, ensuring reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Experiments were too costly to be run multiple times, however, a lot of ablations are available and the model is tested on multiple datasets.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The machine on which we run the experiments, as well as the computing time needed is mentioned in the Appendix

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our work complies with NeurIPS ethics guidelines. No human subjects or sensitive data were involved.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper explicitly discusses potential positive impacts such as advancements in assistive technologies, human-robot interaction, AR, and embodied AI, alongside negative societal impacts including privacy concerns from surveillance, potential for bias propagation, and risks of misuse, while also suggesting the need for ethical guidelines and mitigation strategies

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No model in this paper is with a high risk for misuse. The collected datasets are all open-source and accessible

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and code libraries are properly cited with licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code for the CERES framework and the corresponding trained models are new assets. Documentation is provided with the code release, detailing implementation, training procedures, and dependencies to ensure reproducibility.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not use crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our research did not involve experiments on humans subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: LLMs were only used for proofreading; they did not contribute to methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

# Appendix

## A Limitations

While CERES demonstrates notable advancements in robust Ego-RVOS, we acknowledge certain aspects for future consideration. The LBD module's current definition of confounders ($\mathcal{Z}$) as object-action pairs, while effective for the targeted dataset biases, represents one specific strategy; future work could explore more nuanced or automatically discovered confounder definitions. The additive score assumption used in the back-door adjustment, though common, is an approximation of the true deconfounded score.

For the VFD, its performance is influenced by the capabilities of the underlying pre-trained depth estimator. While we empirically show that depth aids in mitigating visual confounding, formally verifying full front-door conditions for the chosen mediator structure is nontrivial. The Attention-Linear Family (ALF) assumption explicitly trades expressivity for identifiability: restricting fusion weights to be functions of $M_d$ helps prevent leakage of the unobserved confounder $U$ and yields a minimal, identifiable realization of the mediator, but narrows the function class. Empirically, our ALF-based DAttn outperforms a nonlinear MLP fusion (Table 4), indicating superior robustness under egocentric confounding. Exploring richer mediator parameterizations that preserve front-door validity is left for future work.

The MAttn module approximates temporal context using a sliding window, a technique effective for many dynamic egocentric scenes, though its generalization to scenarios with extremely long-range temporal dependencies could be further investigated. The integration of depth features and additional attention mechanisms introduces computational costs relative to simpler baselines, a common trade-off for enhanced robustness, and further optimization could be explored.

Finally, the current evaluation of CERES is focused on the Ego-RVOS task. Extending and rigorously evaluating its applicability and potential adaptations for a broader spectrum of egocentric video understanding challenges, such as egocentric action recognition or long-term activity understanding, presents a valuable avenue for future research.

## B Broader Impact

The advancements in robust Ego-RVOS achieved by CERES hold potential for significant positive impacts. More reliable egocentric video understanding can directly benefit assistive technologies, enhancing contextual awareness for individuals with visual impairments, and can enable more intuitive human-robot interaction by allowing machines to better grasp human-object interactions from a first-person view. In fields like augmented reality, precise segmentation of actively manipulated objects can lead to more seamless and responsive user experiences. Crucially, the causal principles and architectural components developed in CERES, particularly the strategies for mitigating dataset and visual biases, may offer a foundational approach for enhancing robustness and generalizability in other egocentric video analysis tasks. This is particularly relevant for embodied AI, where a nuanced understanding of human actions and object interactions from an egocentric perspective is critical for agents to learn from human demonstrations, predict intentions, and operate safely and intelligently within complex human environments.

However, as with any technology capable of detailed scene analysis, responsible development is paramount. The potential for misuse in surveillance or intrusive monitoring, especially with personal egocentric data, necessitates strong ethical guidelines and privacy-preserving measures. While CERES aims to reduce specific biases, the underlying pre-trained models might still harbor unaddressed biases, emphasizing the need for continuous auditing and fairness considerations in AI system development. The deployment of such technologies should thus proceed with a commitment to ethical practices and ongoing research into comprehensive bias mitigation.

## C More Implementation Details

All experiments were conducted using PyTorch 2.1.2 and CUDA 11.8 on a system with four NVIDIA V100 GPUs. Models were trained for 6 epochs with a total batch size of 4, where each batch item was a single video clip. We initialized the learning rate to $1 \times 10^{-3}$ for our CERES modules and $1 \times 10^{-4}$ for pre-trained components, decaying it by 0.1 at epochs 3 and 5, using the AdamW optimizer. The primary segmentation loss combined bounding box, Dice and Focal losses. Input frames were resized to $448 \times 448$ for both training and inference.

We use ReferFormer pretrained weight on Youtube-VOS dataset. For the VFD, depth features were extracted using a frozen pre-trained Depth Anything V2 encoder. The LBD module defined confounders $\mathcal{Z}$ from unique "verb-noun" pairs in training queries. The VFD's DAttn utilized features from the last three layers of the image and depth encoders, while MAttn employed a temporal window of $W = 5$ recent frames. During inference, an online strategy was adopted, processing frames sequentially without access to future information.

# D Metric Details

This section provides further details on the evaluation metrics used in this work, complementing the descriptions in Section 5.1 of the main paper.

• **Accuracy and F1-score.** These metrics evaluate the binary classification of whether an object is actively involved in the queried action. A prediction is considered a True Positive (TP) if its Intersection over Union (IoU) with the ground-truth active object mask exceeds 0.5. This stricter threshold, compared to some prior works like ActionVOS [36] which might use IoU > 0, ensures a more accurate reflection of identification and localization. True Negatives (TN) are correctly identified non-active objects or correct "no target" predictions. False Positives (FP) are non-active objects misclassified as active, or incorrect "target present" predictions. False Negatives (FN) are active objects missed or misclassified as non-active.

Based on these, Accuracy (Acc) and F1-score are calculated as:

- Accuracy (Acc):
$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$
- F1-score:
$$\text{F1} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

The F1-score is the harmonic mean of Precision and Recall, providing a balanced measure for active object classification.

• **Mean Intersection over Union (mIoU) and Cumulative IoU (cIoU).** These are standard segmentation quality metrics. **mIoU** calculates the average IoU across all individual object instances in the dataset. For each instance, IoU is the ratio of the area of overlap between the predicted mask and the ground-truth mask to the area of their union. **cIoU** computes a single IoU value over the entire dataset (or a subset) by summing all intersection areas and dividing by the sum of all union areas. As stated in the main paper, these metrics are reported separately for positive objects (target objects actively involved in the action: $\text{mIoU}^{\oplus}$, $\text{cIoU}^{\oplus}$) and negative objects (other objects present in the scene but not involved in the queried action: $\text{mIoU}^{\ominus}$, $\text{cIoU}^{\ominus}$). Lower scores for negative objects indicate better discrimination against non-targets.

• **Generalized IoU (gIoU).** Proposed by Liu et al. [26], gIoU offers a combined assessment of both segmentation quality and the model's ability to correctly classify target presence. It is calculated as the mean of per-sample scores, making it less sensitive to object size variations compared to cIoU. The per-sample score for gIoU is determined as follows:

- If a ground-truth target object exists for the query in a given sample: The prediction for this object is first evaluated against the TP criterion (IoU > 0.5 with the ground truth). If it qualifies as a TP, its actual IoU value contributes to the gIoU average for that sample. If the prediction's IoU is $\leq 0.5$, or if no object is predicted by the model, the contribution for that sample is 0.
- If the ground truth indicates no target object for the query in a given sample (a "no-target" sample): If the model correctly predicts that no target object is present, the sample's contribution to gIoU is 1. If the model incorrectly predicts an object, the contribution is 0.

This definition ensures that gIoU comprehensively evaluates the model's performance in segmenting correctly identified targets as well as its ability to correctly handle scenarios where the queried object is absent.

# E Theoretical Details

## E.1 Proof of the Cross–Modal MMSE Estimator

This appendix provides the technical details that underpin Eq. (8) in the main text, showing that the *scaled–dot–product cross-attention*

$$\text{Attn}(Q = \mathbf{M}_d,\ K = V = \mathbf{M}_v) \ = \sum_{j=1}^{L_v} \frac{\exp\big(\langle W_Q \mathbf{M}_d,\ W_K \mathbf{m}_{v,j}\rangle / \sqrt{d}\big)}{\sum_p \exp\big(\langle W_Q \mathbf{M}_d,\ W_K \mathbf{m}_{v,p}\rangle / \sqrt{d}\big)}\, \mathbf{m}_{v,j}, \tag{E.1}$$

is the *minimum–mean–square–error* (MMSE) estimator of the latent mediator $\mathbf{M}$ *within* the attention–linear family $\mathcal{F}_{\text{ALF}}$ defined in Sec. 4.2. The derivation follows the general recipe of simplex–constrained quadratic optimization [4] and re-uses the causal decomposition arguments of Perez et al. [41]. Throughout the appendix all expectations are conditional on the observed pair $(\mathbf{M}_d, \mathbf{M}_v)$ unless stated otherwise.

### E.1.1 Problem Set-Up

Let $\mathbf{M} \in \mathbb{R}^d$ be the *latent* mask mediator that triggers the downstream decoder. For a fixed input we observe

$$\mathbf{M}_d \in \mathbb{R}^d, \qquad \mathbf{M}_v = \{\mathbf{m}_{v,1}, \dots, \mathbf{m}_{v,L_v}\} \subset \mathbb{R}^d.$$

Inside the family $\mathcal{F}_{\text{ALF}}$ every candidate estimator is a convex combination $f_{\boldsymbol{\alpha}}(\mathbf{M}_v) = \sum_j \alpha_j \mathbf{m}_{v,j}$ with non-negative coefficients $\boldsymbol{\alpha} \in \Delta^{L_v}$, where $\Delta^{L_v} := \{\alpha_j \geq 0, \ \sum_j \alpha_j = 1\}$ is the probability simplex. The **restricted MMSE problem** is therefore

$$\boldsymbol{\alpha}^\star \ = \ \text{argmin}_{\boldsymbol{\alpha} \in \Delta^{L_v}} \ \mathcal{L}(\boldsymbol{\alpha}), \quad \mathcal{L}(\boldsymbol{\alpha}) := \mathbb{E}\big[\| \, \mathbf{M} - \textstyle\sum_j \alpha_j \mathbf{m}_{v,j} \|_2^2\big]. \tag{E.2}$$

### E.1.2 Quadratic Form of the Risk

Let $\boldsymbol{\mu} := \mathbb{E}[\mathbf{M}]$ be the (unknown) conditional mean of the latent mask. Expanding the square in Eq. (E.2) gives

$$\mathcal{L}(\boldsymbol{\alpha}) \ = \ \underbrace{\|\boldsymbol{\mu}\|^2}_{\text{const}} - 2\, \boldsymbol{b}^\top \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top G \, \boldsymbol{\alpha}, \tag{E.3}$$

where $b_j = \langle \boldsymbol{\mu}, \mathbf{m}_{v,j} \rangle$ and $G_{ij} = \langle \mathbf{m}_{v,i}, \mathbf{m}_{v,j} \rangle$. Because the first term does not depend on $\boldsymbol{\alpha}$, Eq. (E.2) reduces to the *simplex-constrained quadratic program*

$$\min_{\boldsymbol{\alpha} \in \Delta^{L_v}} \big\{ -2\, \boldsymbol{b}^\top \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} \big\}. \tag{E.4}$$

### E.1.3 Lagrange–Multiplier Solution

We solve Eq. (E.2) by introducing a Lagrangian

$$\mathcal{J}(\boldsymbol{\alpha}, \lambda, \boldsymbol{\eta}) \ = \ -2\, \boldsymbol{b}^\top \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} + \lambda \Big( \sum\nolimits_j \alpha_j - 1 \Big) - \boldsymbol{\eta}^\top \boldsymbol{\alpha}, \tag{E.5}$$

where $\lambda \in \mathbb{R}$ and $\boldsymbol{\eta} \in \mathbb{R}^{L_v}_{\geq 0}$ are dual variables that enforce the simplex constraints. Differentiating Eq. (E.2) and using the KKT conditions [4, Ch. 5] yields

$$2\big(G\boldsymbol{\alpha}\big)_j - 2b_j + \lambda - \eta_j = 0, \quad \forall j, \tag{E.6a}$$

$$\eta_j \, \alpha_j = 0, \quad \alpha_j \geq 0, \ \eta_j \geq 0, \tag{E.6b}$$

$$\sum_j \alpha_j = 1. \tag{E.6c}$$

When $G \succ 0$ (true after layer normalization [48]), the solution lies in the *open* simplex—namely $\alpha_j > 0$, implying $\eta_j = 0$. Subtracting the $i$-th and $j$-th rows of Eq. (E.2) cancels $\lambda$ and gives $b_j - b_i = (G\boldsymbol{\alpha})_j - (G\boldsymbol{\alpha})_i$. Re-ordering yields

$$\alpha_j \ = \ \alpha_i \, \exp\!\Big( \tfrac{b_j - b_i - (G\boldsymbol{\alpha})_j + (G\boldsymbol{\alpha})_i}{\tau} \Big), \tag{E.7}$$

where we have inserted an infinitesimal *temperature* $\tau > 0$ for differentiability (the same trick as entropic regularization [9]). Imposing $\sum_j \alpha_j = 1$ converts Eq. (E.2) into the softmax-style fixed-point

$$\alpha_j \ = \ \frac{\exp\!\big( \frac{2b_j - 2(G\boldsymbol{\alpha})_j}{\tau} \big)}{\sum_p \exp\!\big( \frac{2b_p - 2(G\boldsymbol{\alpha})_p}{\tau} \big)}. \tag{E.8}$$

### E.1.4 Isotropic-Token Approximation

After layer normalization, high-dimensional visual tokens $\{\mathbf{m}_{v,j}\}$ are often nearly orthogonal [15, 21], implying $G_{ij} = \langle \mathbf{m}_{v,i}, \mathbf{m}_{v,j} \rangle \approx \gamma_j \delta_{ij}$, where $\gamma_j = \|\mathbf{m}_{v,j}\|_2^2$. Thus, $(G\boldsymbol{\alpha})_j \approx \gamma_j \alpha_j$. The exponent in Eq. (E.8) becomes $(2b_j - 2\gamma_j \alpha_j)/\tau$.

To achieve the common softmax form based on linear scores, we approximate by assuming the cross-interaction terms $b_j = \langle \boldsymbol{\mu}, \mathbf{m}_{v,j} \rangle$ dominate the self-interaction terms $2\gamma_j \alpha_j / \tau$ in determining the relative attention weights. This simplification, common in deriving attention mechanisms [41], effectively neglects the quadratic self-influence terms or treats their impact as uniform, yielding:

$$\alpha_j \ \approx \ \frac{\exp\!\big(2b_j / \tau\big)}{\sum_p \exp\!\big(2b_p / \tau\big)} \ = \ \text{Softmax}_j\!\big(2 \, \langle \boldsymbol{\mu}, \mathbf{m}_{v,j} \rangle / \tau\big). \tag{E.9}$$

This results in scores linear in $\boldsymbol{\mu}$ and $\mathbf{m}_{v,j}$ within the softmax, aligning with standard attention designs.

### E.1.5 Substituting a Geometric Surrogate for $\mu$

The inner products in Eq. (E.9) still involve the *unknown* mean $\boldsymbol{\mu}$. We therefore introduce a linear surrogate driven by the *geometric cue* $\mathbf{M}_d$:

$$\boldsymbol{\mu} \;\approx\; W_Q \mathbf{M}_d, \qquad \mathbf{m}_{v,j} \;\mapsto\; W_K \mathbf{m}_{v,j}, \tag{E.10}$$

a standard choice in multimodal transformers [47]. Choosing $\tau = \sqrt{d}$ converts Eq. (E.9) into the *scaled dot-product* of Vaswani et al. [48], and re-inserting the value tokens $\mathbf{m}_{v,j}$ finally yields Eq. (E.1).

### E.1.6 Consistency as $\tau \to 0$

Let $\boldsymbol{\alpha}^\star_\tau$ be the solution of the regularized optimization with temperature $\tau$. By the $\Gamma$-convergence of entropic regularization [9, Thm. 1], $\boldsymbol{\alpha}^\star_\tau \to \boldsymbol{\alpha}^\star_0$ as $\tau \to 0$, where $\boldsymbol{\alpha}^\star_0$ is the *exact* Euclidean projection solution of Eq. (E.2). Because $W_Q, W_K$ are trainable and $\mathbf{M}_d$ is fed through a temperature-scaling LayerNorm block, the network can approximate arbitrarily small $\tau$ in practice, so the learned weights converge to the optimal MMSE estimator.

### E.1.7 Connecting Back to Causal Front-Door

Finally, note that our estimator (E.1) uses *only* the bias-free query $\hat{\mathbf{M}}_d(x) \approx \mathbb{E}[\mathbf{M}_d \,|\, \mathcal{X} = x]$ and the raw visual tokens $\mathbf{M}_v(x)$. Because $\mathbf{M}_d \perp\!\!\!\perp \mathcal{U}$ by design, the attention weights $\boldsymbol{\alpha}(x)$ are conditionally independent of the confounder, guaranteeing that $\hat{\mathbf{M}}(x) = \mathrm{Attn}(\hat{\mathbf{M}}_d, \mathbf{M}_v)$ is a *front-door-adjusted* proxy for the latent mediator, exactly as required by Perez et al. [41]. □

## E.2 Derivation of the NWGM Approximation

This section provides a derivation for the Normalized Weighted Geometric Mean (NWGM) approximation used in main text Section 4.1 to implement the back-door adjustment for language de-biasing. The goal is to estimate the causal effect $P(\mathcal{Y} \mid \mathrm{do}(\mathcal{T} = t))$.

The back-door adjustment formula (Eq. 1) is:

$$P(\mathcal{Y} \mid \mathrm{do}(\mathcal{T} = t)) = \sum_z P(\mathcal{Y} \mid \mathcal{T} = t, \mathcal{Z} = z) P(\mathcal{Z} = z) = \mathbb{E}_{\mathcal{Z}}\big[P(\mathcal{Y} \mid t, z)\big]. \tag{E.11}$$

In a typical neural network classifier or segmenter, the conditional probability $P(\mathcal{Y} \mid t, z)$ is obtained by applying a $\mathrm{Softmax}$ function to pre-activation scores (logits), denoted as $s_{\mathcal{Y}}(t, z)$. Thus, we are interested in computing:

$$P(\mathcal{Y} \mid \mathrm{do}(\mathcal{T} = t)) = \mathbb{E}_{\mathcal{Z}}\big[\mathrm{Softmax}(s_{\mathcal{Y}}(t, z))\big]. \tag{E.12}$$

The NWGM approximation (Eq. 2) states:

$$\mathbb{E}_{\mathcal{Z}}\left[\,\mathrm{Softmax}\big(s_{\mathcal{Y}}(t, z)\big)\right] \;\approx\; \mathrm{Softmax}\big(\mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Y}}(t, z)]\big). \tag{E.13}$$

To justify this approximation, we first consider the relationship between the arithmetic mean and the weighted geometric mean (WGM). Let $f(\mathcal{Z})$ be a function of a random variable $\mathcal{Z}$ which takes values $z$ with probabilities $P(z)$. The arithmetic mean is:

$$\mathbb{E}_{\mathcal{Z}}[f(z)] = \sum_z f(z) P(z). \tag{E.14}$$

The weighted geometric mean is:

$$\mathrm{WGM}_{\mathcal{Z}}[f(z)] = \prod_z f(z)^{P(z)}. \tag{E.15}$$

For many distributions, especially when the number of samples for $\mathcal{Z}$ is large or $f(z)$ does not vary excessively, the arithmetic mean can be approximated by the WGM:

$$\mathbb{E}_{\mathcal{Z}}[f(z)] \approx \mathrm{WGM}_{\mathcal{Z}}[f(z)]. \tag{E.16}$$

Now, let $f(z) = \exp(g(z))$ for some function $g(z)$. Substituting this into the WGM definition (Eq. (E.15)):

$$\begin{aligned} \mathrm{WGM}_{\mathcal{Z}}[\exp(g(z))] &= \prod_z (\exp(g(z)))^{P(z)} \\ &= \prod_z \exp(g(z)P(z)) \\ &= \exp\left(\sum_z g(z)P(z)\right) \\ &= \exp(\mathbb{E}_{\mathcal{Z}}[g(z)]). \end{aligned} \tag{E.17}$$

Combining Eq. (E.16) and Eq. (E.17):

$$\mathbb{E}_{\mathcal{Z}}[\exp(g(z))] \approx \exp(\mathbb{E}_{\mathcal{Z}}[g(z)]). \tag{E.18}$$

The $\mathrm{Softmax}$ function for a particular class output, given scores $s$, is proportional to $\exp(s_k)$ (where $s_k$ is the score for class $k$). Specifically, $\mathrm{Softmax}_k(s) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$. The NWGM approximation essentially applies the relationship in Eq. (E.18) to the logits *before* the normalization step inherent in Softmax, effectively moving the expectation inside the exponential terms that dominate the Softmax behavior. This leads to the approximation in Eq. (E.13):

$$\mathbb{E}_{\mathcal{Z}}\big[\,\mathrm{Softmax}\big(s_{\mathcal{Y}}(t,z)\big)\big] \;\approx\; \mathrm{Softmax}\big(\mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Y}}(t,z)]\big). \tag{E.19}$$

This approximation is commonly used and has been discussed in works like Baldi and Sadowski [2] and Liu et al. [25].

With the additive score assumption (Section 4.1), $s_{\mathcal{Y}}(t,z) \simeq s_{\mathcal{T}}(t) + s_{\mathcal{Z}}(z)$, the expected score becomes:

$$\begin{aligned}\mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Y}}(t,z)] &= \mathbb{E}_{\mathcal{Z}}[s_{\mathcal{T}}(t) + s_{\mathcal{Z}}(z)] \\ &= s_{\mathcal{T}}(t) + \mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Z}}(z)].\end{aligned} \tag{E.20}$$

Substituting this into Eq. (E.19) gives:

$$P(\mathcal{Y} \mid \mathrm{do}(\mathcal{T}=t)) \approx \mathrm{Softmax}(s_{\mathcal{T}}(t) + \mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Z}}(z)]). \tag{E.21}$$

The de-confounded score used for prediction is therefore $s'_{\mathcal{Y}}(t) = s_{\mathcal{T}}(t) + \mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Z}}(z)]$, as presented in main text Eq. 3. The term $\mathbb{E}_{\mathcal{Z}}[s_{\mathcal{Z}}(z)]$ is practically estimated by averaging the confounder embeddings $f_{\mathcal{Z}}(z_i)$ weighted by their empirical probabilities $P(z_i)$ from the training set, leading to the de-biased text representation $\mathbf{f}'_{\mathcal{T}}(t)$ in main text Eq. 4.

## E.3 Proof of Consistency of the Memory-Bank Estimator

In this appendix we provide a rigorous justification of the *Memory–bank estimator* defined in Eq. 9 of the main text. Recall that, for a fixed time index $t \in \mathbb{N}$, the estimator is

$$\hat{\mathbf{X}}_t \;=\; \sum_{\tau=1}^{W} w_{t,\tau}\,\mathbf{x}_{t-\tau}, \qquad w_{t,\tau} \;=\; \frac{\exp\big(\mathrm{sim}(\mathbf{x}_t, \mathbf{x}_{t-\tau})\big)}{\sum_{\sigma=1}^{W} \exp\big(\mathrm{sim}(\mathbf{x}_t, \mathbf{x}_{t-\sigma})\big)}. \tag{E.22}$$

Here $\mathbf{x}_t \in \mathbb{R}^d$ is the frame-level embedding for $x_t$, $W$ is the memory horizon, and $\mathrm{sim}(\cdot,\cdot)$ a bounded dot-product similarity.

Our goal is to show $\hat{\mathbf{X}}_t$ is a *consistent* estimator of the front–door expectation $\mathbb{E}_{\mathbf{X}'}[\mathbf{X}']$. The proof proceeds in two steps:

1. Show the *unweighted* empirical mean

$$\bar{\mathbf{X}}_t := \tfrac{1}{W} \sum_{\tau=1}^{W} \mathbf{X}_{t-\tau}$$

is unbiased and converges a.s. to $\mathbb{E}_{\mathbf{X}'}[\mathbf{X}']$ under short–range stationarity.

2. Bound the bias introduced by the importance weights $\{w_{t,\tau}\}$ and prove it vanishes as $W \to \infty$.

### E.3.1 Preliminaries and Assumptions

**(A1) Short–range stationarity.** Over the window of size $W$, the sequence $\{\mathbf{X}_{t-\tau}\}_{\tau=1}^{W}$ is wide–sense stationary, i.e. for any lag $\ell$, the joint law of $(\mathbf{X}_{t-\tau}, \mathbf{X}_{t-\tau-\ell})$ does not depend on $\tau$.

**(A2) Finite second moment.** $\mathbb{E}\big[\|\mathbf{X}_0\|_2^2\big] < \infty$.

**(A3) Bounded similarity.** There is $\kappa > 0$ such that $|\mathrm{sim}(\mathbf{x}, \mathbf{y})| \leq \kappa$ almost surely.

### E.3.2 Unweighted Empirical Mean

Under (A1), the past embeddings $\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-W}$ form an i.i.d. sample from the marginal $P_{\mathbf{X}'}$. Hence

$$\mathbb{E}[\bar{\mathbf{X}}_t] = \mathbb{E}_{\mathbf{X}'}[\mathbf{X}'],$$

and by Kolmogorov's strong law of large numbers [16]

$$\bar{\mathbf{X}}_t \xrightarrow[W\to\infty]{\text{a.s.}} \mathbb{E}_{\mathbf{X}'}[\mathbf{X}']. \tag{E.23}$$

### E.3.3 Effect of Importance Weights

Write the weighted estimator as

$$\hat{\mathbf{X}}_t - \bar{\mathbf{X}}_t = \sum_{\tau=1}^{W} \big(w_{t,\tau} - \tfrac{1}{W}\big)\, \mathbf{X}_{t-\tau}.$$

Then

$$\|\hat{\mathbf{X}}_t - \bar{\mathbf{X}}_t\|_2 \;\le\; \max_{\tau} \|\mathbf{X}_{t-\tau}\|_2 \, \big\|\mathbf{w}_t - \tfrac{1}{W}\mathbf{1}\big\|_1, \tag{E.24}$$

where $\mathbf{w}_t = (w_{t,1}, \dots, w_{t,W})$. From (A3) the softmax weights satisfy

$$\frac{e^{-\kappa}}{We^{\kappa}} \;\le\; w_{t,\tau} \;\le\; \frac{e^{\kappa}}{We^{-\kappa}}, \quad \forall \tau,$$

so

$$\big\|\mathbf{w}_t - \tfrac{1}{W}\mathbf{1}\big\|_1 \;\le\; 2\big(e^{2\kappa} - 1\big)\, W^{-1}. \tag{E.25}$$

Combining Eq. (E.24) with Eq. (E.25) and taking expectation gives

$$\mathbb{E}\big[\|\hat{\mathbf{X}}_t - \bar{\mathbf{X}}_t\|_2\big] \;\le\; 2\big(e^{2\kappa} - 1\big)\, W^{-1}\, \mathbb{E}\big[\max_{\tau} \|\mathbf{X}_{t-\tau}\|_2\big], \tag{E.26}$$

where Doob's maximal inequality and (A2) ensure the last expectation is finite. Thus the weight-induced bias decays at rate $\mathcal{O}(W^{-1})$.

### E.3.4 Consistency

By the triangle inequality,

$$\|\hat{\mathbf{X}}_t - \mathbb{E}_{\mathbf{X}'}[\mathbf{X}']\|_2 \le \|\hat{\mathbf{X}}_t - \bar{\mathbf{X}}_t\|_2 + \|\bar{\mathbf{X}}_t - \mathbb{E}_{\mathbf{X}'}[\mathbf{X}']\|_2.$$

The first term converges to 0 in $L^1$ by Eq. (E.26), the second converges a.s. by Eq. (E.23). Hence

$$\hat{\mathbf{X}}_t \xrightarrow[W\to\infty]{\text{prob.}} \mathbb{E}_{\mathbf{X}'}[\mathbf{X}'],$$

establishing consistency. □

**Remark.** In practice $W$ is fixed (e.g. $W = 8$). Eq. (E.26) shows the residual bias is proportional to $(e^{2\kappa} - 1)/W$, and the model can trade off sharpness vs. bias via the learned scale $\kappa$.

### E.3.5 Necessity of Softmax Weights

In Eq. 9 of main paper the memory–bank feature is a convex combination $\hat{\mathbf{X}}_t = \sum_{\tau=1}^{W} w_\tau \, \mathbf{X}_{t-\tau}, \quad \mathbf{w} \in \Delta^{W-1}$. Why choose the weights with a Softmax?

**Entropy–regularized MMSE derivation.** Let $\mathbf{z}^\star = \mathbb{E}[\mathbf{X}' \mid \mathbf{x}_t]$ be the ideal feature we would predict if the full distribution were known. Inside the attention–linear family we minimize the mean-square error while discouraging a single frame from monopolizing the weight:

$$\min_{\mathbf{w}\in\Delta^{W-1}} \Big\|\textstyle\sum_\tau w_\tau \mathbf{X}_{t-\tau} - \mathbf{z}^\star\Big\|_2^2 \;-\; \lambda\, H(\mathbf{w}), \qquad H(\mathbf{w}) = -\textstyle\sum_\tau w_\tau \log w_\tau, \tag{E.27}$$

where $H(\mathbf{w})$ is the Shannon entropy and $\lambda > 0$ is a temperature balancing *accuracy* and *diversity*. Introducing a Lagrange multiplier for $\sum_\tau w_\tau = 1$ and taking derivatives gives

$$w_\tau = \frac{\exp\big(\langle \mathbf{x}_t, \mathbf{x}_{t-\tau}\rangle / \lambda\big)}{\sum_{\sigma=1}^{W} \exp\big(\langle \mathbf{x}_t, \mathbf{x}_{t-\sigma}\rangle / \lambda\big)},$$

which is exactly the Softmax weight with scale $\lambda = \sqrt{d}/\kappa$ used in multi-head attention. Thus Softmax is the *unique* optimum of the entropy-regularized MMSE problem: it gives higher importance to frames more similar to the current one, yet avoids collapsing onto a single frame.

## F More Qualitative Analysis

This section provides further qualitative evidence to support the claims made in the main paper. We first delve into an analysis of the inherent biases present in the training dataset and then showcase additional comparative segmentation results that highlight the superior performance of our CERES framework.

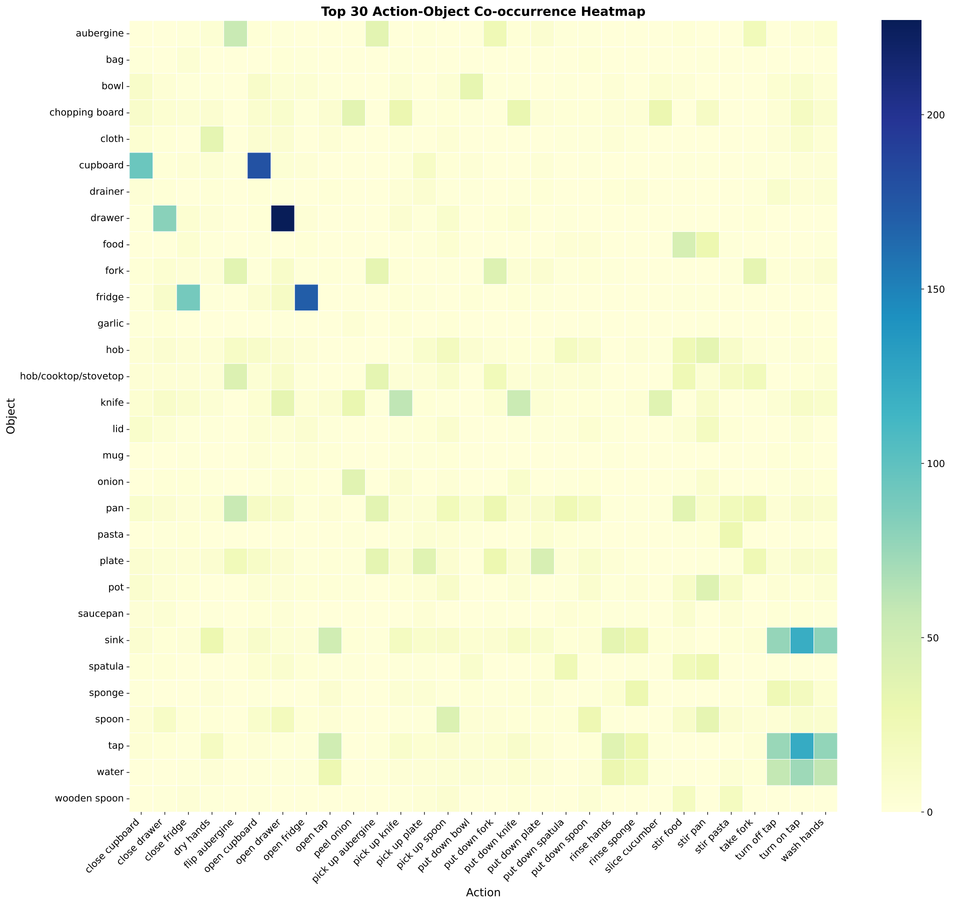


Figure F.1: Co-occurrence frequency of object categories and actions in the VISOR training set. Darker cells indicate more frequent pairings. This visualization clearly shows strong correlations between certain objects and actions, demonstrating the statistical bias our LBD module aims to mitigate.

## F.1 Analysis of Dataset Bias

As discussed in the main paper (Section 1), one of the primary challenges in Ego-RVOS is the presence of dataset biases, where certain object categories frequently co-occur with specific actions. This can lead models to learn spurious correlations rather than truly grounding the language query in the visual scene. To empirically demonstrate this, we conducted a statistical analysis of the action-object pairings within the VISOR training set.

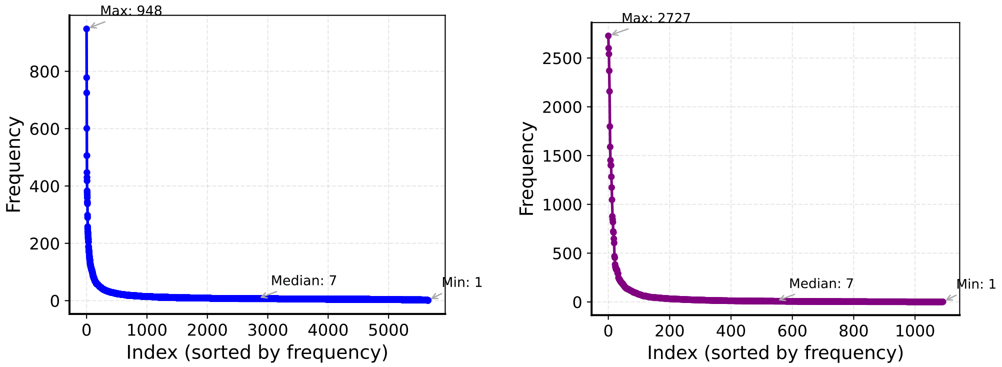
Figure F.2 illustrates the distribution of the most frequent actions (verbs) in the training queries. It is evident that actions such as "cut", "take", and "put" are predominant.

Furthermore, Figure F.1 presents a heatmap (or a co-occurrence matrix analysis) showing the frequency of specific object categories (nouns) appearing with particular actions. This statistical imbalance naturally biases models trained on such data to favor common pairings, potentially failing on queries involving rarer but equally valid combinations.

These distributional statistics underscore the necessity for de-biasing mechanisms like the Linguistic Back-door Deconfounder (LBD) in CERES. By explicitly modeling and adjusting for these confounders (as defined by object-action pairs and their frequencies, see Section 4.1), CERES can achieve more robust language grounding and generalize better to less frequent or novel combinations.

## F.2 More Comparison of Segmentation Results

To further illustrate the robustness and accuracy of CERES, we present additional qualitative comparisons against the strong baseline, ActionVOS [36], on challenging sequences from the VISOR validation set. These examples complement those shown in Figure 4 in the main paper.



(a) Action Distribution

(b) Object Distribution

Figure F.2: Distribution of frequent actions and objects in the VISOR training set queries. This highlights a skewed distribution, where a few actions (objects) are significantly more common.

Table G.1: Performance (%) of CERES (ResNet101 backbone) on VISOR with different depth encoders for DAttn. Results are for the full CERES model. DAv2 refers to Depth Anything V2. The chosen configuration for our main experiments (DAv2 ViT-B) is highlighted. (**Bold** indicates the best performance; underlined indicates the second-best.)

| **Depth Encoder** | **mIoU**$^{\oplus}$↑ | **mIoU**$^{\ominus}$↓ | **gIoU**↑ | **Acc**↑ |
|---|---|---|---|---|
| MiDaS (BEiT-L) | 60.8 | 15.0 | 70.6 | 73.7 |
| DAv2 (ViT-S) | 61.1 | **14.9** | 71.1 | 74.5 |
| DAv2 (ViT-B) | 64.0 | 15.3 | 72.4 | **76.3** |
| DAv2 (ViT-L) | **64.7** | 16.2 | **72.5** | 76.2 |

Figure F.3 showcases scenarios involving (1) significant hand-object occlusion, (2) rapid camera motion leading to motion blur, and (3) objects with subtle state changes.

These additional qualitative results, in conjunction with the quantitative improvements reported in the main paper, reinforce the conclusion that CERES's dual-modal causal intervention strategy effectively addresses key biases and confounding factors, leading to a more robust and reliable Ego-RVOS model. The VFD module enhances resilience to visual distortions common in egocentric video, while the LBD module improves generalization by mitigating reliance on spurious statistical correlations learned from biased training data.

# G More Quantitative Analysis

This section provides additional quantitative analyses to further investigate the components and robustness of our CERES framework. All experiments are conducted on the VISOR dataset using the ResNet101 backbone, unless otherwise specified.

## G.1 Impact of Depth Encoder Choice in DAttn

The Visual Front-door Deconfounder (VFD) utilizes depth features to guide the aggregation of visual semantic features via Depth-guided Attention (DAttn). The quality and nature of these depth features can influence the effectiveness of the mediator construction. To assess this, we evaluated CERES with different pre-trained monocular depth estimation models as the source of depth features. We compared MiDaS (v3.1, BEiT-L backbone) [42] with various sizes of Depth Anything V2 (DAv2) [59] models, specifically those based on ViT-Small (vit-s), ViT-Base (vit-b), and ViT-Large (vit-l). The results are presented in Table G.1.

As shown in Table G.1, the choice of depth encoder impacts performance. Models from the Depth Anything V2 family generally outperform MiDaS in this application. Within the DAv2 family, there is a trend of improved performance with larger model sizes (ViT-S < ViT-B < ViT-L), with DAv2 (ViT-L) achieving the highest mIoU$^{\oplus}$ (64.7%) and gIoU (72.5%). However, DAv2 (ViT-B) provides a strong balance between performance (64.0% mIoU$^{\oplus}$, 72.4% gIoU) and computational cost/model size. Given this trade-off, we selected DAv2 (ViT-B) as the default depth encoder for CERES in our main experiments, as its results are highly competitive while being more resource-efficient than the ViT-L variant. The slightly higher mIoU$^{\ominus}$ for larger models might indicate a

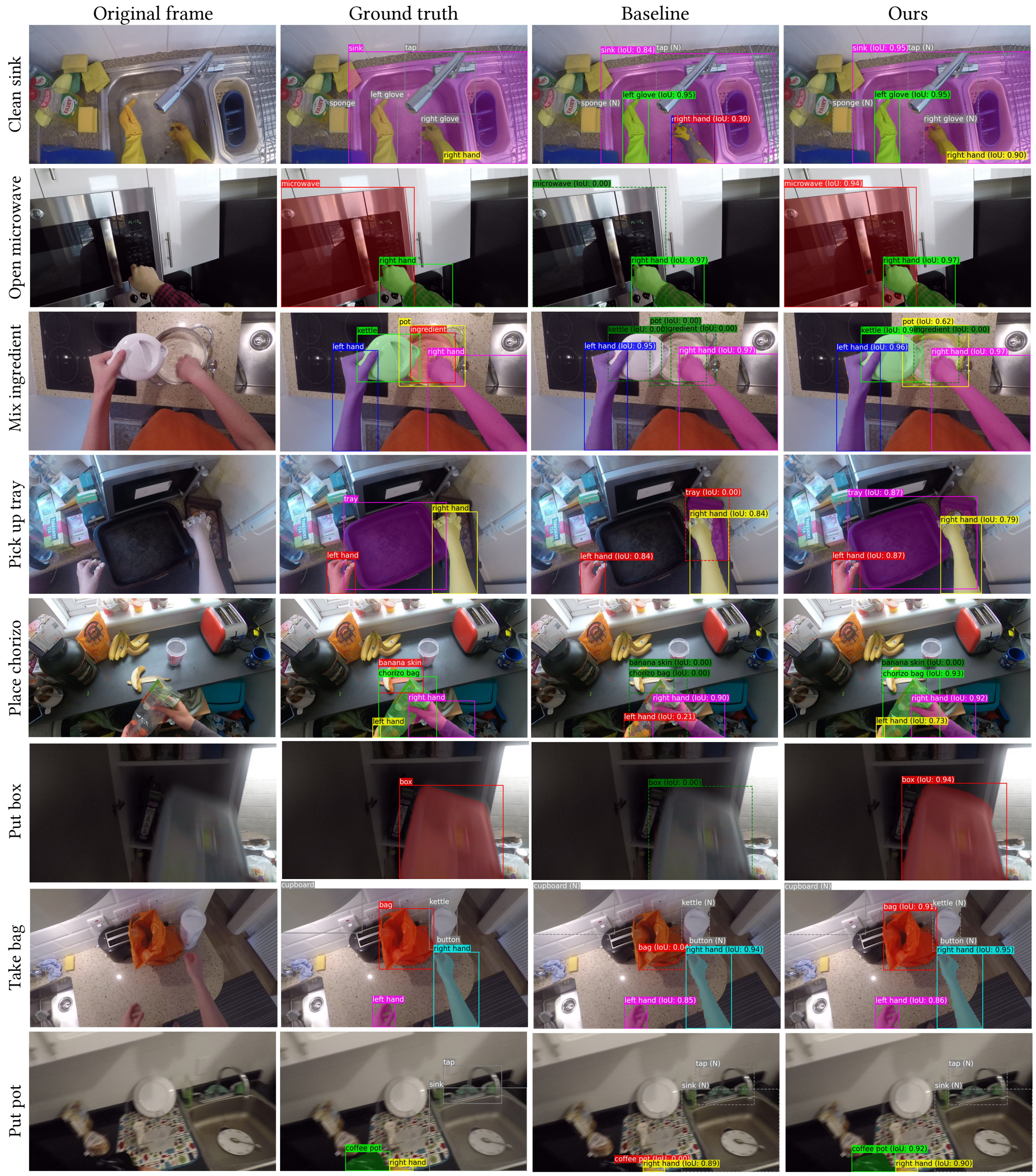


Figure F.3: Additional qualitative comparisons between ActionVOS (Baseline) and CERES on challenging Ego-RVOS scenarios. CERES consistently demonstrates more robust and accurate segmentation in the presence of occlusions, motion blur, and subtle interactions.

more complex feature space that could require further fine-tuning or regularization if negative object suppression is a primary concern. Overall, these results confirm that higher-quality depth information, as provided by more recent and powerful depth estimation models, contributes positively to the VFD module's ability to de-bias visual features.

## G.2 Robustness to Depth Feature Degradation

To further assess the robustness of our DAttn mechanism to imperfections in depth information, we simulated scenarios where depth map quality might be compromised (e.g., due to challenging scenes, sensor noise, or limitations of the depth estimator). We conducted an experiment by adding varying levels of Gaussian noise to the normalized depth features extracted by the DAv2 (ViT-B) encoder before they are fed into the DAttn module. Specifically, for each depth feature vector $f_d$, we added noise $\epsilon \sim \mathcal{N}(0, \sigma_n^2 I)$, where $\sigma_n$ is the standard

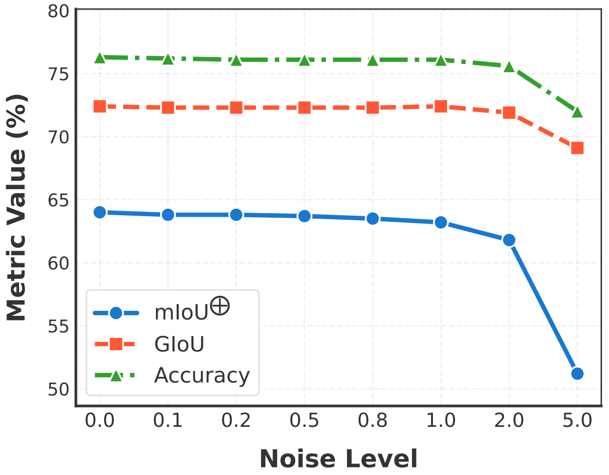


Figure G.1: Performance of CERES (ResNet101, DAv2 ViT-B depth) on VISOR under varying levels of Gaussian noise added to depth features. mIoU$^{\oplus}$ (blue), gIoU (orange), and Accuracy (green) are plotted against the noise standard deviation ($\sigma_n$). ($\sigma_n = 0.0$ means no noise.)

Table G.2: Computational overhead at $448 \times 448$ on RTX 3090. (**Bold** indicates the best; underlined indicates the second-best. Params include the frozen depth encoder (DAv2-B); FPS uses a memory window $W$=5.)

| Method | Backbone | Params (M)↓ | FPS↑ | mIoU$^{\oplus}$↑ |
|---|---|---|---|---|
| ActionVOS | ResNet101 | **195** | **23.8** | 58.4 |
| ActionVOS | VSwin-B | 237 | 15.4 | 62.9 |
| CERES (Ours) | ResNet101 + DAv2-B | 306 | 18.2 | **64.0** |

deviation of the noise. The performance of CERES on VISOR was evaluated across a range of noise levels $\sigma_n \in [0.0, 0.1, 0.2, 0.5, 0.8, 1.0, 2.0, 5.0]$.

The results, depicted in Figure G.1, show a clear trend. For low to moderate noise levels, CERES exhibits notable resilience. Specifically, when the noise standard deviation $\sigma_n$ is less than 1.0 (i.e., for $\sigma_n \leq 0.8$ in our tested discrete levels), the impact on performance is minimal. At $\sigma_n = 0.8$, mIoU$^{\oplus}$ drops from 64.0% to 63.5% (a relative decrease of 0.78%), gIoU drops from 72.4% to 72.3% (a relative decrease of 0.14%), and Accuracy drops from 76.3% to 76.1% (a relative decrease of 0.26%). In all these cases, the relative performance degradation is less than 1% compared to the no-noise baseline. Even at $\sigma_n = 1.0$, gIoU remains remarkably stable (72.4%) and Accuracy only slightly decreases to 76.1%, while mIoU$^{\oplus}$ sees a modest drop to 63.2% (a 1.25% relative decrease).

As the noise intensity increases further (e.g., $\sigma_n = 2.0$ and $\sigma_n = 5.0$), the performance degradation becomes more pronounced, particularly for mIoU$^{\oplus}$, which drops to 61.8% and 51.2% respectively. This indicates that while DAttn can effectively handle minor inaccuracies in depth features, highly corrupted geometric information will naturally lead to a more significant decline in segmentation quality.

This graceful degradation under low to moderate noise levels (with performance loss under 1% for $\sigma_n < 1.0$) demonstrates the robustness of our causally-inspired vision-depth fusion approach. The DAttn mechanism appears capable of leveraging the general structure provided by depth cues even when they are not perfectly accurate, while still underscoring the overall benefit of reasonably high-quality depth information for optimal performance.

### G.3 Overhead Comparison

To quantify the computational overhead introduced by our causal framework relative to established baselines, we report end-to-end throughput and parameter counts at $448 \times 448$ input resolution measured on an NVIDIA RTX 3090. The table below reproduces the setup described in the rebuttal and serves as the basis for overhead comparison.

According to Table G.2, relative to the ResNet101 ActionVOS baseline, our method increases the parameter count by approximately $+56.9\%$ (from 195M to 306M) and reduces throughput by $-23.5\%$ (from $23.8$ to $18.2$ FPS), reflecting the added depth pathway. In contrast, when compared to the stronger VSwin-B ActionVOS baseline, our method carries a smaller parameter overhead of $+29.1\%$ (from 237M to 306M) yet delivers higher throughput by $+18.2\%$ (from $15.4$ to $18.2$ FPS), indicating that the observed efficiency gains stem from the causal design rather than scaling the RGB backbone alone.

Overall, these measurements isolate the overhead attributable to the depth encoder and the lightweight attention modules. While the separate depth pathway increases parameters and reduces FPS versus a ResNet101-only baseline, the framework remains more efficient than upgrading the RGB backbone to Video Swin-B, achieving higher throughput at comparable or better accuracy.