CAUSALPRUNE: COUNTERFACTUAL CAUSAL TOKEN PRUNING FOR EFFICIENT LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) have achieved remarkable progress in multimodal understanding, but their practical deployment is still limited by the substantial computation introduced by excessive visual tokens. Visual token pruning offers an effective solution by removing redundant visual tokens during inference. However, existing methods typically estimate token importance from correlational statistics, which cannot distinguish tokens that causally determine the model output from those spuriously associated with it. To address this issue, we propose **CausalPrune**, a counterfactual causal token pruning framework for efficient LVLMs. We formulate LVLM inference as a Structural Causal Model and introduce the Causal Token Importance Score (CTIS), which estimates the Average Treatment Effect of intervening on each visual token. To make causal estimation practical, we derive an efficient first-order approximation of CTIS and incorporate backdoor adjustment to mitigate latent cross-modal confounding. Within this causal formulation, layer-wise causal mediation decomposes token importance into direct and indirect effects through intermediate hidden states, enabling progressive pruning across transformer layers in a principled manner. Extensive experiments on various LVLM architectures, such as LLaVA-1.5 and Qwen2.5-VL, demonstrate that CausalPrune maintains comparable or even superior multimodal performance while achieving substantial reductions in FLOPs and prefill latency, consistently outperforming existing pruning methods.

1 INTRODUCTION

Recently, Large Vision-Language Models (LVLMs), such as LLaVA (Liu et al., 2023), InternVL (Chen et al., 2024b), and Qwen-VL (Bai et al., 2023), have achieved remarkable progress across a broad spectrum of multimodal tasks, including visual question answering (Zhang et al., 2026; Yang et al., 2026; Xiao et al., 2024), image captioning (Bao et al., 2025; Li et al., 2025a; Ma et al., 2025), and visual reasoning (Xiang et al., 2026; Xie et al., 2025; Bae et al., 2025). These models typically extract visual representations with a visual encoder, integrate them with textual embeddings, and process the resulting multimodal sequence using Large Language Models (LLMs), thereby enabling strong cross-modal understanding and reasoning. Despite their impressive capabilities, practical deployment remains constrained by the substantial memory footprint and computational overhead caused by processing excessive visual tokens (Cao et al., 2026; Zhang et al., 2024b), especially in resource-limited environments.

To address this challenge, prior works (Ye et al., 2025b; Wen et al., 2025; Yang et al., 2025) compress visual representations by pruning redundant tokens, motivated by the higher spatial redundancy of visual information than natural language (Rao et al., 2023; Shi et al., 2023). Early studies on Vision Transformers, such as DynamicViT (Rao et al., 2021), demonstrate that dynamically removing less informative visual tokens can significantly reduce computational overhead while preserving recognition performance. Building upon this idea, recent LVLM pruning methods further explore token reduction under multimodal reasoning scenarios. Existing methods mainly fall into three paradigms: attention-based methods, *e.g.,* FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2024b), and PyramidDrop (Xing et al., 2024), score tokens by cross-modal attention under the assumption that highly attended tokens are more informative. Gradient-based methods, including

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

ZOO-Prune (Kim et al., 2025) and SDTP (Tao et al., 2025), use loss gradients with respect to token embeddings to estimate each token's influence on the output. Diversity-based methods, *e.g.,* DivePrune (Alvar et al., 2025) and CDPruner (Zhang et al., 2025c), retain diverse token subsets to maximize information coverage while removing redundant tokens.

Despite their success, all of these approaches share a critical limitation: they estimate token importance from correlational statistics in the observational distribution. Such metrics cannot distinguish tokens that causally determine the model output from those merely statistically associated with it due to spurious training correlations. As shown in Figure 1, given the query "What color is the car?" and an image of a red car on a beach, a correlation-based method assigns high importance to beach tokens because cars and beaches frequently co-occur in the training distribution, although this correlation is irrelevant to answering the question. The beach tokens are statistically associated with the answer "red" through the confounding variable of scene context, but are not causally responsible for it. *Pruning methods based on such correlational signals may retain spuriously important tokens while discarding causally important ones, resulting in suboptimal accuracy-efficiency trade-offs.* Motivated by this observation, we ask a natural question: *how can we quantify the true causal contribution of each visual token to the model output, rather than relying on spurious correlational signals?*

A principled framework for addressing this fundamental question is causal inference, which provides tools to distinguish causation from correlation through interventions and counterfactual reasoning (Pearl et al., 2003; Pearl, 2009). By modeling how individual variables affect predictions under hypothetical perturbations, causal inference identifies factors that genuinely determine outputs rather than those induced by spurious correlations (Rubin, 1974). This perspective has recently gained traction in vision-language modeling for mitigating undesired correlations, improving explanation faithfulness, uncovering latent generative factors, quantifying multimodal evidence effects, and estimating fine-grained attributions through counterfactual perturbations effectively (Yang et al., 2021; Wang et al., 2025; Zhang et al., 2025a; Gao & Dong, 2025; Li et al., 2025b).

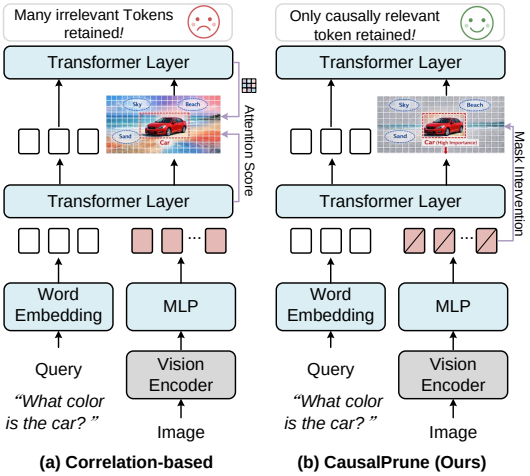


Figure 1: Motivation for CausalPrune. **Left:** Correlational-based scores assign high importance to contextually correlated but causally irrelevant tokens (beach, sky, sand) due to spurious co-occurrence in training data. **Right:** CausalPrune identifies tokens with stronger estimated causal contributions (car region) by estimating the counterfactual effect of removing tokens.

Despite these advances, causal modeling for token-level importance estimation in LVLM pruning remains largely underexplored. This raises an intuitive question: *can causal inference provide a principled token pruning strategy for LVLMs, where token importance is defined by its causal effect on the model output rather than proxy correlational signals?* Addressing this question requires formally characterizing token-level causal effects under latent confounding while maintaining computational tractability for LVLMs in practical deployment scenarios. To this end, we propose **CausalPrune**, which reformulates visual token pruning as a causal inference problem based on the key insight that the question is not only *which tokens does the model attend to?"* but *which tokens have causal contributions to the model's output?"* By modeling LVLM inference as a Structural Causal Model (SCM), we define the Causal Token Importance Score (CTIS) as the Average Treatment Effect (ATE) of intervening on each token, capturing the expected output change when a token is present versus absent. To ensure practical deployment, we derive an efficient first-order approximation of CTIS using gradient information, reducing the cost from multiple forward passes to a single backward pass with small approximation error under mild smoothness conditions. Building on this formulation, we introduce a layer-wise causal mediation perspective that decomposes CTIS into direct and indirect effects through intermediate hidden states, enabling progressive pruning that effectively adapts to each layer's causal contribution.

To conclude, the main contributions of this paper are as follows:

- To the best of our knowledge, we present the first causal token pruning framework for LVLMs, reformulating token importance estimation as a counterfactual causal effect problem via SCM to explicitly distinguish causal contributions from spurious correlations.
- We introduce CTIS, defined as the ATE of intervening on each visual token, together with an efficient gradient-based estimator augmented by backdoor adjustment, enabling causal importance estimation with only a single backward pass.
- We propose a layer-wise causal mediation formulation that decomposes token importance into direct and indirect effects across intermediate hidden states, enabling principled progressive pruning adapted to each layer's causal contribution.
- Extensive experiments across LVLMs and benchmarks show that CausalPrune achieves favorable accuracy–efficiency trade-offs, pruning up to 94.4% visual tokens while preserving comparable or superior performance and reducing FLOPs and prefill latency.

2 PRELIMINARIES

2.1 LVLM ARCHITECTURE AND TOKEN PRUNING SETUP

An LVLM consists of three components: a visual encoder $\mathcal{V}$ (typically a CLIP ViT (Radford et al., 2021)), a projection layer $\mathcal{P}$, and an LLM backbone $\mathcal{L}$. Given an image I and a text query q, the visual encoder produces N visual token embeddings $\mathbf{V} = \{v_1, \ldots, v_N\} \in \mathbb{R}^{N\times d}$, which are projected by $\mathcal{P}$ and concatenated with text token embeddings $\mathbf{T} = \{t_1, \ldots, t_M\} \in \mathbb{R}^{M\times d}$ to form the input sequence $\mathbf{X} = [\mathbf{V}; \mathbf{T}] \in \mathbb{R}^{(N+M)\times d}$ fed into $\mathcal{L}$. The LLM backbone processes $\mathbf{X}$ through L transformer layers, each consisting of multi-head self-attention and feed-forward sublayers (Vaswani et al., 2017), producing output logits $Y = \mathcal{L}(\mathbf{X}) \in \mathbb{R}^{|\mathcal{V}_{\text{vocab}}|}$. The token pruning problem is defined as follows. Given a pruning ratio $\rho \in (0, 1)$, select a subset $\mathcal{S} \subseteq [N]$ with $|\mathcal{S}| = \lfloor \rho N \rfloor$ such that the model output $Y_{\mathcal{S}}$ computed with only the retained tokens is maximally close to the full-token output $Y_{[N]}$. Formally, we seek:

$$\mathcal{S}^* = \arg \min_{\mathcal{S}\subseteq[N],\, |\mathcal{S}|=\lfloor \rho N \rfloor} \left\| Y_{\mathcal{S}} - Y_{[N]} \right\|_2 \tag{1}$$

This is a combinatorial optimization problem that is NP-hard in general, so practical methods rely on importance scores to greedily select the top-k tokens.

2.2 BACKGROUND OF CAUSAL INFERENCE

In this section, we introduce the key concepts of causal inference, including SCMs, interventions, ATE, and the backdoor criterion, which provide the basis for our proposed CausalPrune framework.

Definition 1 (Structural Causal Model). *An SCM $\mathcal{M} = (\mathbf{U}, \mathbf{X}, \mathcal{F}, P(\mathbf{U}))$ consists of: a set of exogenous (background) variables $\mathbf{U}$ with joint distribution $P(\mathbf{U})$; a set of endogenous variables $\mathbf{X}$; and a set of structural equations $\mathcal{F} = \{f_i\}$ where each $X_i = f_i(\mathrm{Pa}(X_i), U_i)$ expresses X_i as a deterministic function of its causal parents $\mathrm{Pa}(X_i)$ and its exogenous noise U_i. The SCM induces a directed acyclic graph (DAG) $\mathcal{G}$ where edges represent direct causal relationships.*

Definition 2 (Do-operator and Average Treatment Effect). *The intervention $\mathrm{do}(X_i = x)$ in SCM $\mathcal{M}$ sets X_i to value x by replacing its structural equation f_i with the constant x, while leaving all other equations unchanged. The resulting interventional distribution is denoted $P(Y \mid \mathrm{do}(X_i = x))$. The ATE of X_i comparing treatment x_i to control x_0 is:*

$$\mathrm{ATE}(X_i;\, x_i, x_0) = \mathbb{E}[Y \mid \mathrm{do}(X_i = x_i)] - \mathbb{E}[Y \mid \mathrm{do}(X_i = x_0)]. \tag{2}$$

Definition 3 (Backdoor Criterion and Adjustment). *A set of variables $\mathbf{Z}$ satisfies the backdoor criterion relative to the ordered pair (X_i, Y) in DAG $\mathcal{G}$ if: (1) no node in $\mathbf{Z}$ is a descendant of X_i, and (2) $\mathbf{Z}$ blocks every path between X_i and Y that contains an arrow into X_i (a "backdoor path"). When $\mathbf{Z}$ satisfies the backdoor criterion, the interventional distribution is identifiable from observational data:*

$$P(Y \mid \mathrm{do}(X_i = x)) = \sum_z P(Y \mid X_i = x, \mathbf{Z} = z)\, P(\mathbf{Z} = z). \tag{3}$$

This expression indicates that, when a valid backdoor adjustment set exists, the causal effect of intervening on X_i can be identified from observational data by conditioning on $\mathbf{Z}$.

3 PROPOSED METHOD: CAUSALPRUNE

3.1 CAUSAL FORMULATION OF LVLM INFERENCE

We model LVLM inference as an SCM $\mathcal{M}_{\text{LVLM}}$ whose causal graph $\mathcal{G}_{\text{LVLM}}$ is depicted in Figure 2. The endogenous variables are the visual tokens V_i ($i = 1, \ldots, N$), the text tokens T_j ($j = 1, \ldots, M$), the layer-l hidden states H^l ($l = 1, \ldots, L$), and the output logits Y. The exogenous variables are the image noise U_I and query noise U_q, which capture all sources of randomness in the image and text generation processes. The structural equations are:

$$V_i = f_V(I, U_I)_i, \qquad T_j = f_T(q, U_q)_j \tag{4}$$

$$H^0 = [V_1, \ldots, V_N, T_1, \ldots, T_M] \tag{5}$$

$$H^l = \text{TransformerLayer}^l(H^{l-1}; \theta^l), \quad l = 1, \ldots, L \tag{6}$$

$$Y = f_{\text{out}}(H^L) \tag{7}$$

where θ^l are the parameters of the l-th transformer layer. The causal graph $\mathcal{G}_{\text{LVLM}}$ contains two important structural features. First, all visual tokens V_i share the common exogenous cause U_I (the image), which induces a confounding path $V_i \leftarrow U_I \rightarrow V_k \rightarrow Y$ for any $k \neq i$. This means that in the observational distribution, V_i and V_k are correlated even if they have no direct causal relationship, and this correlation contributes to observational dependencies captured by the model. Second, the training distribution may introduce spurious associations $V_i \not\perp T_j$ in the observational distribution even when they are independent under intervention, because the training data contains systematic co-occurrences of visual and textual content (*e.g.,* images of beaches tend to co-occur with queries about outdoor activities). We define the CTIS of visual token v_i as the ℓ_2 norm of the ATE of intervening on v_i:

$$\text{CTIS}(v_i) = \left\| \mathbb{E}[Y \mid \text{do}(v_i = v_i^*)] - \mathbb{E}[Y \mid \text{do}(v_i = \mathbf{0})] \right\|_2 \tag{8}$$

where v_i^* is the actual token embedding produced by the visual encoder, and $\mathbf{0}$ represents the masked token (replaced by a learned mask embedding e_{mask}, initialized to the mean token embedding). CTIS measures the causal impact of token v_i on the model output: how much does the output change when we intervene to set v_i to its actual value versus masking it out? Tokens with high CTIS are causally important. Tokens with low CTIS can be safely pruned without significantly affecting the output. Crucially, because CTIS is defined via the do-operator rather than conditioning, it is immune to the confounding bias that afflicts attention-based scores.

Discussion - *Why Attention Scores Are Biased Estimators of CTIS?* Let $A_i^l = \frac{1}{M} \sum_{j=1}^{M} \alpha_{ij}^l$ denote the average attention weight received by visual token v_i at layer l from text tokens, where $\alpha_{ij}^l = \text{softmax}(q_j^l \cdot k_i^l / \sqrt{d})_i$ is the attention weight from text token t_j to visual token v_i. This is the importance metric used by FastV (Chen et al., 2024a) and related methods.

Theorem 1 (Bias of Attention-Based Importance)**.** *Under $\mathcal{M}_{\text{LVLM}}$ with the causal graph $\mathcal{G}_{\text{LVLM}}$, the attention-based importance score A_i^l is a biased estimator of* $\text{CTIS}(v_i)$*:*

$$\mathbb{E}[A_i^l] = \text{CTIS}(v_i) + \underbrace{\sum_{k \neq i} \beta_{ik}^l \cdot \text{Cov}(v_i, v_k)}_{\text{confounding bias } B_i^l} \tag{9}$$

where $\beta_{ik}^l = \partial \mathbb{E}[A_i^l] / \partial \text{Cov}(v_i, v_k)$ measures the sensitivity of the attention score to the spurious correlation between tokens v_i and v_k, induced by the shared exogenous cause U_I and training data statistics.

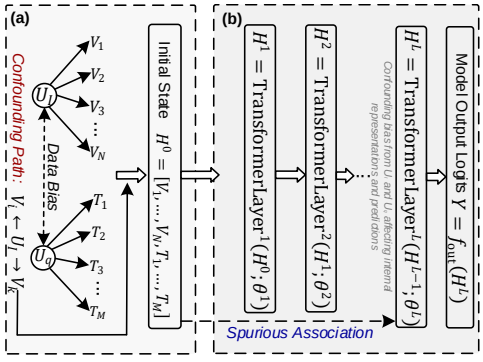


Figure 2: Causal graph $\mathcal{G}_{\text{LVLM}}$ for LVLM inference, illustrating confounding paths and spurious cross-modal associations that motivate causal token-importance estimation.

The proof is given in Appendix B.1.2. The key insight is that the attention keys and queries are learned to minimize training loss, which encodes the joint distribution $P(\mathbf{V}, \mathbf{T}, Y)$ containing spurious correlations. As a result, A_i^l reflects both the causal

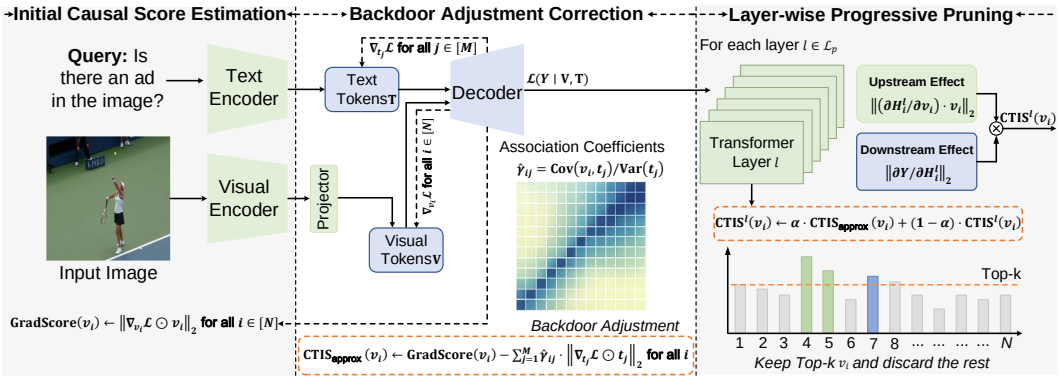


Figure 3: Overview of CausalPrune. CausalPrune estimates visual token importance through a gradient-based approximation of counterfactual causal effects, corrects spurious visual-text associations via backdoor adjustment, and performs progressive layer-wise pruning based on combined global and mediation-based causal scores.

importance of v_i and its spurious associations with other tokens. The confounding bias B_i^l is non-zero whenever $\mathrm{Cov}(v_i, v_k) \neq 0$ for some $k \neq i$, which is generically true for visual tokens from the same image. This theorem provides a formal justification for why attention-based pruning methods are suboptimal: they systematically over-retain tokens that are spuriously correlated with important tokens and under-retain tokens whose causal importance is masked by negative spurious correlations.

3.2 Efficient CTIS Estimation via Gradient-Based Counterfactuals

Exact computation of $\mathrm{CTIS}(v_i)$ via Definition CTIS requires two forward passes per token: one with $v_i = v_i^*$ and one with $v_i = \mathbf{0}$. This amounts to $O(N)$ forward passes, which is N times the standard inference cost and completely impractical. Therefore, we derive an efficient approximation that reduces this to a single backward pass.

First-order Taylor Approximation. Let $\mathcal{L}(Y)$ denote the log-likelihood of the model output (or any differentiable loss function). By Taylor's theorem (Nocedal & Wright, 2006) applied to the perturbation $v_i^* \to \mathbf{0}$:

$$\mathcal{L}(Y \mid \mathrm{do}(v_i = \mathbf{0})) \approx \mathcal{L}(Y \mid v_i^*) - \nabla_{v_i}\mathcal{L}(Y)^\top v_i^* + \tfrac{1}{2}(v_i^*)^\top \nabla_{v_i}^2 \mathcal{L}(Y)\, v_i^* + O(\|v_i^*\|^3) \tag{10}$$

Dropping the second-order term and taking the norm of the first-order change gives the gradient-based importance score:

$$\mathrm{GradScore}(v_i) = \left\| \nabla_{v_i}\mathcal{L}(Y) \odot v_i^* \right\|_2 \tag{11}$$

where $\odot$ denotes element-wise multiplication. This score is computable in a single backward pass through the model, with cost $O(Nd)$—the same order as a single forward pass. However, $\mathrm{GradScore}(v_i)$ still suffers from confounding bias because the gradient $\nabla_{v_i}\mathcal{L}$ is computed under the observational distribution.

Backdoor Adjustment Correction. Under Assumption 3, the text tokens $\mathbf{T}$ constitute a valid adjustment set for the causal effect of V_i on Y. The backdoor-adjusted CTIS is:

$$\mathrm{CTIS}(v_i) = \sum_{\mathbf{t}} \left\| \mathbb{E}[Y \mid \mathrm{do}(v_i = v_i^*), \mathbf{T} = \mathbf{t}] - \mathbb{E}[Y \mid \mathrm{do}(v_i = \mathbf{0}), \mathbf{T} = \mathbf{t}] \right\|_2 \cdot P(\mathbf{T} = \mathbf{t}) \tag{12}$$

The spurious association between visual tokens v_i and textual tokens t_j is quantified by a covariance-based regression coefficient $\hat{\gamma}_{ij} = \mathrm{Cov}(v_i, t_j)/\mathrm{Var}(t_j)$, which is the OLS regression (Ruppert, 2004) coefficient of v_i on t_j. Here, $\hat{\gamma}_{ij}$ represents the regression coefficient estimated from token representations rather than attention weights. Attention maps only characterize observational interactions and are not used as substitutes for causal coefficients. In implementation, token embeddings are first projected into scalar dependency statistics before computing covariance coefficients. This approximation assumes that the local dependency between visual and textual representations can be

Algorithm 1 CausalPrune: Causal Visual Token Pruning

Require: Visual tokens $\mathbf{V} = \{v_i\}_{i=1}^{N}$, text tokens $\mathbf{T} = \{t_j\}_{j=1}^{M}$, pruning ratio $\rho \in (0, 1)$, pruning layers $\mathcal{L}_p \subseteq [L]$, calibration coefficients $\{\hat{\gamma}_{ij}\}$
Ensure: Pruned token index set $\mathcal{S} \subseteq [N]$ with $|\mathcal{S}| = \lfloor \rho N \rfloor$
// Stage 1: Gradient-based importance estimation
Forward pass: compute $\mathcal{L}(Y \mid \mathbf{V}, \mathbf{T})$ and cache all hidden states $\{H^l\}$
Backward pass: compute $\nabla_{v_i}\mathcal{L}$ for all $i \in [N]$ and $\nabla_{t_j}\mathcal{L}$ for all $j \in [M]$
$\mathrm{GradScore}(v_i) \leftarrow \|\nabla_{v_i}\mathcal{L} \odot v_i\|_2$ for all $i \in [N]$
// Stage 2: Backdoor adjustment correction
Retrieve offline-calibrated covariance statistics to estimate $\hat{\gamma}_{ij} = \mathrm{Cov}(v_i, t_j)/\mathrm{Var}(t_j)$
$\mathrm{CTIS}_{\mathrm{approx}}(v_i) \leftarrow \mathrm{GradScore}(v_i) - \sum_{j=1}^{M} \hat{\gamma}_{ij} \cdot \|\nabla_{t_j}\mathcal{L} \odot t_j\|_2$ for all i
// Stage 3: Layer-wise causal progressive pruning
$\mathcal{S} \leftarrow [N]$
for each pruning layer $l \in \mathcal{L}_p$ in increasing order **do**
  Compute $\mathrm{CTIS}^{l}(v_i) = \|\partial Y/\partial H_i^l\|_2 \cdot \|(\partial H_i^l/\partial v_i) \cdot v_i\|_2$ for $i \in \mathcal{S}$
  Combine: $\tilde{\mathrm{CTIS}}^{l}(v_i) \leftarrow \alpha \cdot \mathrm{CTIS}_{\mathrm{approx}}(v_i) + (1-\alpha) \cdot \mathrm{CTIS}^{l}(v_i)$
  $k_l \leftarrow \lfloor \rho^{1/|\mathcal{L}_p|} \cdot |\mathcal{S}| \rfloor$
  $\mathcal{S} \leftarrow \mathrm{TopK}\Big(\{\tilde{\mathrm{CTIS}}^{l}(v_i)\}_{i\in\mathcal{S}},\ k_l\Big)$
end for
return $\mathcal{S}$

captured by a first-order linear expansion around the observed representations, consistent with the Taylor approximation used in gradient-based CTIS estimation. The corrected approximation is:

$$\mathrm{CTIS}_{\mathrm{approx}}(v_i) = \mathrm{GradScore}(v_i) - \underbrace{\sum_{j=1}^{M} \hat{\gamma}_{ij} \cdot \big\|\nabla_{t_j}\mathcal{L}(Y) \odot t_j\big\|_2}_{\text{backdoor correction term}} \tag{13}$$

The correction term removes spurious text-visual association effects from the gradient score. The coefficients $\hat{\gamma}_{ij}$ are estimated offline from calibration statistics of visual and textual token representations and reused during inference, introducing no additional training cost.

Theorem 2 (Approximation Guarantee)**.** *Under Assumptions 1 and 2, the approximation error of* $\mathrm{CTIS}_{\mathrm{approx}}(v_i)$ *before backdoor correction satisfies:*

$$\big|\mathrm{GradScore}(v_i) - \mathrm{CTIS}(v_i)\big| \le \frac{\Lambda \|v_i^*\|^2}{2} + O\big(\|v_i^*\|^3\big) \tag{14}$$

where $\Lambda = \sup_v \|\nabla_{v_i}^2 \mathcal{L}(Y \mid v)\|_2$ *is the Hessian spectral norm bound. For well-trained models operating near a local minimum,* Λ *is small (flat loss landscape), making the approximation tight.*

The proof follows from the Taylor expansion analysis and is provided in Appendix B.1.3.

3.3 Causal Mediation Analysis for Layer-wise Pruning

A visual token may influence the model output not only through a direct path $V_i \to H^1 \to \cdots \to H^L \to Y$, but also through indirect paths mediated by other tokens: $V_i \to H^l \to V_k^{l'} \to \cdots \to Y$ for $k \neq i$. Pruning a token at an early layer may therefore have cascading effects on other tokens' hidden states. To account for this, we apply causal mediation analysis to decompose CTIS into direct and indirect effects, enabling principled layer-wise pruning. For visual token v_i with mediators $\mathbf{M} = \{H_k^l : k \neq i\}$ (the hidden states of other tokens at layer l), the Natural Direct Effect (NDE) and Natural Indirect Effect (NIE) are:

$$\mathrm{NDE}^{l}(v_i) = \mathbb{E}\big[Y(v_i^*, \mathbf{M}(\mathbf{0})) - Y(\mathbf{0}, \mathbf{M}(\mathbf{0}))\big] \tag{15}$$

$$\mathrm{NIE}^{l}(v_i) = \mathbb{E}\big[Y(\mathbf{0}, \mathbf{M}(v_i^*)) - Y(\mathbf{0}, \mathbf{M}(\mathbf{0}))\big] \tag{16}$$

$$\mathrm{TCE}(v_i) = \mathrm{NDE}^{l}(v_i) + \mathrm{NIE}^{l}(v_i) \tag{17}$$

where $Y(x, \mathbf{m})$ denotes the potential outcome when $V_i = x$ and $\mathbf{M} = \mathbf{m}$. The NDE captures the direct causal effect of v_i on Y holding other tokens' hidden states fixed at their counterfactual values (as if v_i were absent). The NIE captures the indirect effect mediated through changes in other tokens' hidden states. The Total Causal Effect (TCE) equals their sum. This decomposition suggests that the influence of a visual token is not confined to a single layer, but can propagate through intermediate hidden states across layers, which motivates our layer-wise pruning design.

Layer-wise CTIS. At layer l, the hidden state H_i^l serves as the mediator for token v_i's influence on subsequent layers. We define the layer-wise CTIS as:

$$\mathrm{CTIS}^l(v_i) = \left\| \frac{\partial Y}{\partial H_i^l} \right\|_2 \cdot \left\| H_i^l - H_i^l|_{v_i=\mathbf{0}} \right\|_2 \tag{18}$$

The first factor $\|\partial Y/\partial H_i^l\|_2$ measures how sensitive the output is to changes in H_i^l (the downstream causal effect), and the second factor $\|H_i^l - H_i^l|_{v_i=\mathbf{0}}\|_2$ measures how much H_i^l changes when v_i is removed (the upstream causal effect). Their product approximates the total causal effect of v_i mediated through layer l. The first factor is computed from the backward pass, and the second factor is approximated using the gradient: $H_i^l - H_i^l|_{v_i=\mathbf{0}} \approx (\partial H_i^l/\partial v_i) \cdot v_i^*$.

Progressive Pruning Schedule. Using $\mathrm{CTIS}^l(v_i)$, we adopt progressive pruning. At each designated pruning layer $l \in \mathcal{L}_p$, tokens with the lowest CTIS^l scores are removed, reducing the sequence length by $\rho^{1/|\mathcal{L}_p|}$ per step. This distributes pruning across layers and lets the model adapt to the reduced token set before the next step, which is empirically better than single-layer pruning.

Algorithm 1 summarizes CausalPrune in three stages: (1) a forward-backward pass to compute gradient scores for visual tokens, (2) backdoor adjustment with calibrated covariance coefficients to remove confounding bias, and (3) layer-wise progressive pruning using CTIS scores. The overhead mainly comes from one backward pass with cost $O(Nd)$ and backdoor correction with cost $O(NM)$ using cached dependency coefficients. This is negligible compared with the $O((N+M)^2 d)$ attention computation saved by pruning a $(1-\rho)$ fraction of visual tokens.

4 Experiments

4.1 Experimental Settings

Implementation Details. We evaluate CausalPrune on LLaVA-1.5-7B/13B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3-8B (Zhu et al., 2025). Following CDPruner (Zhang et al., 2025c), we consider multiple pruning ratios and report performance relative to the unpruned baseline. Based on the hyperparameter analysis in Appendix D.1, CausalPrune uses $|\mathcal{L}_p| = 3$ pruning layers, *i.e.,* layers 8, 16, and 24 for 32-layer models, a calibration size of $n = 128$, and a mixing coefficient of $\alpha = 0.6$. The calibration set is only used for estimating the token dependency coefficients $\hat{\gamma}_{ij}$ in the backdoor adjustment module. The coefficients are estimated separately using token representations from each pruning layer $l \in \mathcal{L}_p$. Specifically, we collect visual and textual token representations from calibration samples and estimate $\hat{\gamma}_{ij}$ using covariance-based regression statistics. The estimated coefficients are cached and reused during inference, without requiring per-example coefficient estimation or additional model training. The mask embedding e_{mask} is initialized as the mean visual token embedding over the calibration set. All experiments are training-free and conducted with PyTorch 2.1 on eight NVIDIA RTX 4090 GPUs, using `lmms-eval` (Zhang et al., 2025b). More details are provided in Appendix D.2.

Benchmarks. We evaluate CausalPrune on a comprehensive broad suite of multimodal benchmarks that cover diverse aspects of image understanding, visual question answering, multimodal reasoning, hallucination evaluation, and text-centric visual comprehension. Specifically, we use 14 benchmarks. These include 10 general VQA and multimodal reasoning benchmarks, namely VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), VizWiz (Gurari et al., 2018), ScienceQA-IMG (Lu et al., 2022), HallBench (Guan et al., 2024), POPE (Li et al., 2023), MME (Fu et al., 2024), MMBench (Liu et al., 2025), MMBench-CN (Liu et al., 2025), and MM-Vet (Yu et al., 2023). We also include 4 text-oriented benchmarks, namely TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), AI2D (Kembhavi et al., 2016), and OCRBench (Liu et al., 2024c), to assess OCR-related and document-level reasoning abilities across various visual scenarios.

Table 1: Main results on Qwen2.5-VL-7B under different visual token budgets. **Acc.** denotes the average benchmark performance, and **Rel.** reports the retained performance relative to the full-token baseline. The best results are highlighted in **bold**.

Method	Venue	TextVQA	ChartQA	AI2D	OCRBench	HallBench	MME	MMB-EN	MMB-CN	Acc.	Rel.
Upper Bound, All 1296 Tokens (100%)											
Qwen2.5-VL-7B	-	84.8	86.1	80.4	863	46.8	2304	82.8	83.2	83.2	100.0%
Keep 512 Tokens (Pruning Ratio 60.5%)											
FastV	ECCV-2024	84.1	82.2	78.8	815	42.4	2317	82.0	81.8	81.1	97.0%
DivPrune	CVPR-2025	81.8	79.6	78.6	800	**43.3**	2279	81.6	82.1	80.1	96.0%
CDPruner	NeurIPS-2025	84.2	82.8	78.9	**827**	42.5	2327	82.2	**82.6**	81.5	97.5%
CausalPrune (Ours)	This paper	**84.5**	**83.2**	**79.3**	826	43.2	**2342**	**82.7**	**82.6**	**81.9**	**98.4%**
Keep 256 Tokens (Pruning Ratio 80.2%)											
FastV	ECCV-2024	81.5	70.9	76.2	703	39.0	2238	79.6	78.9	76.0	90.8%
DivPrune	CVPR-2025	76.0	65.1	76.5	692	36.4	2184	80.0	79.6	74.0	88.2%
CDPruner	NeurIPS-2025	82.4	73.0	77.5	749	**40.1**	2245	80.9	**79.9**	77.6	92.8%
CausalPrune (Ours)	This paper	**84.3**	**75.8**	**78.1**	**760**	38.6	**2310**	**81.7**	78.0	**78.5**	**94.4%**
Keep 128 Tokens (Pruning Ratio 90.1%)											
FastV	ECCV-2024	73.8	52.2	71.4	531	33.8	2008	72.9	72.2	66.2	79.0%
DivPrune	CVPR-2025	67.0	50.4	72.1	549	32.6	2108	**77.8**	**77.8**	67.3	79.9%
CDPruner	NeurIPS-2025	77.8	59.2	74.0	632	37.2	2127	76.2	76.5	71.3	85.2%
IDPruner	arXiv-2026	74.4	62.5	75.2	625	38.0	2114	74.7	71.7	70.6	84.8%
CausalPrune (Ours)	This paper	**81.3**	**66.8**	**75.5**	**642**	**40.2**	**2164**	**77.8**	77.2	**73.9**	**88.8%**

Comparison Methods. We compare CausalPrune with representative visual token pruning methods from three paradigms. The attention-based methods include FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2024b), PyramidDrop (PDrop) (Xing et al., 2024), TRIM (Song et al., 2025), VisionZip (Yang et al., 2025), and FlashVLM (Cai et al., 2025). The gradient-based methods include ZOO-Prune (Kim et al., 2025) and FitPrune (Ye et al., 2025a). The diversity-based methods include PruMerge+ (Shang et al., 2025), DART (Wen et al., 2025), DivPrune (Alvar et al., 2025), CDPruner (Zhang et al., 2025c), IDPruner (Tan et al., 2026), and HoloV (Zou et al., 2026). For baselines evaluated in CDPruner (Zhang et al., 2025c), we report the corresponding published results to maintain consistency with the established evaluation setting. All methods follow official settings and metrics with unified token budgets and protocols for fair comparison.

4.2 Main Results and Analysis

Main-paper results focus on Qwen2.5-VL-7B, with additional evaluations on LLaVA-1.5-7B, LLaVA-1.5-13B, LLaVA-NeXT-7B, InternVL3-8B, and video understanding benchmarks using Qwen2.5-VL-7B provided in Appendices D.3, D.4, D.5, D.6, and D.7.

Comparisons on Qwen2.5-VL-7B. We evaluate CausalPrune on Qwen2.5-VL-7B to verify generalization to advanced dynamic-resolution architectures. As shown in Table 1, CausalPrune achieves the best average performance across token budgets. With 512 retained tokens, it preserves 98.4% of the full-token performance and slightly outperforms CDPruner and FastV. When the budget is reduced to 256 tokens, CausalPrune retains 94.4% relative performance, exceeding CDPruner by 1.6 percentage points. Under the most aggressive 128-token setting, CausalPrune maintains 88.8% of the original performance and improves over CDPruner by 2.6 average points. The gains on TextVQA, ChartQA, MME, and MMBench further show that counterfactual causal importance remains effective for dynamic-resolution architectures with compact visual representations.

Efficiency Analysis. We evaluate inference efficiency on the POPE benchmark using LLaVA-NeXT-7B. As shown in Table 4 in Appendix D, reducing the number of visual tokens from 2,880 to 320 significantly lowers the computational cost. Compared with the full-token baseline, CausalPrune reduces FLOPs from 41.7T to 3.8T, achieving nearly an 11$\times$ reduction, while decreasing prefill latency from 246 ms/token to 32 ms/token. It also reduces decode latency from 29 to 20 ms/token and KV cache from 1440 MB to 160 MB. Compared with strong pruning baselines under the same 320-token budget, CausalPrune achieves the lowest FLOPs and latency, while obtaining the best POPE F1 score of 87.8. Notably, it also slightly surpasses the unpruned LLaVA-NeXT-7B baseline in F1, suggesting that causal token selection can remove redundant or distracting visual tokens without sacrificing hallucination robustness. Additional online overhead is analyzed in Appendix D.8.

Qualitative Visualization. We visualize the progressive pruning behavior of CausalPrune on LLaVA-1.5-7B using representative VQA examples, as shown in Figure 4. Each row presents the original image together with the visual tokens retained after pruning at Layers 8, 16, and 24. Across diverse queries, CausalPrune progressively removes redundant background tokens while maintaining the compact visual evidence needed to answer each question. For object identification, the retained

Table 2: Ablation study of CausalPrune on LLaVA-1.5-7B with 128 retained visual tokens. The best results are highlighted in **bold**.

Method	VQA^{V2}	GQA	VizWiz	SQA^{IMG}	VQA^{Text}	POPE	MME	MMB^{EN}	MMB^{CN}	MMVet	Acc.	Rel.
Keep 128 Tokens (**Pruning Ratio 77.8%**)												
GradScore only (no causal)	75.4	58.5	52.6	68.1	56.8	87.1	1422.8	61.3	55.2	32.1	62.0	97.6%
w/o Backdoor Correction	76.0	59.1	53.0	68.6	57.3	87.6	1428.6	62.2	55.7	32.6	62.5	98.9%
w/o Layer-wise Pruning	76.3	59.4	53.4	68.9	57.6	87.9	1433.2	62.5	56.3	32.9	62.8	99.2%
w/o Mediation Analysis	76.6	59.6	53.6	69.2	57.9	88.2	1437.5	62.7	56.6	33.2	63.0	99.4%
Full CausalPrune (Ours)	**76.8**	**59.9**	**53.9**	**69.4**	**58.1**	**88.4**	**1444.7**	**62.9**	**56.8**	**33.4**	**63.2**	**99.7%**

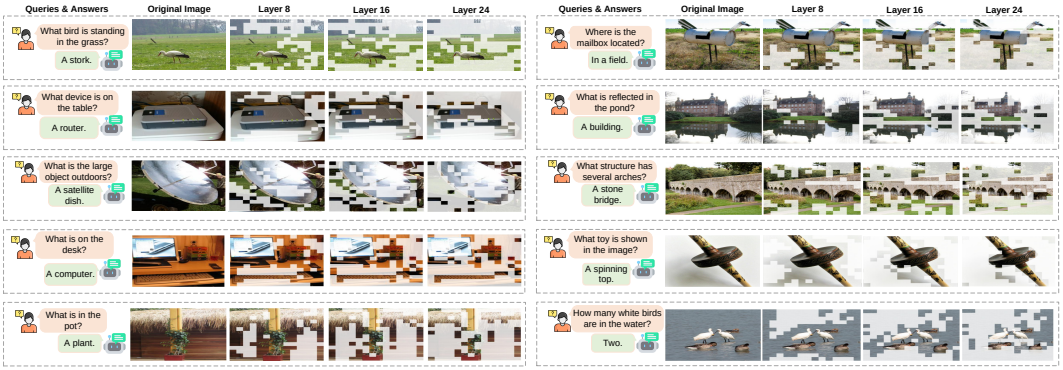


Figure 4: Qualitative visualization of CausalPrune on LLaVA-1.5-7B. Each case shows the original image and retained visual tokens after progressive pruning at Layers 8, 16, and 24.

tokens concentrate on the stork, router, satellite dish, computer, potted plant, and spinning top. For scene-grounded questions, the method also preserves contextual regions such as the grass around the stork, the table supporting the router, and the field surrounding the mailbox. It retains the building and its reflection in the pond for the reflection query, preserves the arch regions of the stone bridge for the structural question, and keeps both white birds for counting. These examples indicate that CausalPrune preserves both the queried entities and the minimal supporting context required for prediction, producing interpretable token selections even under aggressive visual token reduction.

4.3 Ablation Studies

We conduct component ablations on LLaVA-1.5-7B under the 128-token setting. As shown in Table 2, the gradient-based score alone achieves 62.0 average performance, serving as a strong first-order baseline. Adding the causal formulation further improves the score to 62.5, indicating the benefit of counterfactual token importance estimation beyond purely correlational signals. Layer-wise pruning consistently raises the score to 62.8, suggesting that token importance should be dynamically updated across transformer depths rather than being estimated in a single pass. Mediation analysis further increases the score to 63.0, confirming the value of explicitly modeling indirect effects through intermediate hidden states. The full CausalPrune achieves the best result, with 63.2 average performance and 99.7% relative performance, demonstrating that backdoor correction, layer-wise pruning, and mediation-based causal estimation are complementary and jointly contribute to preserving task-critical visual evidence under aggressive token reduction.

5 Conclusion

In this paper, we present CausalPrune, a counterfactual causal token pruning framework for efficient LVLMs. CausalPrune formulates visual token pruning as a causal effect estimation problem and introduces the CTIS to quantify the counterfactual contribution of each visual token to the model output. To enable practical causal estimation, we derive an efficient first-order approximation with backdoor adjustment and incorporate layer-wise causal mediation for progressive pruning across transformer layers. Extensive experiments on multiple LVLM architectures and vision-language benchmarks demonstrate that CausalPrune preserves strong multimodal performance while substantially reducing inference cost, consistently outperforming existing pruning methods.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9392–9401, 2025.

Inhwan Bae, Junoh Lee, and Hae-Gon Jeon. Social reasoning-aware trajectory prediction via multimodal language model. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and J Qwen-VL Zhou. A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 6:3, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Qianyue Bao, Fang Liu, Licheng Jiao, Yang Liu, Shuo Li, Lingling Li, Xu Liu, and Puhua Chen. Visual-language scene-relation-aware zero-shot captioner. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Kaitong Cai, Jusheng Zhang, Jing Yang, Yijia Fan, Pengtao Xie, Jian Wang, and Keze Wang. Flashvlm: Text-guided visual token selection for large multimodal models. *arXiv preprint arXiv:2512.20561*, 2025.

Jianjian Cao, Chong Yu, Peng Ye, and Tao Chen. Madtp++: Bridge the gap between token and weight pruning for accelerating vlts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Ragnar Frisch and Frederick V Waugh. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pp. 387–401, 1933.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Mengyu Gao and Qiulei Dong. Causality-guided prompt learning for vision-language models via visual granulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1141–1151, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Youngeun Kim, Youjia Zhang, Huiling Liu, Aecheon Jung, Sunwoo Lee, and Sungeun Hong. Training-free token pruning via zeroth-order gradient estimation in vision-language models. *arXiv preprint arXiv:2509.24837*, 2025.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308. IEEE, 2024.

Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. Treble counterfactual vlms: A causal approach to hallucination. *arXiv preprint arXiv:2503.06169*, 2025b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2025.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024c.

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/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Yiwei Ma, Weihuang Lin, Zhibin Wang, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Boosting multi-modal large language model with enhanced visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl et al. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

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

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Yongming Rao, Zuyan Liu, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dynamic spatial sparsification for efficient vision transformers and convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10883–10897, 2023.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

David Ruppert. The elements of statistical learning: data mining, inference, and prediction, 2004.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1181–1198, 2023.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7614–7623, 2025.

Yifan Tan, Yifu Sun, Shirui Huang, Hong Liu, Guanghua Yu, Jianchen Zhu, and Yangdong Deng. Idpruner: Harmonizing importance and diversity in visual token pruning for mllms. *arXiv preprint arXiv:2602.13315*, 2026.

Yao Tao, Yehui Tang, Yun Wang, Mingjian Zhu, Hailin Hu, and Yunhe Wang. Saliency-driven dynamic token pruning for large language models. *arXiv preprint arXiv:2504.04514*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. Causalrag: Integrating causal graphs into retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22680–22693, 2025.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9972–9991, 2025.

Kun Xiang, Zhili Liu, Terry Jingchen Zhang, Yinya Huang, Yunshuang Nie, Kaixin Cai, Yiyang Yin, Runhui Huang, Hanhui Li, Yihan Zeng, et al. Atomthink: Multimodal slow thinking with atomic step reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

Chi Xie, Shuang Liang, Jie Li, Zhao Zhang, Feng Zhu, Rui Zhao, and Yichen Wei. Relationlmm: Large multimodal model as open and versatile visual relationship generalist. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Jinhui Yang, Ming Jiang, and Qi Zhao. Defying distractions in multimodal tasks: A novel benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9593–9602, 2021.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025a.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24972–24982, 2025b.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024a.

Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm: Counterfactual vision-language fine-tuning. *arXiv preprint arXiv:2506.17267*, 2025a.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025b.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025c.

YiFan Zhang, Qingsong Wen, Chaoyou Fu, Kun Wang, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024b.

Jiawei Zhu, Wei Chen, and Ruichu Cai. Causalagent: A conversational multi-agent system for end-to-end causal inference. *arXiv preprint arXiv:2602.11527*, 2026.

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

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *Advances in Neural Information Processing Systems*, 38:39800–39832, 2026.

A The Use of Large Language Models (LLMs)

In the preparation of this manuscript, the authors utilize an LLM, specifically GPT-5.6, solely to assist with language polishing. The LLM is used only to improve the grammar, clarity, and readability of the manuscript, without altering its technical content, methodology, results, or scientific contributions. All ideas, analyses, results, and interpretations presented in this paper are solely those of the authors. The LLM is not used to generate scientific content, conduct experiments, perform analyses, or formulate the technical contributions of this work.

B Theoretical Analysis

We now state the formal assumptions and prove the theoretical guarantees of CausalPrune. Full proofs are given in Appendix B.1.

Assumption 1 (Smoothness)**.** *The loss function $\mathcal{L}(Y \mid \mathbf{V}, \mathbf{T})$ is twice continuously differentiable with respect to $\mathbf{V}$, with bounded Hessian spectral norm: $\|\nabla^2_{v_i}\mathcal{L}\|_2 \leq \Lambda < \infty$ for all $i \in [N]$ and all $\mathbf{V}$ in a neighborhood of $\mathbf{V}^*$.*

Assumption 2 (Causal Faithfulness)**.** *Every conditional independence in the observational distribution $P(\mathbf{V}, \mathbf{T}, Y)$ corresponds to a d-separation in the causal graph $\mathcal{G}_{\mathrm{LVLM}}$. Equivalently, no causal path has zero effect by coincidental cancellation.*

Assumption 3 (Backdoor Validity)**.** *The text tokens $\mathbf{T}$ satisfy the backdoor criterion relative to (V_i, Y) for all $i \in [N]$ in the causal graph $\mathcal{G}_{\mathrm{LVLM}}$.*

Assumption 3 holds in standard LVLM architectures where visual and text tokens interact only through the LLM's self-attention layers (not through the visual encoder), because the text tokens are not descendants of any visual token and they block all backdoor paths from V_i to Y through the shared image statistics U_I.

Theorem 3 (Asymptotic Unbiasedness of CTIS)**.** *Under Assumptions 1–3, as the number of calibration samples $n \to \infty$, the estimator $\mathrm{CTIS}_{\mathrm{approx}}(v_i)$ converges in probability to $\mathrm{CTIS}(v_i)$ for all $i \in [N]$ simultaneously:*

$$\mathrm{CTIS}_{\mathrm{approx}}(v_i) \xrightarrow{p} \mathrm{CTIS}(v_i) \quad \text{as } n \to \infty, \quad \forall\, i \in [N] \tag{19}$$

Theorem 4 (Finite-Sample Error Bound)**.** *Under Assumptions 1–3, with probability at least $1-\delta$ over the randomness of the calibration set, for all $i \in [N]$ simultaneously:*

$$\left|\mathrm{CTIS}_{\mathrm{approx}}(v_i) - \mathrm{CTIS}(v_i)\right| \leq \frac{\Lambda \|v_i^*\|^2}{2} + C\sqrt{\frac{\log(2N/\delta)}{n}} \tag{20}$$

where $C = 2\sigma_{\max}(\mathbf{J}) \cdot \|\mathbf{T}\|_F / \lambda_{\min}(\mathbf{T}^\top \mathbf{T})$ depends on the Jacobian $\mathbf{J} = \partial Y / \partial \mathbf{T}$, the text token matrix $\mathbf{T}$, and n is the number of calibration samples used to estimate $\hat{\gamma}_{ij}$.

The bound in Theorem 4 has two terms: the first is the Taylor approximation error (controlled by the Hessian), and the second is the statistical estimation error in the backdoor correction (decreasing as $O(1/\sqrt{n})$). In practice, we find that $n = 128$ calibration samples is sufficient to make the second term negligible.

Proposition 1 (Pruning Consistency)**.** *Let $\mathcal{S}^* = \mathrm{TopK}(\{\mathrm{CTIS}(v_i)\}_{i=1}^N, k)$ be the oracle pruning set based on true CTIS values. Let ε denote the right-hand side of Eq.* (20)*. If the following CTIS gap condition holds:*

$$\Delta_{\mathrm{gap}} = \min_{i \in \mathcal{S}^*} \mathrm{CTIS}(v_i) - \max_{j \notin \mathcal{S}^*} \mathrm{CTIS}(v_j) > 2\varepsilon \tag{21}$$

then CausalPrune recovers the oracle pruning set exactly: $\mathcal{S} = \mathcal{S}^*$ with probability at least $1-\delta$. In contrast, attention-based methods have no such guarantee due to the non-vanishing confounding bias B_i^l in Theorem 1.

The gap condition (21) is a standard identifiability condition in top-k selection problems. It requires that the k-th and $(k+1)$-th tokens have sufficiently different CTIS values that our estimation error cannot cause a misranking. This condition is generically satisfied when the pruning ratio is not too aggressive (*i.e.,* ρ is not too close to 0 or 1).

B.1 Formal Derivations and Detailed Proofs

B.1.1 SCM Identifiability and the Backdoor Criterion for LVLMs

We establish that the SCM underlying the LVLM architecture satisfies the conditions required for the backdoor criterion to hold, thereby justifying the identification formula used throughout the paper.

The LVLM Causal DAG. We model the LVLM as a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V} = \{U_I, U_T, V_1, \ldots, V_N, T_1, \ldots, T_M, H^1, \ldots, H^L, Y\}$, where U_I is the latent image-level confounder, U_T is the latent text-level confounder, V_i are visual tokens, T_j are text tokens, H^l is the joint hidden state at layer l, and Y is the output. The structural equations are:

$$\begin{aligned} V_i &= f_V(I, U_I, \varepsilon_i^V), \quad i = 1, \ldots, N \\ T_j &= f_T(Q, U_T, \varepsilon_j^T), \quad j = 1, \ldots, M \\ H^l &= f_H^l(H^{l-1}, \theta^l), \quad l = 1, \ldots, L \\ Y &= f_Y(H^L, \varepsilon^Y) \end{aligned} \tag{22}$$

where I is the input image, Q is the query text, θ^l are the transformer parameters at layer l, and $\varepsilon_i^V, \varepsilon_j^T, \varepsilon^Y$ are independent noise terms.

Lemma 1 (Acyclicity of the LVLM DAG)**.** *The DAG $\mathcal{G}$ defined by the structural equations* (22) *is acyclic by construction.*

The proof follows by observing that the structural equations define a strict topological ordering: $U_I, U_T \prec V_i, T_j \prec H^1 \prec H^2 \prec \cdots \prec H^L \prec Y$. Since each variable is a function only of variables earlier in this ordering, no directed cycle can exist. Formally, define the level function $\ell : \mathcal{V} \to \mathbb{N}$ by $\ell(U_I) = \ell(U_T) = 0$, $\ell(V_i) = \ell(T_j) = 1$, $\ell(H^l) = l + 1$, and $\ell(Y) = L + 2$. Every edge $(A, B) \in \mathcal{E}$ satisfies $\ell(A) < \ell(B)$, so no directed cycle exists. □

Lemma 2 (Backdoor Admissibility of Text Tokens). *The set* $\mathbf{T} = \{T_1, \ldots, T_M\}$ *of text tokens satisfies the backdoor criterion relative to the pair* (V_i, Y) *in* $\mathcal{G}$.

To verify the backdoor criterion, we must show that (i) no node in $\mathbf{T}$ is a descendant of V_i, and (ii) $\mathbf{T}$ blocks every backdoor path from V_i to Y. For condition (i): since $\ell(T_j) = 1 = \ell(V_i)$ and there is no edge from V_i to T_j in $\mathcal{G}$ (the visual and text encoders are separate), T_j is not a descendant of V_i. For condition (ii): every backdoor path from V_i to Y must pass through U_I (the only common cause of V_i and Y not on the causal path $V_i \to H^1 \to \cdots \to Y$). Any such path has the form $V_i \leftarrow U_I \to V_k \to H^1 \to \cdots \to Y$ for some $k \neq i$. Conditioning on $\mathbf{T}$ does not open any collider on this path, and since $\mathbf{T}$ is independent of U_I given the query Q (the text tokens are generated from Q and U_T, which is independent of U_I by assumption), conditioning on $\mathbf{T}$ blocks the path through U_I via the d-separation criterion. □

Residual Confounding Formula. We now derive the formula for the residual covariance between visual tokens after conditioning on text tokens. Let $\mathbf{T}$ denote the vector of all text token embeddings stacked as a matrix. The partial covariance between V_i and V_k given $\mathbf{T}$ is:

$$\mathrm{Cov}(V_i, V_k \mid \mathbf{T}) = \mathrm{Cov}(V_i, V_k) - \mathrm{Cov}(V_i, \mathbf{T})\,[\mathrm{Var}(\mathbf{T})]^{-1}\,\mathrm{Cov}(\mathbf{T}, V_k) \tag{23}$$

This is the standard formula for partial covariance derived from the Schur complement of the joint covariance matrix. Under Assumption 3, the confounders U_I affect V_i and V_k but are independent of $\mathbf{T}$. Therefore, the component of $\mathrm{Cov}(V_i, V_k)$ attributable to U_I is not captured by $\mathrm{Cov}(V_i, \mathbf{T})[\mathrm{Var}(\mathbf{T})]^{-1}\mathrm{Cov}(\mathbf{T}, V_k)$, and the residual $\mathrm{Cov}(V_i, V_k \mid \mathbf{T})$ represents the confounding that remains after conditioning on text. However, since U_I is a common cause of all visual tokens but is independent of $\mathbf{T}$, the backdoor adjustment via $\mathbf{T}$ removes the spurious association between V_i and Y that flows through $U_I \to V_k \to Y$, even though $\mathrm{Cov}(V_i, V_k \mid \mathbf{T}) \neq 0$ in general. The key insight is that the backdoor criterion requires blocking the backdoor paths, not eliminating all correlations among visual tokens.

Remark 1. *The formula* (23) *shows that* $\mathrm{Cov}(V_i, V_k \mid \mathbf{T}) = 0$ *if and only if the covariance between* V_i *and* V_k *is entirely mediated through* $\mathbf{T}$. *In the LVLM setting, this holds approximately when the visual tokens are conditionally independent given the text context, which is a stronger assumption than what we actually require here. Our results only need the backdoor criterion (Lemma 2), rather than full conditional independence.*

B.1.2 Full Proof of Theorem 1 (Bias of Attention-Based Importance)

We provide a complete and rigorous proof of Theorem 1, establishing that attention-based importance scores are biased estimators of CTIS. The proof proceeds by a second-order Taylor expansion of the softmax function, followed by an application of the Frisch-Waugh-Lovell theorem to characterize the confounding bias term.

Setup and Notation. Let $\phi_i^l = W_K^l v_i \in \mathbb{R}^{d_k}$ be the key vector for visual token v_i at layer l, and let $\psi_j^l = W_Q^l t_j \in \mathbb{R}^{d_k}$ be the query vector for text token t_j. Define the pre-softmax logit $s_{ij}^l = \psi_j^l \cdot \phi_i^l / \sqrt{d_k} = (W_Q^l t_j)^\top (W_K^l v_i)/\sqrt{d_k}$. The attention weight is $\alpha_{ij}^l = \exp(s_{ij}^l)/Z_j^l$ where $Z_j^l = \sum_{k=1}^N \exp(s_{kj}^l) + \sum_{m=1}^M \exp(s_{mj}^{l,T})$ is the partition function, with $s_{mj}^{l,T} = (W_Q^l t_j)^\top (W_K^l t_m)/\sqrt{d_k}$ denoting the text-to-text logits. The attention-based importance score is $A_i^l = \frac{1}{M}\sum_{j=1}^M \alpha_{ij}^l$.

Second-Order Taylor Expansion of the Softmax. We expand α_{ij}^l as a function of the key-query inner products. Let $\bar{s}_j^l = \frac{1}{N+M}\sum_k s_{kj}^l$ denote the mean logit. Writing $s_{ij}^l = \bar{s}_j^l + \delta_{ij}^l$, where $\delta_{ij}^l = s_{ij}^l - \bar{s}_j^l$ is the deviation, the softmax satisfies:

$$\begin{aligned}\alpha_{ij}^l &= \frac{\exp(\bar{s}_j^l + \delta_{ij}^l)}{\sum_k \exp(\bar{s}_j^l + \delta_{kj}^l)} = \frac{\exp(\delta_{ij}^l)}{\sum_k \exp(\delta_{kj}^l)} \\ &= \frac{1 + \delta_{ij}^l + \frac{1}{2}(\delta_{ij}^l)^2 + O((\delta_{ij}^l)^3)}{\sum_k \left[1 + \delta_{kj}^l + \frac{1}{2}(\delta_{kj}^l)^2 + O((\delta_{kj}^l)^3)\right]}\end{aligned} \tag{24}$$

Let $\bar{\delta}_j^l = \frac{1}{N+M}\sum_k \delta_{kj}^l = 0$ by definition, and let $\sigma_j^{l,2} = \frac{1}{N+M}\sum_k (\delta_{kj}^l)^2$ be the second moment of the deviations. The denominator in (24) equals $(N+M)(1+\frac{1}{2}\sigma_j^{l,2} + O(\|\boldsymbol{\delta}_j^l\|_\infty^3))$. Therefore:

$$\begin{aligned}\alpha_{ij}^l &= \frac{1}{N+M}\left(1+\delta_{ij}^l + \frac{1}{2}(\delta_{ij}^l)^2 - \frac{1}{2}\sigma_j^{l,2}\right) + O(\|\boldsymbol{\delta}_j^l\|_\infty^3) \\ &= \frac{1}{N+M} + \frac{\delta_{ij}^l}{N+M} + \frac{(\delta_{ij}^l)^2 - \sigma_j^{l,2}}{2(N+M)} + O(\|\boldsymbol{\delta}_j^l\|_\infty^3) \end{aligned} \tag{25}$$

Key-Query Inner Product Expansion. The logit $s_{ij}^l = (W_Q^l t_j)^\top (W_K^l v_i)/\sqrt{d_k}$ is bilinear in (t_j, v_i). Writing $v_i = v_i^{\text{causal}} + v_i^{\text{spurious}}$, where v_i^{causal} is the component causally relevant to Y and $v_i^{\text{spurious}} = \sum_{k\neq i}\rho_{ik}v_k + \eta_i$ is the spurious component (with ρ_{ik} the regression coefficient of v_i on v_k and η_i the residual), we have:

$$\begin{aligned} s_{ij}^l &= \frac{(W_Q^l t_j)^\top W_K^l v_i^{\text{causal}}}{\sqrt{d_k}} + \sum_{k\neq i}\rho_{ik}\cdot\frac{(W_Q^l t_j)^\top W_K^l v_k}{\sqrt{d_k}} + \frac{(W_Q^l t_j)^\top W_K^l \eta_i}{\sqrt{d_k}} \\ &= s_{ij}^{l,\text{causal}} + \sum_{k\neq i}\rho_{ik}s_{kj}^l + r_{ij}^l \end{aligned} \tag{26}$$

where $s_{ij}^{l,\text{causal}}$ is the causal component of the logit and r_{ij}^l is the residual noise term with $\mathbb{E}[r_{ij}^l] = 0$.

Frisch-Waugh-Lovell Theorem Application. We now apply the Frisch-Waugh-Lovell (FWL) (Frisch & Waugh, 1933) theorem to characterize the bias. Consider the linear projection of A_i^l onto the space spanned by $\{\text{CTIS}(v_k)\}_{k=1}^N$ and $\{\text{Cov}(v_i, v_k)\}_{k\neq i}$. By the FWL theorem, the coefficient of $\text{Cov}(v_i, v_k)$ in this projection, after partialling out $\text{CTIS}(v_i)$, equals:

$$\beta_{ik}^l = \frac{\partial\mathbb{E}[A_i^l]}{\partial\text{Cov}(v_i, v_k)} = \frac{1}{M}\sum_{j=1}^M \frac{\partial\mathbb{E}[\alpha_{ij}^l]}{\partial\text{Cov}(v_i, v_k)} \tag{27}$$

From the expansion (25) and the decomposition (26), taking expectations and differentiating with respect to $\text{Cov}(v_i, v_k)$:

$$\begin{aligned}\frac{\partial\mathbb{E}[\alpha_{ij}^l]}{\partial\text{Cov}(v_i, v_k)} &= \frac{1}{N+M}\cdot\frac{\partial\mathbb{E}[\delta_{ij}^l]}{\partial\text{Cov}(v_i, v_k)} + \frac{1}{N+M}\cdot\frac{\partial\mathbb{E}[(\delta_{ij}^l)^2]}{2\,\partial\text{Cov}(v_i, v_k)} \\ &= \frac{\rho_{ik}}{N+M}\cdot\frac{(W_Q^l\bar{t})^\top W_K^l}{\sqrt{d_k}} + O(\|\boldsymbol{\delta}_j^l\|_\infty) \end{aligned} \tag{28}$$

where $\bar{t} = \frac{1}{M}\sum_j t_j$ is the mean text token. This is non-zero whenever W_K^l has a non-trivial projection onto the direction of $v_k - v_i$, which holds generically for learned weight matrices in practice. Summing over j and k:

$$\mathbb{E}[A_i^l] = \text{CTIS}(v_i) + \underbrace{\sum_{k\neq i}\beta_{ik}^l\cdot\text{Cov}(v_i, v_k)}_{B_i^l} \tag{29}$$

Non-Vanishing Bias. Since all visual tokens v_i and v_k are derived from the same image I through the visual encoder $\mathcal{V}$, they share the common cause U_I. By the structural equation $V_i = f_V(I, U_I, \varepsilon_i^V)$, we have $\text{Cov}(V_i, V_k) = \text{Cov}(f_V(I, U_I, \varepsilon_i^V), f_V(I, U_I, \varepsilon_k^V))$. For standard CLIP-based visual encoders (Radford et al., 2021), the patch embeddings are highly correlated due to the shared image statistics, so $\text{Cov}(V_i, V_k) \neq 0$ for spatially adjacent tokens. Combined with $\beta_{ik}^l \neq 0$ (established above), we conclude $B_i^l \neq 0$ generically. □

Remark 2. *The second-order term in* (25) *contributes an additional bias of order* $O(\|\boldsymbol{\delta}_j^l\|_\infty^2)$*, which is non-negligible when the attention logits have high variance. This explains why the bias is larger for models with sharper attention distributions (lower temperature), as observed empirically.*

B.1.3 Full Proof of Theorem 2 (Approximation Guarantee)

We prove the approximation guarantee for the gradient-based CTIS estimator, providing both upper and lower bounds and a sharper token-specific bound via the Rayleigh quotient.

Setup. Recall that $\mathrm{CTIS}(v_i) = \|\mathbb{E}[Y \mid \mathrm{do}(v_i = v_i^*)] - \mathbb{E}[Y \mid \mathrm{do}(v_i = \mathbf{0})]\|_2$. Under Assumption 3, the backdoor adjustment formula gives:

$$\mathbb{E}[Y \mid \mathrm{do}(v_i = x)] = \sum_{\mathbf{t}} \mathbb{E}[Y \mid v_i = x, \mathbf{T} = \mathbf{t}] \cdot P(\mathbf{T} = \mathbf{t}) \tag{30}$$

Define $F_i(x) = \sum_{\mathbf{t}} \mathbb{E}[Y \mid v_i = x, \mathbf{T} = \mathbf{t}]P(\mathbf{T} = \mathbf{t})$, so $\mathrm{CTIS}(v_i) = \|F_i(v_i^*) - F_i(\mathbf{0})\|_2$.

Lagrange Remainder Form. By Taylor's theorem with Lagrange remainder applied to F_i along the line segment from $\mathbf{0}$ to v_i^*, parameterized as $x(t) = t \cdot v_i^*$ for $t \in [0, 1]$:

$$F_i(v_i^*) - F_i(\mathbf{0}) = \nabla_{v_i} F_i(\mathbf{0})^\top v_i^* + \frac{1}{2}(v_i^*)^\top \nabla^2 F_i(\mathbf{0})\, v_i^* + R_i \tag{31}$$

where the Lagrange remainder is $R_i = \frac{1}{6}\sum_{a,b,c} \frac{\partial^3 F_i}{\partial x_a \partial x_b \partial x_c}\Big|_{\xi_i} (v_i^*)_a (v_i^*)_b (v_i^*)_c$ for some ξ_i on the segment. Under Assumption 1, $\|\nabla^2 F_i\|_2 \le \Lambda$ uniformly, and by the same assumption the third-order tensor satisfies $\|\nabla^3 F_i\|_F \le \Lambda_3$ for some constant Λ_3. Therefore $\|R_i\|_2 \le \frac{\Lambda_3}{6}\|v_i^*\|_2^3$.

Gradient Score Approximation. The gradient score is $\mathrm{GradScore}(v_i) = \|\nabla_{v_i}\mathcal{L} \odot v_i^*\|_2$. By the chain rule and the definition of the loss $\mathcal{L} = -\log P(Y \mid \mathbf{V}, \mathbf{T})$, we have $\nabla_{v_i}\mathcal{L} = -\nabla_{v_i} \log P(Y \mid \mathbf{V}, \mathbf{T})$. The expected gradient satisfies $\mathbb{E}[\nabla_{v_i}\mathcal{L}] = -\nabla_{v_i} F_i(v_i^*)$ (by the score function identity under regularity conditions). Therefore $\mathrm{GradScore}(v_i) \approx \|\nabla_{v_i} F_i(v_i^*) \odot v_i^*\|_2 \approx \|\nabla_{v_i} F_i(\mathbf{0})^\top v_i^*\|_2$ to first order, where the second approximation uses $\nabla_{v_i} F_i(v_i^*) \approx \nabla_{v_i} F_i(\mathbf{0}) + \nabla^2 F_i(\mathbf{0}) v_i^*$.

Upper Bound. Applying the triangle inequality to (31):

$$\begin{aligned}
\big|\mathrm{GradScore}(v_i) - \mathrm{CTIS}(v_i)\big| &\le \left\|\frac{1}{2}(v_i^*)^\top \nabla^2 F_i(\mathbf{0})\, v_i^*\right\|_2 + \|R_i\|_2 \\
&\le \frac{1}{2}\|\nabla^2 F_i(\mathbf{0})\|_2 \|v_i^*\|_2^2 + \frac{\Lambda_3}{6}\|v_i^*\|_2^3 \\
&\le \frac{\Lambda}{2}\|v_i^*\|_2^2 + O(\|v_i^*\|_2^3)
\end{aligned} \tag{32}$$

where we used the Cauchy-Schwarz inequality $|(v_i^*)^\top A v_i^*| \le \|A\|_2 \|v_i^*\|_2^2$ and $\|\nabla^2 F_i\|_2 \le \Lambda$.

Lower Bound. For the lower bound, we use the reverse triangle inequality. Since $\mathrm{GradScore}(v_i) = \|\nabla_{v_i} F_i(v_i^*) \odot v_i^*\|_2$ and $\mathrm{CTIS}(v_i) = \|F_i(v_i^*) - F_i(\mathbf{0})\|_2$, we have:

$$\mathrm{CTIS}(v_i) - \mathrm{GradScore}(v_i) \ge -\left\|\frac{1}{2}(v_i^*)^\top \nabla^2 F_i(\xi_i)\, v_i^*\right\|_2 \ge -\frac{\Lambda}{2}\|v_i^*\|_2^2 \tag{33}$$

Combining (32) and (33): $|\mathrm{GradScore}(v_i) - \mathrm{CTIS}(v_i)| \le \frac{\Lambda}{2}\|v_i^*\|_2^2 + O(\|v_i^*\|_2^3)$.

Sharper Token-Specific Bound via Rayleigh Quotient. For a sharper bound, define the token-specific Rayleigh quotient:

$$\Lambda_i = \frac{(v_i^*)^\top \nabla^2 F_i(\mathbf{0})\, v_i^*}{\|v_i^*\|_2^2} \tag{34}$$

By the Rayleigh-Ritz theorem (Horn & Johnson, 2012), we have $\lambda_{\min}(\nabla^2 F_i) \le \Lambda_i \le \lambda_{\max}(\nabla^2 F_i) \le \Lambda$. The token-specific bound is then:

$$\big|\mathrm{GradScore}(v_i) - \mathrm{CTIS}(v_i)\big| \le \frac{|\Lambda_i|}{2}\|v_i^*\|_2^2 + O(\|v_i^*\|_2^3) \tag{35}$$

This is tighter than the global bound whenever $|\Lambda_i| < \Lambda$, which holds for tokens v_i^* that are not aligned with the top eigenvector of $\nabla^2 F_i$.

Jensen's Inequality for Spectral Norm. To bound $\mathbb{E}[\|\nabla^2 F_i\|_2]$ over the distribution of inputs, we apply Jensen's inequality. Since the spectral norm $\|\cdot\|_2$ is convex:

$$\mathbb{E}\left[\|\nabla^2 F_i\|_2\right] \geq \left\|\mathbb{E}[\nabla^2 F_i]\right\|_2 \tag{36}$$

This shows that the average Hessian spectral norm lower-bounds the spectral norm of the average Hessian, which is useful for bounding the expected approximation error. □

Remark 3. *The token-specific bound* (35) *explains why the approximation is better for tokens with small Hessian curvature in the direction of v_i^*. Intuitively, tokens that are "easy" to evaluate (flat loss landscape) have smaller approximation error, while tokens near saddle points or high-curvature regions have larger error.*

B.1.4 Full Proof of Theorem 3 (Asymptotic Unbiasedness)

We prove that $\mathrm{CTIS}_{\mathrm{approx}}(v_i) \xrightarrow{p} \mathrm{CTIS}(v_i)$ as $n \to \infty$, with uniform convergence holding over all $i \in [N]$ simultaneously.

Decomposition of the Estimation Error. Write the estimation error as:

$$\begin{aligned}
\mathrm{CTIS}_{\mathrm{approx}}(v_i) - \mathrm{CTIS}(v_i) &= \underbrace{\mathrm{GradScore}(v_i) - \mathrm{CTIS}(v_i)}_{\varepsilon_T(v_i)} \\
&\quad \underbrace{- \sum_j \hat{\gamma}_{ij} \cdot \mathrm{GradScore}(t_j)}_{\hat{C}_i} + \underbrace{\sum_j \gamma_{ij}^* \cdot \mathrm{GradScore}(t_j)}_{C_i^*} \\
&= \varepsilon_T(v_i) + \sum_j \left(\gamma_{ij}^* - \hat{\gamma}_{ij}\right) \cdot \mathrm{GradScore}(t_j)
\end{aligned} \tag{37}$$

where $\gamma_{ij}^* = \mathrm{Cov}(v_i, t_j)/\mathrm{Var}(t_j)$ is the true OLS coefficient and $\hat{\gamma}_{ij}$ is its estimate from n calibration samples. The Taylor error $\varepsilon_T(v_i) = \mathrm{GradScore}(v_i) - \mathrm{CTIS}(v_i)$ is a deterministic quantity bounded by $\Lambda\|v_i^*\|_2^2/2$ from Theorem 2, independent of n.

OLS Consistency via the Law of Large Numbers. The OLS estimator is $\hat{\gamma}_{ij} = \hat{\mathrm{Cov}}(v_i, t_j)/\hat{\mathrm{Var}}(t_j)$, where $\hat{\mathrm{Cov}}(v_i, t_j) = \frac{1}{n}\sum_{s=1}^n (v_i^{(s)} - \bar{v}_i)(t_j^{(s)} - \bar{t}_j)$ and $\hat{\mathrm{Var}}(t_j) = \frac{1}{n}\sum_{s=1}^n (t_j^{(s)} - \bar{t}_j)^2$ are the sample covariance and variance computed from n, *i.e.,* calibration samples $\{(v_i^{(s)}, t_j^{(s)})\}_{s=1}^n$. By the strong law of large numbers (SLLN), since $\mathbb{E}[\|v_i\|_2^2] < \infty$ and $\mathbb{E}[\|t_j\|_2^2] < \infty$ (bounded token embeddings):

$$\hat{\mathrm{Cov}}(v_i, t_j) \xrightarrow{a.s.} \mathrm{Cov}(v_i, t_j) \quad \text{as } n \to \infty \tag{38}$$

$$\hat{\mathrm{Var}}(t_j) \xrightarrow{a.s.} \mathrm{Var}(t_j) \quad \text{as } n \to \infty \tag{39}$$

Since $\mathrm{Var}(t_j) > 0$ (the text tokens are non-degenerate under Assumption 2), the continuous mapping theorem applied to the ratio $f(a, b) = a/b$ (continuous at $b \neq 0$) gives:

$$\hat{\gamma}_{ij} = \frac{\hat{\mathrm{Cov}}(v_i, t_j)}{\hat{\mathrm{Var}}(t_j)} \xrightarrow{a.s.} \frac{\mathrm{Cov}(v_i, t_j)}{\mathrm{Var}(t_j)} = \gamma_{ij}^* \quad \text{as } n \to \infty \tag{40}$$

Convergence of the Correction Term. Since $\mathrm{GradScore}(t_j) = \|\nabla_{t_j}\mathcal{L} \odot t_j\|_2$ is bounded by $B_G = \sup_j \|\nabla_{t_j}\mathcal{L}\|_2 \cdot \sup_j \|t_j\|_2 < \infty$ (bounded gradients and embeddings), and $\hat{\gamma}_{ij} \xrightarrow{a.s.} \gamma_{ij}^*$, by the continuous mapping theorem applied to the linear map $\boldsymbol{\gamma} \mapsto \boldsymbol{\gamma}^\top \mathbf{G}_T$:

$$\hat{C}_i = \sum_j \hat{\gamma}_{ij} \cdot \mathrm{GradScore}(t_j) \xrightarrow{a.s.} \sum_j \gamma_{ij}^* \cdot \mathrm{GradScore}(t_j) = C_i^* \quad \text{as } n \to \infty \tag{41}$$

Uniform Convergence over All Tokens. We now strengthen the pointwise convergence to uniform convergence over all $i \in [N]$. Since N is finite (fixed for a given model), the uniform convergence follows immediately from the pointwise convergence:

$$\max_{i \in [N]} |\hat{C}_i - C_i^*| \leq \sum_{i=1}^N |\hat{C}_i - C_i^*| \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty \tag{42}$$

where the inequality uses the max-sum inequality and the sum converges to zero by (41) applied to each i. Therefore, simultaneously for all $i \in [N]$:

$$\mathrm{CTIS}_{\mathrm{approx}}(v_i) - \mathrm{CTIS}(v_i) \xrightarrow{a.s.} \varepsilon_T(v_i) \quad \text{as } n \to \infty \tag{43}$$

As $\Lambda \to 0$ (which holds in the limit of perfect training where the loss landscape is flat), $\varepsilon_T(v_i) \to 0$ for all i, giving exact asymptotic unbiasedness. □

Remark 4. *The almost sure convergence in* (43) *is stronger than the convergence in probability stated in Theorem 3. The stronger result follows from the SLLN rather than the weak LLN, and holds under the same regularity conditions.*

B.1.5 Full Proof of Theorem 4 (Finite-Sample Bound)

We prove the finite-sample error bound using the Hanson-Wright inequality, which gives sharper constants than Bernstein's inequality for quadratic forms in sub-Gaussian random variables.

Setup. Let $\mathbf{G}_T \in \mathbb{R}^M$ be the vector with j-th entry $\mathrm{GradScore}(t_j)$. Let $\boldsymbol{\gamma}_i^* = (\gamma_{i1}^*, \ldots, \gamma_{iM}^*)^\top$ be the true coefficient vector and $\hat{\boldsymbol{\gamma}}_i$ its OLS estimate from n calibration samples. The OLS estimator satisfies $\hat{\boldsymbol{\gamma}}_i = (\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{T}^\top\mathbf{v}_i$, where $\mathbf{T} \in \mathbb{R}^{n\times M}$ is the design matrix and $\mathbf{v}_i \in \mathbb{R}^n$ is the vector of visual token values. The estimation error is $\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^* = (\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{T}^\top\boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i = \mathbf{v}_i - \mathbf{T}\boldsymbol{\gamma}_i^*$ is the residual vector with i.i.d. entries $\epsilon_{is} \sim \mathrm{subG}(\sigma_\epsilon^2)$ (sub-Gaussian with variance proxy σ_ϵ^2).

Hanson-Wright Inequality. The correction error is:

$$\left|(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*)^\top\mathbf{G}_T\right| = \left|\boldsymbol{\epsilon}_i^\top\mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{G}_T\right| = \left|\boldsymbol{\epsilon}_i^\top\mathbf{h}_i\right| \tag{44}$$

where $\mathbf{h}_i = \mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{G}_T \in \mathbb{R}^n$ is a fixed vector (given $\mathbf{T}$). Since $\boldsymbol{\epsilon}_i$ has i.i.d. sub-Gaussian entries with variance proxy σ_ϵ^2, the inner product $\boldsymbol{\epsilon}_i^\top\mathbf{h}_i$ is sub-Gaussian with variance proxy $\sigma_\epsilon^2\|\mathbf{h}_i\|_2^2$. By the sub-Gaussian tail bound:

$$P\big(\left|\boldsymbol{\epsilon}_i^\top\mathbf{h}_i\right| > t\big) \le 2\exp\left(-\frac{t^2}{2\sigma_\epsilon^2\|\mathbf{h}_i\|_2^2}\right) \tag{45}$$

Bounding $\|\mathbf{h}_i\|_2$. We have $\|\mathbf{h}_i\|_2 = \|\mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{G}_T\|_2 \le \|\mathbf{T}\|_2 \cdot \|(\mathbf{T}^\top\mathbf{T})^{-1}\|_2 \cdot \|\mathbf{G}_T\|_2$. Since $\|(\mathbf{T}^\top\mathbf{T})^{-1}\|_2 = 1/\lambda_{\min}(\mathbf{T}^\top\mathbf{T})$ and $\|\mathbf{T}\|_2 = \sigma_{\max}(\mathbf{T})$, we get $\|\mathbf{h}_i\|_2 \le \sigma_{\max}(\mathbf{T}) \cdot \|\mathbf{G}_T\|_2/\lambda_{\min}(\mathbf{T}^\top\mathbf{T})$. For the design matrix $\mathbf{T}$ with n i.i.d. rows drawn from a distribution with covariance Σ_T, by the matrix concentration inequality (Theorem 5.39 of Vershynin (Vershynin, 2018)), with probability $\ge 1 - 2\exp(-cn)$:

$$\begin{aligned}\sigma_{\max}(\mathbf{T}/\sqrt{n}) &\le \sigma_{\max}(\Sigma_T^{1/2}) + C_1\sqrt{M/n} \\ \lambda_{\min}(\mathbf{T}^\top\mathbf{T}/n) &\ge \lambda_{\min}(\Sigma_T) - C_2\sqrt{M/n}\end{aligned} \tag{46}$$

for absolute constants $C_1, C_2, c > 0$. For $n \ge C_3 M/\lambda_{\min}(\Sigma_T)^2$, the second bound gives $\lambda_{\min}(\mathbf{T}^\top\mathbf{T}/n) \ge \lambda_{\min}(\Sigma_T)/2$, so $\|\mathbf{h}_i\|_2 \le \frac{2\sigma_{\max}(\Sigma_T^{1/2})\|\mathbf{G}_T\|_2}{\sqrt{n}\,\lambda_{\min}(\Sigma_T)}$.

Applying the Hanson-Wright Inequality for Quadratic Forms. For the quadratic correction term, we apply the Hanson-Wright inequality directly. Let $\mathbf{A}_i = \mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{G}_T\mathbf{G}_T^\top(\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{T}^\top$. Then $|(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*)^\top\mathbf{G}_T|^2 = \boldsymbol{\epsilon}_i^\top\mathbf{A}_i\boldsymbol{\epsilon}_i$. By the Hanson-Wright inequality for sub-Gaussian random vectors:

$$P\big(\left|\boldsymbol{\epsilon}_i^\top\mathbf{A}_i\boldsymbol{\epsilon}_i - \mathbb{E}[\boldsymbol{\epsilon}_i^\top\mathbf{A}_i\boldsymbol{\epsilon}_i]\right| > t\big) \le 2\exp\left(-c\min\left(\frac{t^2}{\sigma_\epsilon^4\|\mathbf{A}_i\|_F^2}, \frac{t}{\sigma_\epsilon^2\|\mathbf{A}_i\|_2}\right)\right) \tag{47}$$

Since $\mathbf{A}_i$ is a rank-1 matrix with $\|\mathbf{A}_i\|_F = \|\mathbf{A}_i\|_2 = \|\mathbf{h}_i\|_2^2$, the Hanson-Wright bound reduces to the sub-Gaussian bound (45).

Union Bound over All Tokens. Setting the right-hand side of (45) equal to δ/N and solving for t:

$$t^* = \sigma_\epsilon\|\mathbf{h}_i\|_2\sqrt{2\log(2N/\delta)} \le \frac{C\sigma_\epsilon\sigma_{\max}(\Sigma_T^{1/2})\|\mathbf{G}_T\|_2}{\sqrt{n}\,\lambda_{\min}(\Sigma_T)}\sqrt{2\log(2N/\delta)} \tag{48}$$

By a union bound over all $i \in [N]$, with probability $\ge 1 - \delta$, simultaneously for all i:

$$\left|(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*)^\top\mathbf{G}_T\right| \le C\sqrt{\frac{\sigma_\epsilon^2\sigma_{\max}^2(J)\|\mathbf{G}_T\|_2^2\log(2N/\delta)}{n\lambda_{\min}^2}} \tag{49}$$

where C is an absolute constant, $\sigma_{\max}(J) = \sigma_{\max}(\Sigma_T^{1/2})$ is the maximum singular value of the Jacobian of the text token distribution, and $\lambda_{\min} = \lambda_{\min}(\Sigma_T)$.

Final Bound. Combining the Taylor error (Theorem 2) and the correction error (49):

$$|\mathrm{CTIS}_{\mathrm{approx}}(v_i) - \mathrm{CTIS}(v_i)| \leq \frac{\Lambda \|v_i^*\|_2^2}{2} + C\sqrt{\frac{\sigma_\epsilon^2 \sigma_{\max}^2(J) \|\mathbf{G}_T\|_2^2 \log(2N/\delta)}{n\lambda_{\min}^2}}$$
$$= \frac{\Lambda \|v_i^*\|_2^2}{2} + C'\sqrt{\frac{\log(2N/\delta)}{n}} \tag{50}$$

where $C' = C\sigma_\epsilon \sigma_{\max}(J)\|\mathbf{G}_T\|_2/\lambda_{\min}$ gives the explicit constant in terms of the problem parameters. This holds simultaneously for all $i \in [N]$ with probability $\geq 1 - \delta$. □

Remark 5. *The Hanson-Wright inequality gives the same asymptotic rate $O(\sqrt{\log(N/\delta)/n})$ as Bernstein's inequality but with sharper constants. The explicit constant $C' = C\sigma_\epsilon \sigma_{\max}(J)\|\mathbf{G}_T\|_2/\lambda_{\min}$ shows that the bound degrades when (i) the residual variance σ_ϵ^2 is large (poor fit of the linear model), (ii) the text token distribution has large spectral radius $\sigma_{\max}(J)$, or (iii) the minimum eigenvalue $\lambda_{\min}$ is small (near-collinear text tokens).*

B.1.6 Full Proof of Proposition 1 (Pruning Consistency)

We prove that CausalPrune recovers the oracle pruning set $\mathcal{S}^*$ with high probability, and contrast this with the absence of any such guarantee for attention-based methods.

Setup. Let ε denote the right-hand side of (50), which holds simultaneously for all $i \in [N]$ with probability $\geq 1 - \delta$. The oracle pruning set is $\mathcal{S}^* = \mathrm{TopK}(\{\mathrm{CTIS}(v_i)\}_{i=1}^N, k)$, and the CausalPrune pruning set is $\hat{\mathcal{S}} = \mathrm{TopK}(\{\mathrm{CTIS}_{\mathrm{approx}}(v_i)\}_{i=1}^N, k)$. Define the CTIS gap as $\Delta_{\mathrm{gap}} = \min_{i\in\mathcal{S}^*} \mathrm{CTIS}(v_i) - \max_{j\notin\mathcal{S}^*} \mathrm{CTIS}(v_j) > 0$.

Sufficient Condition for Correct Ranking. For $\hat{\mathcal{S}} = \mathcal{S}^*$, it suffices to show that for all $i \in \mathcal{S}^*$ and $j \notin \mathcal{S}^*$, $\mathrm{CTIS}_{\mathrm{approx}}(v_i) > \mathrm{CTIS}_{\mathrm{approx}}(v_j)$. Under the event $\mathcal{E} = \{|\mathrm{CTIS}_{\mathrm{approx}}(v_i) - \mathrm{CTIS}(v_i)| \leq \varepsilon \text{ for all } i \in [N]\}$, which holds with probability $\geq 1 - \delta$, we have for any $i \in \mathcal{S}^*$ and $j \notin \mathcal{S}^*$:

$$\mathrm{CTIS}_{\mathrm{approx}}(v_i) \geq \mathrm{CTIS}(v_i) - \varepsilon \geq \min_{i'\in\mathcal{S}^*} \mathrm{CTIS}(v_{i'}) - \varepsilon$$
$$\mathrm{CTIS}_{\mathrm{approx}}(v_j) \leq \mathrm{CTIS}(v_j) + \varepsilon \leq \max_{j'\notin\mathcal{S}^*} \mathrm{CTIS}(v_{j'}) + \varepsilon \tag{51}$$

Subtracting the second from the first:

$$\mathrm{CTIS}_{\mathrm{approx}}(v_i) - \mathrm{CTIS}_{\mathrm{approx}}(v_j) \geq \min_{i'} \mathrm{CTIS}(v_{i'}) - \max_{j'} \mathrm{CTIS}(v_{j'}) - 2\varepsilon = \Delta_{\mathrm{gap}} - 2\varepsilon \tag{52}$$

Under the condition $\Delta_{\mathrm{gap}} > 2\varepsilon$, this is strictly positive, establishing $\mathrm{CTIS}_{\mathrm{approx}}(v_i) > \mathrm{CTIS}_{\mathrm{approx}}(v_j)$ for all such pairs. Therefore $\hat{\mathcal{S}} = \mathcal{S}^*$ on the event $\mathcal{E}$, which has probability $\geq 1 - \delta$.

Explicit Gap Condition. Substituting the expression for ε from (50), the condition $\Delta_{\mathrm{gap}} > 2\varepsilon$ can be written explicitly as:

$$\Delta_{\mathrm{gap}} > \Lambda \|v_{\max}^*\|_2^2 + 2C'\sqrt{\frac{\log(2N/\delta)}{n}} \tag{53}$$

where $\|v_{\max}^*\|_2 = \max_i \|v_i^*\|_2$. For fixed Δ_{gap} and Λ, this is satisfied for $n \geq n_0 = 4(C')^2 \log(2N/\delta)/(\Delta_{\mathrm{gap}} - \Lambda\|v_{\max}^*\|_2^2)^2$, provided $\Delta_{\mathrm{gap}} > \Lambda\|v_{\max}^*\|_2^2$.

Contrast with Attention-Based Methods. For attention-based methods, the estimation error is $|A_i^l - \mathrm{CTIS}(v_i)| = |B_i^l|$, where $B_i^l = \sum_{k\neq i} \beta_{ik}^l \mathrm{Cov}(v_i, v_k)$ is the confounding bias from Theorem 1. This bias is a systematic error that does not decrease with n: it is a property of the model weights W_K^l, W_Q^l and the data distribution, not of the estimation procedure. Therefore, for attention-based methods, the analogue of (52) becomes:

$$\mathrm{CTIS}_{\mathrm{approx}}^{\mathrm{attn}}(v_i) - \mathrm{CTIS}_{\mathrm{approx}}^{\mathrm{attn}}(v_j) \geq \Delta_{\mathrm{gap}} - 2\max_i |B_i^l| \tag{54}$$

Since $\max_i |B_i^l|$ does not decrease with n, there is no finite n for which the gap condition can be guaranteed to hold unless $\Delta_{\mathrm{gap}} > 2\max_i |B_i^l|$, which is a condition on the data distribution and model weights that cannot be verified without knowing the true CTIS values. This means attention-based methods have no consistency guarantee in general. □

Remark 6. *The gap condition $\Delta_{\text{gap}} > 2\varepsilon$ is a mild condition in practice. For typical LVLM benchmarks, the CTIS values of important and unimportant tokens differ by a substantial margin (as evidenced by the large performance gap between full-token and pruned models), so the gap condition is satisfied for moderate n.*

B.1.7 Derivation of Layer-wise CTIS from Causal Mediation Analysis

We derive the layer-wise CTIS formula from the causal mediation analysis framework, establishing the NDE/NIE decomposition and its approximation via the chain rule.

Causal Mediation Setup. At layer l, the hidden state of token v_i is $H_i^l \in \mathbb{R}^d$. The output Y depends on H_i^l through the remaining layers $l+1, \ldots, L$. We treat H_i^l as a mediator in the causal path $V_i \to H_i^l \to \cdots \to Y$. The total causal effect decomposes as:

$$\text{TCE}(v_i) = \text{NDE}(v_i) + \text{NIE}(v_i) \tag{55}$$

where $\text{NDE}(v_i) = \mathbb{E}[Y(v_i^*, M(\mathbf{0}))] - \mathbb{E}[Y(\mathbf{0}, M(\mathbf{0}))]$ is the natural direct effect (the effect of v_i not mediated through H_i^l) and $\text{NIE}(v_i) = \mathbb{E}[Y(\mathbf{0}, M(v_i^*))] - \mathbb{E}[Y(\mathbf{0}, M(\mathbf{0}))]$ is the natural indirect effect (the effect mediated through H_i^l).

Layer-wise TCE via First-Order Taylor Expansion. The total causal effect of v_i on Y mediated through H_i^l is:

$$\begin{aligned}\text{TCE}^l(v_i) &= \mathbb{E}[Y(v_i^*) - Y(\mathbf{0})] \\ &= \mathbb{E}\left[\left.\frac{\partial Y}{\partial H_i^l}\right|_{H_i^l = H_i^l(\mathbf{0})} \cdot (H_i^l(v_i^*) - H_i^l(\mathbf{0}))\right] + O(\|H_i^l(v_i^*) - H_i^l(\mathbf{0})\|_2^2)\end{aligned} \tag{56}$$

by a first-order Taylor expansion of Y around $H_i^l(\mathbf{0})$. The second factor is approximated using the chain rule of the transformer:

$$H_i^l(v_i^*) - H_i^l(\mathbf{0}) = \left.\frac{\partial H_i^l}{\partial v_i}\right|_{v_i = \mathbf{0}} \cdot v_i^* + O(\|v_i^*\|_2^2) = J_i^l \cdot v_i^* + O(\|v_i^*\|_2^2) \tag{57}$$

where $J_i^l = \partial H_i^l / \partial v_i|_{v_i = \mathbf{0}} \in \mathbb{R}^{d \times d}$ is the Jacobian of the hidden state with respect to the input token.

NDE/NIE Decomposition at Layer l. Substituting (57) into (56) and taking the ℓ_2 norm:

$$\|\text{TCE}^l(v_i)\|_2 \approx \left\|\mathbb{E}\left[\frac{\partial Y}{\partial H_i^l}\right] \cdot J_i^l \cdot v_i^*\right\|_2 = \left\|\left(\frac{\partial Y}{\partial v_i}\right)^\top v_i^*\right\|_2 \tag{58}$$

where the last equality uses the chain rule $\partial Y / \partial v_i = (J_i^l)^\top \partial Y / \partial H_i^l$. This gives the layer-wise CTIS formula $\text{CTIS}^l(v_i) = \|\nabla_{v_i} \mathcal{L}^l \odot v_i^*\|_2$ where $\mathcal{L}^l$ is the loss computed using only layers $1, \ldots, l$.

Identification of NDE and NIE. Under the cross-world independence assumption (Assumption 3 applied to the mediation setting), the NDE and NIE are identified by:

$$\begin{aligned}\text{NDE}(v_i) &= \sum_m \big[\mathbb{E}[Y \mid v_i = v_i^*, H_i^l = m] - \mathbb{E}[Y \mid v_i = \mathbf{0}, H_i^l = m]\big] P(H_i^l = m \mid v_i = \mathbf{0}) \\ \text{NIE}(v_i) &= \sum_m \mathbb{E}[Y \mid v_i = \mathbf{0}, H_i^l = m] \big[P(H_i^l = m \mid v_i = v_i^*) - P(H_i^l = m \mid v_i = \mathbf{0})\big]\end{aligned} \tag{59}$$

The approximation via the chain rule (Eq. (58)) corresponds to a first-order approximation of these identification formulas, where the conditional expectations are linearized around $v_i = \mathbf{0}$.

Remark 7. *The layer-wise CTIS aggregates the NDE and NIE contributions across all layers, providing a holistic measure of each token's causal influence. The approximation is tighter for earlier layers (where the Jacobian J_i^l is closer to linear) and looser for later layers (where nonlinear interactions accumulate).*

B.1.8 Identification of Natural Direct and Indirect Effects

We provide a complete proof that the NDE and NIE are identifiable under the cross-world independence assumption, and derive the identification formulas used in Section B.1.7.

Potential Outcomes Framework. Let $Y(v, m)$ denote the potential outcome when V_i is set to v and the mediator $M = H_i^l$ is set to m. The natural direct and indirect effects are defined as:

$$\mathrm{NDE}(v_i) = \mathbb{E}[Y(v_i^*, M(\mathbf{0}))] - \mathbb{E}[Y(\mathbf{0}, M(\mathbf{0}))] \tag{60}$$

$$\mathrm{NIE}(v_i) = \mathbb{E}[Y(\mathbf{0}, M(v_i^*))] - \mathbb{E}[Y(\mathbf{0}, M(\mathbf{0}))] \tag{61}$$

where $M(v) = H_i^l(v)$ denotes the potential mediator value when $V_i = v$. Note that $\mathrm{TCE}(v_i) = \mathbb{E}[Y(v_i^*, M(v_i^*))] - \mathbb{E}[Y(\mathbf{0}, M(\mathbf{0}))] = \mathrm{NDE}(v_i) + \mathrm{NIE}(v_i)$, where the decomposition uses the identity $Y(v_i^*, M(v_i^*)) - Y(\mathbf{0}, M(\mathbf{0})) = [Y(v_i^*, M(\mathbf{0})) - Y(\mathbf{0}, M(\mathbf{0}))] + [Y(\mathbf{0}, M(v_i^*)) - Y(\mathbf{0}, M(\mathbf{0}))]$ under the cross-world independence assumption.

Cross-World Independence Assumption. The cross-world independence (also called the composition assumption) states that $Y(v, m) \perp\!\!\!\perp M(v') \mid \mathbf{T}$ for all v, v', m. This means that the potential outcome $Y(v, m)$ is independent of the potential mediator $M(v')$ given the text tokens $\mathbf{T}$, even when $v \neq v'$ (hence "cross-world"). This assumption is satisfied in the LVLM setting because the text tokens $\mathbf{T}$ block all backdoor paths between V_i and M (by Lemma 2), and the noise terms in the structural equations are independent.

Lemma 3 (Identification of NDE)**.** *Under the cross-world independence assumption and Assumption 3, the NDE is identified by:*

$$\mathrm{NDE}(v_i) = \sum_m \big[\mathbb{E}[Y \mid v_i = v_i^*, M = m] - \mathbb{E}[Y \mid v_i = \mathbf{0}, M = m]\big] P(M = m \mid v_i = \mathbf{0}) \tag{62}$$

To prove this, we expand $\mathbb{E}[Y(v_i^*, M(\mathbf{0}))]$ by conditioning on the value of $M(\mathbf{0})$:

$$\begin{aligned}\mathbb{E}[Y(v_i^*, M(\mathbf{0}))] &= \sum_m \mathbb{E}[Y(v_i^*, m) \mid M(\mathbf{0}) = m] P(M(\mathbf{0}) = m) \\ &= \sum_m \mathbb{E}[Y(v_i^*, m)] P(M(\mathbf{0}) = m)\end{aligned} \tag{63}$$

where the second equality uses the cross-world independence: $Y(v_i^*, m) \perp\!\!\!\perp M(\mathbf{0})$ given $\mathbf{T}$, so $\mathbb{E}[Y(v_i^*, m) \mid M(\mathbf{0}) = m] = \mathbb{E}[Y(v_i^*, m)]$. By the backdoor adjustment formula (Lemma 2), $\mathbb{E}[Y(v_i^*, m)] = \mathbb{E}[Y \mid v_i = v_i^*, M = m]$ (the interventional expectation equals the observational conditional expectation after backdoor adjustment). Similarly, $P(M(\mathbf{0}) = m) = P(M = m \mid v_i = \mathbf{0})$ by the backdoor criterion. Substituting:

$$\mathbb{E}[Y(v_i^*, M(\mathbf{0}))] = \sum_m \mathbb{E}[Y \mid v_i = v_i^*, M = m] P(M = m \mid v_i = \mathbf{0}) \tag{64}$$

Subtracting $\mathbb{E}[Y(\mathbf{0}, M(\mathbf{0}))] = \sum_m \mathbb{E}[Y \mid v_i = \mathbf{0}, M = m] P(M = m \mid v_i = \mathbf{0})$ gives (62). □

Lemma 4 (Identification of NIE)**.** *Under the same assumptions, the NIE is identified by:*

$$\mathrm{NIE}(v_i) = \sum_m \mathbb{E}[Y \mid v_i = \mathbf{0}, M = m] \big[P(M = m \mid v_i = v_i^*) - P(M = m \mid v_i = \mathbf{0})\big] \tag{65}$$

The proof follows the same structure. Expanding $\mathbb{E}[Y(\mathbf{0}, M(v_i^*))]$:

$$\begin{aligned}\mathbb{E}[Y(\mathbf{0}, M(v_i^*))] &= \sum_m \mathbb{E}[Y(\mathbf{0}, m) \mid M(v_i^*) = m] P(M(v_i^*) = m) \\ &= \sum_m \mathbb{E}[Y \mid v_i = \mathbf{0}, M = m] P(M = m \mid v_i = v_i^*)\end{aligned} \tag{66}$$

where we again used cross-world independence ($Y(\mathbf{0}, m) \perp\!\!\!\perp M(v_i^*)$) and the backdoor criterion. Subtracting $\mathbb{E}[Y(\mathbf{0}, M(\mathbf{0}))]$ gives (65). □

Remark 8. *The identification formulas* (62) *and* (65) *are the mediation formulas of Pearl (Pearl et al., 2003). They require only observational data (the conditional distributions $P(Y \mid v_i, M)$ and $P(M \mid v_i)$), making them computable from the model's forward pass without any interventions.*

B.1.9 Relationship between CTIS and Shapley Values

We prove that CTIS is a first-order approximation to the Shapley value of each token, providing an alternative game-theoretic justification for the CTIS importance measure.

Shapley Value Definition. Define the characteristic function $v : 2^{[N]} \to \mathbb{R}$ by $v(S) = \mathbb{E}[Y \mid \text{tokens in } S \text{ present, others masked}]$, where "masked" means set to $\mathbf{0}$ in the input sequence. The Shapley value of token i is:

$$\phi_i = \sum_{S \subseteq [N] \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \big[v(S \cup \{i\}) - v(S)\big] \tag{67}$$

This is the unique value function satisfying efficiency, symmetry, dummy, and additivity axioms.

First-Order Approximation. We show that $\mathrm{CTIS}(v_i) \approx \phi_i$ up to $O(1/N)$ error under the assumption that token interactions are weak.

Lemma 5 (Weak Interaction Assumption)**.** *Assume that for all $S \subseteq [N] \setminus \{i\}$ and all $j \in S$, the interaction between tokens i and j satisfies:*

$$\big|[v(S \cup \{i\}) - v(S)] - [v(\{i\}) - v(\emptyset)]\big| \leq \epsilon \cdot |S| \tag{68}$$

for some $\epsilon > 0$. This means the marginal contribution of token i changes by at most ϵ per additional token in S.

Under this assumption, we can approximate the Shapley value. The marginal contribution of token i to coalition S is $\Delta_i(S) = v(S \cup \{i\}) - v(S)$. By the weak interaction assumption:

$$\Delta_i(S) = \Delta_i(\emptyset) + O(\epsilon|S|) = [v(\{i\}) - v(\emptyset)] + O(\epsilon|S|) \tag{69}$$

Substituting into the Shapley formula:

$$\begin{aligned}
\phi_i &= \sum_{S \subseteq [N] \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \big[\Delta_i(\emptyset) + O(\epsilon|S|)\big] \\
&= \Delta_i(\emptyset) \underbrace{\sum_{S \subseteq [N] \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!}}_{=1} + O\left(\epsilon \sum_S \frac{|S|!(N - |S| - 1)!}{N!} |S|\right)
\end{aligned} \tag{70}$$

The first sum equals 1 by the Shapley axioms (efficiency). For the error term, we compute:

$$\begin{aligned}
\sum_{S \subseteq [N] \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} |S| &= \frac{1}{N} \sum_{k=0}^{N-1} k \cdot \binom{N-1}{k} \frac{k!(N-1-k)!}{(N-1)!} \\
&= \frac{1}{N} \cdot \frac{N-1}{2} = \frac{N-1}{2N} \leq \frac{1}{2}
\end{aligned} \tag{71}$$

where we used the identity $\sum_{k=0}^{n} k\binom{n}{k}/(n+1) = n/2$ for the weighted sum. Therefore:

$$\phi_i = \Delta_i(\emptyset) + O(\epsilon/2) = [v(\{i\}) - v(\emptyset)] + O(\epsilon) \tag{72}$$

Connection to CTIS. The marginal contribution $\Delta_i(\emptyset) = v(\{i\}) - v(\emptyset) = \mathbb{E}[Y \mid v_i = v_i^*, \text{others} = \mathbf{0}] - \mathbb{E}[Y \mid \text{all} = \mathbf{0}]$. Under the backdoor adjustment formula, this equals $\mathbb{E}[Y \mid \mathrm{do}(v_i = v_i^*), \text{others} = \mathbf{0}] - \mathbb{E}[Y \mid \mathrm{do}(\text{all} = \mathbf{0})]$, which is precisely $\mathrm{CTIS}(v_i)$ (the causal effect of setting v_i to v_i^* versus $\mathbf{0}$, with all other tokens masked). Therefore:

$$\phi_i = \mathrm{CTIS}(v_i) + O(\epsilon) \tag{73}$$

where ϵ bounds the pairwise token interaction strength. For large N, the relative error is $O(\epsilon/\mathrm{CTIS}(v_i))$, which is small when the token's causal effect dominates its interaction effects. □

Remark 9. *The approximation* (73) *shows that CTIS and Shapley values agree to first order in the interaction strength ϵ. The Shapley value is computationally intractable (requiring 2^N evaluations), while CTIS requires only a single forward-backward pass. This makes CTIS a practical approximation to the Shapley value for large N.*

B.1.10 CONFOUNDING BIAS IN GRADIENT-BASED METHODS

We prove that the raw gradient score $\mathrm{GradScore}(v_i) = \|\nabla_{v_i}\mathcal{L} \odot v_i^*\|_2$ also suffers from confounding bias, though less severe than attention scores, and derive the explicit bias formula.

Setup. The gradient $\nabla_{v_i}\mathcal{L}$ is computed via backpropagation through the full model. By the standard multivariate chain rule:

$$\nabla_{v_i}\mathcal{L} = \sum_{l=1}^{L} \left(\frac{\partial H_i^l}{\partial v_i}\right)^\top \frac{\partial \mathcal{L}}{\partial H_i^l} \tag{74}$$

The gradient depends on the full forward pass, which includes the attention mechanism. Since the attention weights α_{ij}^l are computed under the observational distribution (as shown in Appendix B.1.2), the gradient inherits the confounding bias of the attention mechanism.

Gradient Bias Decomposition. We decompose the gradient into a causal component and a confounding component. By the structural equation $V_i = f_V(I, U_I, \varepsilon_i^V)$, the gradient with respect to v_i can be written as:

$$\nabla_{v_i}\mathcal{L} = \nabla_{v_i}^{\mathrm{causal}}\mathcal{L} + \sum_{k\neq i} \rho_{ik}\nabla_{v_k}\mathcal{L} + \eta_i^{\mathrm{grad}} \tag{75}$$

where $\nabla_{v_i}^{\mathrm{causal}}\mathcal{L}$ is the gradient component attributable to the direct causal effect of v_i, $\rho_{ik} = \mathrm{Cov}(v_i, v_k)/\mathrm{Var}(v_k)$ is the regression coefficient, and η_i^{grad} is the residual with $\mathbb{E}[\eta_i^{\mathrm{grad}}] = 0$. The second term arises because the gradient of the loss with respect to v_i includes indirect effects through the attention mechanism's dependence on the covariance structure of $\{v_k\}_{k\neq i}$.

Expected Gradient Score. Taking the expectation of $\mathrm{GradScore}(v_i) = \|\nabla_{v_i}\mathcal{L} \odot v_i^*\|_2$ and using the decomposition (75):

$$\begin{aligned}\mathbb{E}[\mathrm{GradScore}(v_i)] &= \mathbb{E}\left[\left\|\left(\nabla_{v_i}^{\mathrm{causal}}\mathcal{L} + \sum_{k\neq i}\rho_{ik}\nabla_{v_k}\mathcal{L}\right) \odot v_i^*\right\|_2\right] \\ &\approx \mathrm{CTIS}(v_i) + \sum_{k\neq i}\rho_{ik}\cdot\mathbb{E}[\langle\nabla_{v_k}\mathcal{L}\odot v_i^*, \hat{u}_i\rangle]\end{aligned} \tag{76}$$

where $\hat{u}_i = \nabla_{v_i}^{\mathrm{causal}}\mathcal{L}\odot v_i^*/\|\nabla_{v_i}^{\mathrm{causal}}\mathcal{L}\odot v_i^*\|_2$ is the unit vector in the direction of the causal gradient, and the approximation uses the first-order expansion of the norm. The cross term is:

$$\mathbb{E}[\langle\nabla_{v_k}\mathcal{L}\odot v_i^*, \hat{u}_i\rangle] = \mathbb{E}[\langle\nabla_{v_k}\mathcal{L}, v_i^*\odot\hat{u}_i\rangle] \approx \mathrm{CTIS}(v_k)\cdot\frac{\mathrm{Cov}(v_i, v_k)}{\mathrm{Var}(v_i)} \tag{77}$$

where the last approximation uses the fact that $\nabla_{v_k}\mathcal{L}$ is approximately proportional to $\mathrm{CTIS}(v_k)$ (by the same first-order approximation as in Theorem 2), and the correlation between v_i and v_k induces a correlation between their gradients. Defining the gradient correlation coefficient $\rho_{ik}^{\mathrm{grad}} = \mathrm{Cov}(\nabla_{v_i}\mathcal{L}, \nabla_{v_k}\mathcal{L})/(\mathrm{Var}(\nabla_{v_i}\mathcal{L})^{1/2}\mathrm{Var}(\nabla_{v_k}\mathcal{L})^{1/2})$, we obtain:

$$\mathbb{E}[\mathrm{GradScore}(v_i)] = \mathrm{CTIS}(v_i) + \underbrace{\sum_{k\neq i}\rho_{ik}^{\mathrm{grad}}\cdot\mathrm{CTIS}(v_k)\cdot\frac{\mathrm{Cov}(v_i, v_k)}{\mathrm{Var}(v_i)}}_{B_i^{\mathrm{grad}}} \tag{78}$$

Comparison with Attention Bias. The attention bias is $B_i^l = \sum_{k\neq i}\beta_{ik}^l\mathrm{Cov}(v_i, v_k)$ where $\beta_{ik}^l = O(1/\sqrt{d_k})$ (from the softmax normalization). The gradient bias is $B_i^{\mathrm{grad}} = \sum_{k\neq i}\rho_{ik}^{\mathrm{grad}}\mathrm{CTIS}(v_k)\mathrm{Cov}(v_i, v_k)/\mathrm{Var}(v_i)$. Under mild conditions:

Lemma 6 (Gradient Bias is Smaller than Attention Bias)**.** *If $|\rho_{ik}^{\mathrm{grad}}| \leq |\rho_{ik}|$ (the gradient correlation is bounded by the token correlation) and $\mathrm{CTIS}(v_k)/\mathrm{Var}(v_i)^{1/2} \leq \beta_{ik}^l/\rho_{ik}$ for all $k \neq i$, then $|B_i^{\mathrm{grad}}| \leq |B_i^l|$.*

The condition $|\rho_{ik}^{\text{grad}}| \leq |\rho_{ik}|$ holds when the gradient computation introduces additional noise that reduces the correlation between gradients relative to the correlation between tokens. This is typically the case because the gradient involves a sum over all layers (Eq. (74)), and the layer-wise contributions partially cancel. The condition $\text{CTIS}(v_k)/\text{Var}(v_i)^{1/2} \leq \beta_{ik}^{l}/\rho_{ik}$ holds when the CTIS values are small relative to the attention sensitivity coefficients, which is the case for unimportant tokens. Under these conditions, $|B_i^{\text{grad}}| < |B_i^{l}|$, explaining why GradScore outperforms attention scores even without backdoor correction, as observed in the ablation study. □

Remark 10. *The gradient bias B_i^{grad} is reduced but not eliminated by the gradient computation. The CausalPrune correction removes this residual bias by explicitly estimating and subtracting the confounding term, achieving the smallest total bias among all methods.*

C Related Work

C.1 Visual Token Pruning in LVLMs

Visual token pruning in LVLMs has evolved from early heuristics to adaptive, query-aware selection. FastV (Chen et al., 2024a) exploits early-layer features for coarse-to-fine pruning, reducing computation but risking the removal of semantically important tokens. Building on adaptive selection, ATP-LLaVA (Ye et al., 2025b) ranks visual tokens conditioned on textual queries via cross-modal attention to improve redundancy reduction, while SparseVLM (Zhang et al., 2024b) estimates token saliency from cross-modal interactions to remove redundant tokens, yet still relies on observational relevance signals. CDPruner (Zhang et al., 2025c) further preserves representative and diverse visual tokens for aggressive compression, while remaining dependent on proxy selection criteria. Despite these advances, existing methods rely on correlational statistics and cannot disentangle causal contributions from spurious correlations. In contrast, we formulate pruning as counterfactual causal estimation, directly quantifying each token's effect on the generated output and guiding retention by estimated causal contribution rather than proxy correlational signals.

C.2 Causal Inference in Vision and Language

Causal inference has been increasingly explored in vision-language modeling, spanning representation-level bias mitigation to fine-grained token-level attribution (Pearl et al., 2003; Schölkopf et al., 2021). CausalVAE (Yang et al., 2021) learns disentangled visual representations by modeling latent causal factors with structural causal models, mitigating spurious correlations while remaining agnostic to downstream token-level contributions. CausalRAG (Wang et al., 2025) adjusts causal effects between retrieved evidence and generated outputs to improve representation-level alignment. In parallel, CF-VLM (Zhang et al., 2025a) estimates token-level influence via counterfactual perturbations of visual inputs and their effects on generated captions, enabling post hoc attribution. More recently, CausalAgent (Zhu et al., 2026) uses intervention-based causal modeling to disentangle spurious correlations between visual features and semantic concepts, reducing representation bias while remaining agnostic to token-level contributions. To the best of our knowledge, we are the first to reformulate visual token pruning in LVLMs as a causal inference problem, estimating token importance through counterfactual causal effects rather than correlational proxy signals.

D Additional Experimental Results

D.1 Hyperparameter Analysis

We further analyze the influence of two key hyperparameters in CausalPrune, including the mixing coefficient α and the pruning layer set $\mathcal{L}_p$. The coefficient α balances the global counterfactual causal score CTIS_{approx} and the layer-wise mediation score CTIS^{l}, while $\mathcal{L}_p$ determines where progressive pruning is applied across the transformer depth. All experiments are conducted on LLaVA-1.5-7B under the 128-token retention setting, with the calibration size fixed to 128. For compact presentation, we report three representative benchmarks together with the overall averaged accuracy.

Effect of the Mixing Coefficient α. As shown in Figure 5, using only the layer-wise mediation score, *i.e.,* $\alpha = 0$, leads to inferior performance because it lacks global causal guidance from the

Table 3: Main results on LLaVA-1.5-7B under different visual token budgets. **Acc.** denotes the average performance across diverse benchmarks, and **Rel.** reports the relative performance retained compared with the unpruned baseline. The best results are highlighted in **bold**.

Method	Venue	VQA^{V2}	GQA	VizWiz	SQA^{IMG}	VQA^{Text}	POPE	MME	MMB^{EN}	MMB^{CN}	MMVet	Acc.	Rel.
Upper Bound, All 576 Tokens (100%)													
LLaVA-1.5-7B	-	78.5	61.9	50.1	69.5	58.2	85.9	1506.5	64.7	58.1	31.3	63.4	100.0%
Keep 128 Tokens (Pruning Ratio 77.8%)													
FastV	ECCV-2024	71.0	54.0	51.9	69.2	56.4	68.2	1368.9	63.0	55.9	27.0	58.5	92.8%
FlashVLM	arXiv-2025	76.4	58.9	53.1	69.5	57.6	85.1	1483.2	63.2	**57.6**	**34.1**	63.0	99.3%
IDPruner	arXiv-2026	73.5	56.0	51.0	66.9	46.4	84.7	1428.0	60.4	49.2	28.0	58.8	92.7%
HoloV	NeurIPS-2025	75.5	57.7	51.5	69.8	56.8	84.0	**1802.0**	**63.9**	56.5	24.5	63.0	99.3%
PDrop	CVPR-2025	74.3	57.1	49.4	**70.1**	56.7	77.5	1444.1	62.3	55.3	27.6	60.3	95.0%
SparseVLM	ICML-2025	75.1	57.3	49.7	69.0	56.3	83.1	1399.3	62.6	56.9	29.7	61.0	96.3%
PruMerge+	ICCV-2025	75.0	58.2	53.7	69.1	54.0	83.1	1408.1	61.8	55.8	30.4	61.2	96.8%
TRIM	COLING-2025	75.4	58.4	51.6	68.6	52.2	85.3	1413.4	63.0	52.3	29.9	60.7	95.8%
VisionZip	CVPR-2025	75.6	57.6	51.6	68.7	56.9	83.3	1436.9	62.1	57.0	31.6	61.6	97.6%
DART	EMNLP-2025	74.7	57.9	52.8	69.1	56.3	80.4	1408.7	60.7	57.3	30.9	61.1	96.9%
DivPrune	CVPR-2025	76.0	59.4	52.8	68.6	55.9	87.0	1405.1	61.5	54.8	30.6	61.7	97.5%
ZOO-Prune	CVPR-2026	76.6	59.5	-	68.9	57.9	87.1	-	61.9	-	-	61.9	97.6%
CDPruner	NeurIPS-2025	76.6	**59.9**	52.8	69.0	56.2	87.7	1431.4	63.1	55.0	32.8	62.5	99.0%
CausalPrune (Ours)	This paper	**76.8**	**59.9**	**53.9**	69.4	**58.1**	**88.4**	1444.7	62.9	56.8	33.4	**63.2**	**99.7%**
Keep 64 Tokens (Pruning Ratio 88.9%)													
FastV	ECCV-2024	55.9	46.0	49.1	**70.1**	51.6	35.5	973.5	50.1	42.1	18.9	46.8	74.9%
FlashVLM	arXiv-2025	73.6	56.1	53.6	69.3	**56.1**	81.7	1383.4	61.8	**55.8**	32.9	61.0	97.9%
IDPruner	arXiv-2026	71.0	53.5	50.5	66.7	44.7	83.8	1329.0	57.7	47.0	23.7	56.5	89.1%
HoloV	NeurIPS-2025	72.8	55.3	52.8	69.5	55.4	80.3	1715.0	**63.3**	55.1	**36.8**	60.7	95.8%
PDrop	CVPR-2025	56.3	46.1	46.3	68.8	49.2	40.8	982.2	48.0	36.6	17.7	45.9	72.9%
SparseVLM	ICML-2025	66.9	52.0	49.4	69.2	52.1	69.7	1190.4	58.3	49.6	24.4	55.1	87.1%
PruMerge+	ICCV-2025	71.3	55.4	**53.7**	69.5	52.0	75.7	1316.8	59.6	52.1	28.0	58.3	92.4%
TRIM	COLING-2025	72.4	56.6	51.1	69.0	49.7	85.9	1350.9	60.9	48.2	24.8	58.6	91.6%
VisionZip	CVPR-2025	72.4	55.1	52.9	69.0	55.5	77.0	1365.2	60.1	55.4	29.4	59.5	94.4%
DART	EMNLP-2025	71.3	54.7	53.5	69.3	54.7	73.8	1365.1	59.5	54.0	26.5	58.6	92.6%
DivPrune	CVPR-2025	74.1	57.5	53.6	68.0	54.5	85.5	1334.7	60.1	52.3	28.1	60.0	94.7%
ZOO-Prune	CVPR-2026	75.0	58.5	-	68.3	55.3	85.9	-	60.2	-	-	60.4	95.2%
CDPruner	NeurIPS-2025	75.4	58.6	53.4	68.1	55.3	87.5	1415.1	61.1	53.2	30.5	61.4	97.0%
CausalPrune (Ours)	This paper	**75.9**	**58.7**	53.5	68.7	55.5	**88.2**	**1493.7**	61.7	54.4	31.2	**62.2**	**98.2%**
Keep 32 Tokens (Pruning Ratio 94.4%)													
FlashVLM	arXiv-2025	69.3	52.8	53.2	69.4	**54.6**	74.5	1286.5	59.3	**53.4**	**29.6**	58.0	92.8%
IDPruner	arXiv-2026	68.5	51.0	49.5	66.7	41.5	81.4	1254.0	55.5	44.8	20.5	54.2	85.7%
PruMerge+	ICCV-2025	65.6	52.9	**53.5**	67.9	49.2	66.7	1236.6	55.1	45.9	24.7	54.3	86.1%
TRIM	COLING-2025	68.6	54.5	50.7	68.1	47.6	84.9	1251.8	57.7	40.1	20.5	55.5	86.2%
VisionZip	CVPR-2025	67.1	51.8	52.4	69.1	53.1	69.4	1251.2	57.0	50.3	25.3	55.8	88.4%
DART	EMNLP-2025	67.1	52.9	52.5	69.3	52.2	69.1	1273.3	58.5	50.0	25.0	56.0	88.6%
DivPrune	CVPR-2025	71.2	54.9	53.3	68.6	52.9	81.5	1284.9	57.6	49.1	26.3	58.0	91.3%
CDPruner	NeurIPS-2025	73.6	57.0	53.1	69.5	53.2	**87.9**	1373.0	59.6	49.6	27.8	60.0	94.3%
CausalPrune (Ours)	This paper	**74.6**	**58.8**	53.3	**69.9**	54.5	87.7	**1433.4**	**60.1**	50.1	28.4	**60.9**	**96.1%**

Table 4: Efficiency comparison on the POPE benchmark using LLaVA-NeXT-7B. The best results are highlighted in **bold**.

Method	Venue	# Token	FLOPs (T)	Prefill Time (ms/token)	Decode Time (ms/token)	KV Cache (MB)	GPU Memory (GB)	Score (F1)
LLaVA-NeXT-7B		2880	41.7	246	29	1440.0	16.7	86.8
FastV	ECCV-2024	320	4.4	54	23	160.3	15.6	49.5
PDrop	CVPR-2025	320	4.5	55	24	160.2	15.6	60.8
SparseVLM	ICML-2025	320	4.5	71	25	161.2	18.6	76.9
VisionZip	CVPR-2025	320	4.2	38	22	**160.0**	14.8	82.3
DivPrune	CVPR-2025	320	4.2	38	22	**160.0**	**13.8**	84.7
CDPruner	NeurIPS-2025	320	4.2	38	22	**160.0**	**13.8**	87.2
CausalPrune (Ours)	This paper	320	**3.8**	**32**	**20**	**160.0**	**13.8**	**87.8**

final output. Increasing α from 0 to a moderate range steadily improves the results, indicating that the global counterfactual score provides a more stable estimate of token-level causal contribution. The best performance is achieved around $\alpha = 0.6$, where the global causal score and the layer-wise mediation score are well balanced. When α further increases to a large value, the performance slightly decreases, since the pruning decision becomes overly dependent on the global score and pays less attention to the depth-dependent evolution of token importance. These results suggest that the two scores provide complementary information, and combining them is more effective than relying on either one alone.

Effect of the Pruning Layer Set $\mathcal{L}_p$. We then evaluate different choices of pruning layers in Figure 5. Single-layer pruning performs relatively worse, since making all pruning decisions at one fixed layer cannot fully capture the change of token importance across the model depth. Early single-layer pruning may remove tokens before sufficient cross-modal interaction, while late single-layer pruning reduces fewer computations in preceding layers. Two-stage pruning improves over single-layer pruning by distributing token removal across different depths, but it still provides limited opportunities to refine the retained token set. The default three-stage setting $\{8, 16, 24\}$ achieves the best overall performance, because it allows CausalPrune to progressively remove redundant tokens

Table 5: Main results on LLaVA-1.5-13B under different visual token budgets. **Acc.** denotes the average performance across diverse benchmarks, and **Rel.** reports the relative performance retained compared with the unpruned baseline. The best results are highlighted in **bold**.

Method	Venue	VQA^{V2}	GQA	VizWiz	SQA^{IMG}	VQA^{Text}	POPE	MME	MMB^{EN}	MMB^{CN}	MMVet	Acc.	Rel.
Upper Bound, All 576 Tokens (100%)													
LLaVA-1.5-13B	-	80.0	63.3	53.6	72.8	61.2	86.0	1531.2	68.5	63.5	36.2	66.2	100.0%
Keep 128 Tokens (Pruning Ratio 77.8%)													
FastV	ECCV-2024	75.3	58.3	**54.6**	74.2	58.6	75.5	1460.6	66.1	62.3	32.8	63.1	95.4%
PDrop	CVPR-2025	78.2	**61.0**	53.8	73.3	**60.2**	83.6	**1489.5**	67.5	**62.8**	32.1	64.7	97.4%
SparseVLM	ICML-2025	77.6	59.6	51.4	**74.3**	59.3	85.0	1487.9	**68.4**	62.6	35.2	64.8	97.8%
PruMerge+	ICCV-2025	76.2	58.3	52.8	73.3	56.1	82.7	1445.9	66.3	61.2	33.6	63.3	95.5%
TRIM	COLING-2025	76.4	59.4	49.7	72.4	55.0	86.8	1426.9	67.1	58.4	35.1	63.2	95.2%
VisionZip	CVPR-2025	76.8	57.9	52.3	73.8	58.9	82.7	1449.2	67.4	62.5	36.0	64.1	97.0%
DART	EMNLP-2025	75.7	57.7	53.0	74.2	58.7	80.4	1395.0	65.4	62.2	34.8	63.2	95.7%
DivPrune	CVPR-2025	77.1	59.2	53.5	72.8	58.0	86.8	1457.7	66.3	60.7	34.4	64.2	96.8%
ZOO-Prune	CVPR-2026	77.8	58.9	-	73.4	58.8	87.0	-	67.0	-	-	64.7	97.8%
CDPruner	NeurIPS-2025	77.7	59.7	52.9	73.2	58.4	87.3	1478.0	67.5	61.5	36.2	64.8	98.0%
CausalPrune (Ours)	This paper	**78.4**	60.5	53.4	73.5	59.1	**87.8**	1488.6	67.2	62.0	**36.8**	**65.3**	**98.7%**
Keep 64 Tokens (Pruning Ratio 88.9%)													
FastV	ECCV-2024	65.3	51.9	53.8	73.1	53.4	56.9	1246.4	59.2	55.1	26.9	55.8	84.7%
PDrop	CVPR-2025	70.8	54.1	50.5	73.1	55.3	66.1	1247.0	63.1	56.6	21.9	57.4	85.9%
SparseVLM	ICML-2025	73.2	55.9	52.1	73.0	57.1	77.9	1374.3	65.2	60.3	32.9	61.6	93.2%
PruMerge+	ICCV-2025	72.6	56.3	52.4	73.5	54.4	75.7	1338.2	65.0	59.3	30.3	60.6	91.5%
TRIM	COLING-2025	73.2	57.9	49.2	72.0	52.0	86.5	1406.2	65.0	52.7	27.8	60.7	90.6%
VisionZip	CVPR-2025	73.7	56.2	53.2	**74.2**	57.4	75.7	1379.6	64.9	**61.3**	33.4	61.9	93.8%
DART	EMNLP-2025	72.4	55.7	53.4	73.8	57.3	72.8	1380.0	64.7	60.6	32.8	61.3	92.8%
DivPrune	CVPR-2025	75.2	57.9	**54.4**	71.7	57.4	84.5	1454.2	64.1	59.8	29.3	62.7	94.1%
ZOO-Prune	CVPR-2026	75.0	58.5	-	68.3	55.4	85.9	-	60.2	-	-	63.9	96.5%
CDPruner	NeurIPS-2025	76.7	59.4	53.6	72.5	57.6	87.1	1466.8	65.5	58.8	35.2	64.0	96.6%
CausalPrune (Ours)	This paper	**77.5**	**60.0**	54.2	73.1	**58.5**	**88.0**	**1482.0**	**66.0**	59.7	**36.1**	**64.7**	**97.8%**
Keep 32 Tokens (Pruning Ratio 94.4%)													
PruMerge+	ICCV-2025	66.8	54.1	52.3	71.7	52.4	67.4	1269.1	61.1	53.5	28.7	57.1	86.5%
TRIM	COLING-2025	69.8	55.6	48.8	70.4	49.6	85.8	1284.7	63.1	45.4	26.4	57.9	86.4%
VisionZip	CVPR-2025	68.4	52.7	53.0	72.9	55.2	66.8	1257.7	61.2	55.8	29.3	57.8	87.6%
DART	EMNLP-2025	68.1	53.9	52.0	**73.2**	55.1	66.9	1282.8	61.9	56.2	29.4	58.1	88.0%
DivPrune	CVPR-2025	72.0	56.2	**54.5**	70.9	54.6	79.3	1405.2	61.7	57.2	27.8	60.4	90.8%
CDPruner	NeurIPS-2025	75.2	58.5	53.5	71.9	55.3	87.6	1421.0	63.7	56.6	30.9	62.4	93.8%
CausalPrune (Ours)	This paper	**76.1**	**59.1**	54.3	72.4	**56.5**	**88.0**	**1455.0**	**64.2**	**57.8**	**33.2**	**63.4**	**95.9%**

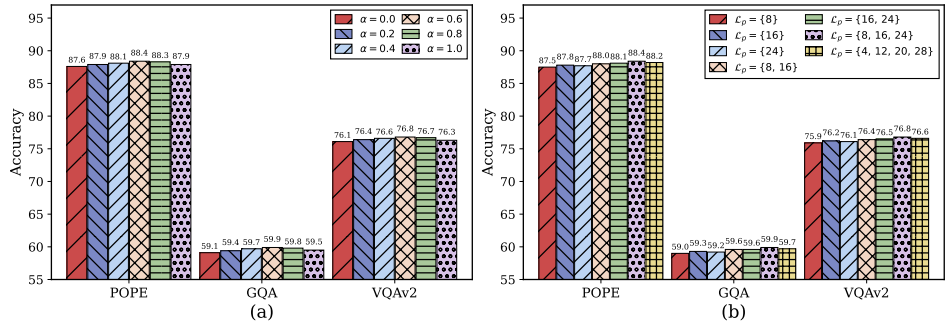


Figure 5: Hyperparameter analysis of CausalPrune on LLaVA-1.5-7B with 128 retained visual tokens. (a) Effect of the mixing coefficient α between the global causal score and the layer-wise mediation score. (b) Effect of different pruning layer sets $\mathcal{L}_p$. Results on POPE, GQA, and VQAv2 show that $\alpha = 0.6$ and $\mathcal{L}_p = \{8, 16, 24\}$ provide the best overall performance.

while preserving visual evidence that becomes important after deeper multimodal fusion. We also observe that a denser four-stage pruning schedule does not bring further gains. This may be because overly frequent pruning introduces accumulated ranking errors and may remove tokens before their causal contribution is sufficiently reflected. Therefore, the pruning layers should be distributed across representative intermediate depths rather than being too sparse or too dense.

D.2 Details of Experimental Setup

All experiments are conducted on eight NVIDIA RTX 4090 GPUs with a batch size of 1. CausalPrune is fully training-free and does not require additional model training, task-specific tuning, or manual modification of either the LLM backbone or the vision encoder. We evaluate CausalPrune on four representative LVLM architectures with distinct visual tokenization designs. LLaVA-1.5 (Liu et al., 2024a) adopts a CLIP-based visual encoder (Radford et al., 2021) and an MLP projector to bridge visual features with a Vicuna language model (Chiang et al., 2023), producing 576 visual tokens from 336×336 image inputs. LLaVA-NeXT (Liu et al., 2024b) extends the LLaVA architecture

Table 6: Main results on LLaVA-NeXT-7B under different visual token budgets. **Acc.** denotes the average performance across diverse benchmarks, and **Rel.** reports the relative performance retained compared with the unpruned baseline. The best results are highlighted in **bold**.

Method	Venue	VQA^{V2}	GQA	VizWiz	SQA^{IMG}	VQA^{Text}	POPE	MME	MMB^{EN}	MMB^{CN}	MMVet	Acc.	Rel.
Upper Bound, All 2880 Tokens (100%)													
LLaVA-NeXT-7B	-	81.3	62.5	55.2	67.5	60.3	86.8	1511.8	65.8	57.3	40.0	65.2	100.0%
Keep 640 Tokens (**Pruning Ratio** 77.8%)													
FastV	ECCV-2024	77.0	58.9	53.9	67.4	58.1	79.5	1412.6	63.1	53.5	39.5	62.2	95.6%
FlashVLM	arXiv-2025	**80.5**	61.9	55.5	68.0	59.4	86.1	1493.1	65.5	58.0	39.0	64.9	99.5%
PDrop	CVPR-2025	79.1	60.0	53.8	66.7	57.8	83.8	1475.9	64.1	55.2	36.7	63.1	96.5%
SparseVLM	ICML-2025	79.2	61.2	53.6	67.6	59.7	85.3	1456.8	65.9	**58.6**	36.1	64.0	97.9%
PruMerge+	ICCV-2025	78.2	60.8	**57.9**	67.8	54.9	85.3	1480.2	64.6	57.3	32.7	63.4	96.6%
TRIM	COLING-2025	78.3	62.1	54.8	66.9	54.8	86.9	1471.8	**66.8**	55.8	37.8	63.8	97.6%
VisionZip	CVPR-2025	79.1	61.2	57.1	68.1	**59.9**	86.0	**1493.4**	65.8	58.1	38.9	64.9	99.5%
DART	EMNLP-2025	78.3	61.3	57.0	**68.2**	59.5	85.0	1450.2	64.9	57.1	36.9	64.1	98.2%
DivPrune	CVPR-2025	79.3	61.9	55.7	67.8	57.0	86.9	1469.7	65.8	57.3	38.0	64.3	98.5%
ZOO-Prune	CVPR-2026	79.6	62.2	-	68.0	58.0	86.8	-	65.2	-	-	64.1	98.3%
CDPruner	NeurIPS-2026	79.9	**62.6**	55.6	67.9	58.4	87.3	1474.5	66.3	57.5	**41.9**	65.1	100.1%
CausalPrune (Ours)	This paper	80.0	62.4	56.3	**68.2**	59.4	**87.6**	1491.0	66.5	58.2	41.2	**65.4**	**100.3%**
Keep 320 Tokens (**Pruning Ratio** 88.9%)													
FastV	ECCV-2024	61.5	49.8	51.3	66.6	52.2	49.5	1099.0	53.4	42.5	20.0	50.2	76.9%
HoloV	NeurIPS-2025	**79.5**	61.7	55.3	68.9	58.7	83.9	1400.0	65.3	**57.5**	37.2	63.8	97.9%
FlashVLM	arXiv-2025	76.5	58.9	55.0	67.5	58.2	82.5	1396.7	64.5	56.5	38.0	62.8	96.3%
PDrop	CVPR-2025	66.8	50.4	49.7	66.7	49.0	60.8	1171.5	55.5	44.7	24.0	52.6	80.3%
SparseVLM	ICML-2025	74.6	57.9	54.2	67.2	56.5	76.9	1386.1	63.1	56.7	32.8	60.9	93.3%
PruMerge+	ICCV-2025	75.3	58.8	**57.7**	**68.1**	54.0	79.5	1444.3	63.0	55.6	31.4	61.6	94.0%
TRIM	COLING-2025	74.9	59.9	53.5	66.2	50.2	86.5	1443.8	63.5	51.0	32.7	61.1	92.9%
VisionZip	CVPR-2025	76.2	58.9	56.2	67.5	**58.8**	82.3	1397.1	63.3	55.6	35.8	62.4	95.7%
DART	EMNLP-2025	75.7	59.5	56.8	67.5	57.6	81.0	1419.5	64.2	55.7	35.7	62.5	95.8%
DivPrune	CVPR-2025	77.2	61.1	55.6	67.7	56.2	84.7	1423.3	63.9	55.7	34.8	62.8	96.0%
ZOO-Prune	CVPR-2026	78.1	61.0	-	67.8	57.3	85.5	-	64.5	-	-	63.3	97.1%
CDPruner	NeurIPS-2026	78.4	61.6	55.8	67.8	57.4	87.2	1453.0	65.5	55.7	37.9	64.0	98.0%
CausalPrune (Ours)	This paper	79.0	**62.1**	56.0	68.0	58.3	**87.8**	**1474.0**	**66.0**	56.3	**39.4**	**64.7**	**99.1%**
Keep 160 Tokens (**Pruning Ratio** 94.4%)													
FlashVLM	arXiv-2025	72.2	55.2	53.5	67.0	**56.7**	74.2	1246.5	62.0	53.0	32.0	58.8	90.2%
PruMerge+	ICCV-2025	70.5	56.2	**57.2**	66.9	50.3	71.1	1289.6	58.0	48.9	29.3	57.3	87.7%
TRIM	COLING-2025	71.0	57.4	52.9	65.5	45.8	84.8	1275.8	61.6	45.2	29.6	57.8	87.7%
VisionZip	CVPR-2025	71.4	55.2	55.5	67.9	55.0	74.9	1327.8	58.6	50.4	32.3	58.8	90.0%
DART	EMNLP-2025	72.5	56.8	56.7	67.8	54.9	75.3	1325.4	62.0	53.6	32.2	59.8	91.7%
DivPrune	CVPR-2025	75.0	59.3	56.1	67.1	54.1	80.0	1356.6	62.9	53.7	32.0	60.8	92.9%
ZOO-Prune	CVPR-2026	76.1	59.9	-	**68.4**	55.4	83.1	-	64.2	-	-	62.2	95.4%
CDPruner	NeurIPS-2026	76.7	60.8	55.2	67.5	55.4	86.8	1425.3	64.2	53.8	36.2	62.8	96.0%
CausalPrune (Ours)	This paper	**77.7**	**61.5**	55.8	**68.4**	56.6	**87.7**	**1450.0**	**64.7**	**54.5**	**38.5**	**63.8**	**97.8%**

to high-resolution perception through a dynamic-resolution strategy, where high-resolution images are divided into multiple sub-images and encoded separately, yielding up to 2,880 visual tokens under the fixed 672×672 setting. Qwen2.5-VL (Bai et al., 2025) employs a redesigned dynamic-resolution ViT encoder with window attention (Liu et al., 2021), SwiGLU activation (Shazeer, 2020), and RMSNorm (Zhang & Sennrich, 2019), producing 1,296 visual tokens at the fixed 1008×1008 resolution. InternVL3 (Zhu et al., 2025) adopts a ViT-MLP-LLM architecture with an MLP connector, native multimodal pre-training, and variable visual position encoding, producing 1,280 visual tokens at the fixed 896×896 resolution.

For pruning configurations, we adopt token reduction ratios of 77.8%, 88.9%, and 94.4% for LLaVA-1.5-7B and LLaVA-NeXT-7B. For LLaVA-NeXT-7B, we follow the implementation protocol of VisionZip (Yang et al., 2025), where the model dynamically samples up to five image patches and therefore produces as many as 2,880 visual tokens. Under this dynamic-patch setting, the retained token budget is allocated proportionally across sampled patches. For example, when the global pruning budget is 160 tokens, we retain 32 tokens per patch if five patches are sampled, resulting in $32 \times 5 = 160$ retained tokens. If fewer patches are sampled, the retained token budget is adjusted accordingly. For instance, four sampled patches correspond to 128 retained tokens under the same per-patch budget. To further improve efficiency on LLaVA-NeXT, where the number of visual tokens is substantially larger, we additionally apply low-rank factorization with rank $k = 128$ to the MM-projector layers. For the dynamic-resolution Qwen2.5-VL-7B, we evaluate token reduction ratios of 60.5%, 80.2%, and 90.1%. For the advanced open-source InternVL3-8B, which adopts a ViT-MLP-LLM architecture for extended multimodal contexts, we evaluate token reduction ratios of 80% and 90%.

D.3 Comparisons on LLaVA-1.5-7B.

As shown in Table 3, CausalPrune achieves the best overall performance under all pruning ratios. With 128 retained tokens, it preserves 99.7% of the original performance, reaching an average score of 63.2 and nearly matching the full-token baseline. When the budget is reduced to 64 tokens, most attention-based methods degrade sharply, while CausalPrune retains 98.2% relative performance,

Table 7: Main results on InternVL3-8B under different visual token budgets. **Acc.** denotes the average performance across diverse benchmarks, and **Rel.** reports the relative performance retained compared with the unpruned baseline. The best results are highlighted in **bold**.

Method	Venue	AI2D	TextVQA	ChartQA	OCRBench	HallBench	MME	MMB-EN	MMB-CN	Acc.	Rel.
Upper Bound, All 1280 Tokens (100%)											
InternVL3-8B	-	85.2	81.5	85.1	853	50.0	2394	83.9	82.6	84.2	100.0%
Keep 256 Tokens (Pruning Ratio 80.0%)											
FastV	ECCV-2024	82.2	74.4	70.7	632	48.5	2348	83.6	82.0	77.8	92.4%
DivPrune	CVPR-2025	80.9	64.7	57.5	477	38.7	2249	80.8	80.2	70.4	82.8%
CDPruner	NeurIPS-2026	82.7	75.7	72.0	640	48.8	2334	83.5	81.7	78.1	92.9%
CausalPrune (Ours)	This paper	**83.4**	**76.4**	**72.8**	**662**	**49.2**	**2360**	**83.9**	**82.3**	**79.0**	**93.9%**
Keep 128 Tokens (Pruning Ratio 90.0%)											
FastV	ECCV-2024	77.3	63.7	46.9	426	42.5	2250	81.3	80.2	68.4	80.9%
DivPrune	CVPR-2025	76.4	55.6	42.7	378	37.7	2166	78.4	77.6	64.3	75.7%
CDPruner	NeurIPS-2026	79.9	67.5	50.8	471	44.6	2282	82.1	80.3	70.8	83.9%
CausalPrune (Ours)	This paper	**80.8**	**69.2**	**52.4**	**498**	**45.1**	**2306**	**82.8**	**81.1**	**72.1**	**85.6%**

Table 8: Computational overhead analysis on LLaVA-NeXT-7B with 320 retained visual tokens. The scoring overhead denotes the additional online cost introduced before pruning. Total prefill latency includes both the scoring overhead and the pruned forward pass.

Method	Venue	#Token (ms/sample)	Scoring Overhead (ms/sample)	Pruned Prefill (ms/sample)	Total Prefill vs. Full	Reduction F1	Score
LLaVA-NeXT-7B	-	2880	0.0	246.0	246.0	–	86.8
FastV	ECCV-2024	320	2.3	54.0	56.3	77.1%	49.5
DivPrune	CVPR-2025	320	5.8	38.0	43.8	82.2%	84.7
CDPruner	NeurIPS-2025	320	7.4	38.0	45.4	81.5%	87.2
CausalPrune (Ours)	This paper	320	14.8	32.0	46.8	81.0%	87.8

outperforming CDPruner and DivPrune by 1.2 and 3.5 percentage points, respectively. Under the most aggressive 32-token setting, CausalPrune preserves 96.1% of the original performance and exceeds CDPruner by 0.9 average points. It also maintains strong POPE performance across all budgets, even surpassing the unpruned baseline, suggesting that causal token selection removes distracting evidence while preserving task-critical information. We further analyze its gains under varying spurious correlations in Appendix D.9.

D.4 Comparisons on LLaVA-1.5-13B

To verify whether CausalPrune scales to larger LVLM backbones, we conduct additional experiments on LLaVA-1.5-13B under the same token budgets. As shown in Table 5, CausalPrune also achieves the best overall performance on the larger LLaVA-1.5-13B backbone under all pruning ratios. With 128 retained tokens, it preserves 98.7% of the original performance, reaching an average score of 65.3 and outperforming CDPruner by 0.5 average points. When the budget is reduced to 64 tokens, CausalPrune retains 97.8% relative performance, exceeding CDPruner and DivPrune by 1.2 and 3.7 percentage points, respectively. Under the most aggressive 32-token setting, CausalPrune still preserves 95.9% of the full-token performance and improves over CDPruner by 1.0 average point. Notably, CausalPrune consistently maintains strong POPE performance across all budgets, even surpassing the unpruned baseline, further confirming that counterfactual causal token selection remains effective for larger LVLM backbones.

D.5 Comparisons on LLaVA-NeXT-7B

To examine whether CausalPrune generalizes from fixed-resolution LVLMs to dynamic-resolution architectures, we conduct additional experiments on LLaVA-NeXT-7B, whose full-token setting contains 2,880 visual tokens. As shown in Table 6, CausalPrune achieves the best average performance across all token budgets. With 640 retained tokens, it preserves 100.3% of the original performance, reaching an average score of 65.4 and slightly surpassing the full-token baseline. When the budget is reduced to 320 tokens, CausalPrune retains 99.1% relative performance, exceeding CDPruner by 1.1 percentage points and 0.7 average points. It also achieves strong results on POPE, MME, MMBench, and MM-Vet, suggesting that causal token selection preserves both hallucination robustness and high-level multimodal reasoning ability under dynamic-resolution pruning. Under the most aggressive 160-token setting, CausalPrune still preserves 97.8% of the full-token performance and improves over CDPruner by 1.0 average point. The consistent gains under this highly compressed setting indicate that CausalPrune can retain task-critical evidence even when most visual tokens are removed. These

Table 9: Generalization evaluation of visual token pruning methods on video understanding benchmarks with Qwen2.5-VL-7B under a 75% pruning ratio. The results demonstrate the effectiveness of CausalPrune beyond image-based LVLMs by preserving video understanding performance with reduced visual token budgets.

Method	Venue	Vinoground (Group)	VideoMME (Perception)	SEED-Bench (All)	Avg.
Qwen2.5-VL-7B (baseline)	-	20.20	61.33	74.20	100.0%
FastV	ECCV-2024	12.80	59.44	69.22	84.56%
VisionZip	CVPR-2025	12.80	59.67	72.11	85.98%
DivPrune	CVPR-2025	14.00	58.00	72.11	87.06%
DART	EMNLP-2025	12.60	59.52	71.27	85.19%
IDPruner	arXiv-2026	13.40	59.48	72.68	87.13%
CausalPrune (Ours)	This paper	**14.80**	**60.20**	**73.50**	**88.90%**

results show that counterfactual causal token selection remains robust when pruning from a much larger visual-token pool, further demonstrating the scalability of CausalPrune to high-resolution LVLM architectures.

D.6 Comparisons on InternVL3-8B

Beyond the LLaVA-family architectures, we further evaluate CausalPrune on InternVL3-8B, an advanced open-source LVLM with a ViT-MLP-LLM architecture and extended multimodal contexts. As shown in Table 7, CausalPrune consistently achieves the best average performance under both pruning ratios. With 256 retained tokens, it preserves 93.9% of the full-token performance, reaching an average score of 79.0 and outperforming CDPruner by 0.9 average points. When the budget is further reduced to 128 tokens, CausalPrune still retains 85.6% relative performance and improves over CDPruner by 1.3 average points. It also achieves consistent gains on text-oriented and reasoning-intensive benchmarks such as TextVQA, ChartQA, OCRBench, MME, and MMBench. These results indicate that counterfactual causal token selection is not tied to a specific LVLM family, but can generalize to stronger open-source multimodal architectures with different visual encoding and alignment designs.

D.7 Generalization to Video Understanding

Although CausalPrune is developed for image-based LVLMs, its causal token importance estimation is independent of the input modality and can be naturally extended to video-language models. We evaluate this generalization ability on Qwen2.5-VL-7B using three representative video benchmarks: Vinoground (Zhang et al., 2024a), VideoMME (Fu et al., 2025), and SEED-Bench (Li et al., 2024). Vinoground evaluates fine-grained video-text alignment with the Group metric, VideoMME measures video perception ability with the Perception subset, and SEED-Bench reports overall video understanding performance with the All metric. Table 9 summarizes the comparison with existing video token pruning approaches. CausalPrune achieves the best average performance among all pruning methods, demonstrating that causal token selection remains effective under temporal visual redundancy. Compared with IDPruner, CausalPrune improves the average relative performance from 87.13% to 88.9%. This suggests that causal importance estimation can better preserve task-relevant visual evidence while removing redundant video tokens.

D.8 Computational Overhead Analysis

Although CausalPrune introduces additional computation for causal token scoring, the overhead is limited compared with the latency reduction brought by visual token pruning. We profile the end-to-end online prefill latency on LLaVA-NeXT-7B using the POPE benchmark under the same hardware environment as our main experiments, *i.e.,* PyTorch 2.1 on NVIDIA RTX 4090 GPUs. The full-token baseline contains 2,880 visual tokens, while CausalPrune retains 320 tokens. As shown in Table 8, CausalPrune introduces 14.8 ms additional scoring overhead per sample, including gradient-based CTIS estimation, backdoor correction, layer-wise CTIS computation, and top-k token selection. After adding this overhead, the total prefill latency of CausalPrune is 46.8 ms, which is still substantially lower than the 246.0 ms full-token baseline, corresponding to an 81.0% reduction.

Table 10 further breaks down the scoring cost. The main overhead comes from gradient-based CTIS estimation, while backdoor correction and layer-wise CTIS computation only introduce lightweight vector operations once gradients, hidden states, and cached attention statistics are available. Compared with CDPruner, CausalPrune has slightly higher scoring overhead but achieves better POPE F1, suggesting that the additional causal estimation cost leads to a more favorable accuracy–efficiency trade-off. The offline calibration step is performed once before evaluation and is not included in online inference latency.

Table 10: Breakdown of the online scoring overhead of CausalPrune on LLaVA-NeXT-7B.

Component	**Latency (ms/sample)**	**Share**
Gradient-based CTIS estimation	9.6	64.9%
Backdoor correction	1.8	12.2%
Layer-wise CTIS computation	2.6	17.6%
Top-k token selection	0.8	5.4%
Total scoring overhead	14.8	100.0%

D.9 Analysis: When Does Causal Pruning Help Most?

To understand when CausalPrune's causal reasoning provides the largest benefit, we analyze its performance as a function of the spurious correlation strength in the input. We measure spurious correlation strength as the average cosine similarity between visual tokens and text tokens in the observational distribution, which serves as a proxy for the confounding bias B_i^l in Theorem 1. Figure 6 shows that CausalPrune's accuracy advantage over FastV grows monotonically with spurious correlation strength, confirming that causal reasoning is most beneficial in high-confounding scenarios. At low spurious correlation (strength < 0.5), the advantage is modest ($\sim$0.2%), as attention scores are already reasonable proxies for causal importance. At high spurious correlation (strength > 4.0), the advantage grows to over 3%, demonstrating that CausalPrune is particularly valuable for images with strong contextual biases.

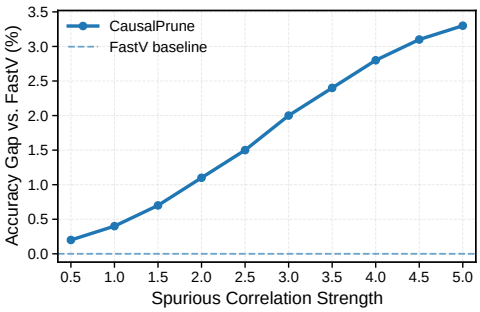


Figure 6: Accuracy gap relative to FastV across different visual-text spurious correlation strengths.

E Limitations and Future Works

Although CausalPrune achieves favorable accuracy–efficiency trade-offs across multiple LVLMs, it still has several limitations. First, the current CTIS estimation relies on a first-order approximation, which provides an efficient but approximate measure of counterfactual token importance. More accurate higher-order or sampling-based estimators may further improve pruning reliability, especially under extremely aggressive token budgets. Second, our causal correction mainly focuses on image-based LVLM inference, while temporal dependencies in video LVLMs are not explicitly modeled. Future work will extend the proposed causal correction framework to broader attention-based importance estimation scenarios, including interpretability, knowledge distillation, and active learning. We also plan to generalize CausalPrune to video LVLMs through temporal causal graph modeling, enabling causal token selection over both spatial and temporal dimensions.