ADVERSARIAL BOTTLENECK EXPLANATIONS FOR SINGLE-PAIR CLIP-LIKE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explaining a frozen vision–language model at the level of one image–text pair requires attributing a matching score without retraining or auxiliary examples. We study a bottleneck approach that separates feature selection from attribution readout. The Adversarial Bottleneck Method (ABM) searches over bounded intermediate-feature gates using the paired cosine score and assigns importance from a final-state Gaussian KL score. This removes the explicit compression–relevance coefficient used by M2IB while retaining a configurable perturbation path. We derive a finite-step bound for the projected gate formulation that accounts for boundary clipping and provides an explicit sufficient condition for one-step surrogate ascent. The bound does not imply optimal compression or attribution faithfulness. Reported experiments on Conceptual Captions, ImageNet, and Flickr8k show favorable image-side Confidence Drop point estimates under the inherited evaluation protocol, with complementary ROAD, component, and AltCLIP results. The submitted text wrapper has unresolved validity defects, and the historical measurements lack an exact run-to-implementation binding; they do not yet validate the reference formulation. ABM is a method for inspecting a model's matching score, with theoretical support for its update geometry rather than for causal or human-semantic explanations.

1 INTRODUCTION

CLIP (Radford et al., 2021) and related dual encoders compare image and text embeddings through a cosine score. An explanation of a particular match should identify evidence in both modalities while keeping the target model fixed. We study this single-pair setting: no backbone retraining, auxiliary contrastive examples, or concept vocabulary is required at explanation time. It is useful when inspecting a retrieval error or auditing why a frozen model associates one image with one caption.

Bottleneck explanations perturb intermediate features rather than train a separate explanatory predictor. M2IB (Wang et al., 2023) trades relevance against compression through a coefficient β. NIB (Zhu et al., 2025) addresses closely related sensitivity and randomness concerns through a deterministic narrowing construction. These methods motivate a narrower question: can a pair-conditioned vector gate be searched directly, and can its final state provide an attribution readout separate from the search objective? This is the question addressed by ABM, rather than a claim to introduce information bottlenecks or parameter-free attribution.

ABM combines a feature–noise interpolation, projected sign-gradient search on paired cosine similarity, and a final-state KL score. The term *adversarial* refers to a gradient-directed perturbation of intermediate features; it does not denote min–max training or certified robustness. Direct gates avoid the sigmoid parameterization's chain-rule attenuation, but the box constraint does not itself impose sparsity. Likewise, a final-state readout does not accumulate finite-step integration error, but it is not a completeness identity for the original prediction.

Our contributions are: (i) a single-pair gate-search and readout construction; (ii) an explicit displacement-dependent bound separating feasible first-order alignment from finite-step ascent; and (iii) a protocol-specific assessment of the reported confidence, ROAD, throughput, and component results. We distinguish the analyzed gate formulation from the archived execution record

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.

(Section 4.1). This distinction matters because a proof about one parameterization cannot certify measurements produced by an unverified implementation of another.

2 Related work

2.1 Attribution and perturbation paths

Saliency, Grad-CAM, RISE, and LIME explain predictions through gradients, activation maps, random masks, or local surrogates (Simonyan et al., 2013; Selvaraju et al., 2017; Petsiuk et al., 2018; Ribeiro et al., 2016). Applying them to a dual encoder requires specifying the paired score and modality-specific perturbation. Integrated Gradients provides an exact path-integral perspective with a chosen baseline (Sundararajan et al., 2017). AGI, MFABA, and AttEXplore instead construct perturbation paths using gradients (Pan et al., 2021; Zhu et al., 2024b;a). ABM borrows sign-directed search but uses an endpoint readout. We distinguish finite-step quadrature error from exact line integration; reversals alone do not invalidate a signed line integral.

2.2 Single-pair vision–language explanations

M2IB (Wang et al., 2023) and NIB (Zhu et al., 2025) are close bottleneck predecessors; Grad-ECLIP (Zhao et al., 2024) directly attributes CLIP matching scores. NIB uses deterministic narrowing and an information-change construction, whereas ABM searches coordinate-wise gates and scores the final gate. The submitted implementation refreshes noise and freezes its amplitude during each gradient estimate; the fixed-noise formulation below is an analytical reference. Thus, removing β is not by itself a novelty claim. The present tables contain M2IB and Grad-ECLIP but no verified matched NIB result; the method-level distinction does not establish an empirical advantage over NIB.

Other approaches study different explanatory interfaces. COCOA uses contrastive examples, LICO trains language-consistent explanations, and representation decomposition analyzes CLIP components (Lin et al., 2023; Lei et al., 2023; Gandelsman et al., 2023). CLIP Surgery and ClearCLIP modify dense-inference behavior (Li et al., 2025; Lan et al., 2024); fine-tuning and concept surrogates introduce additional learned components (Gong et al., 2025; Hoang-Xuan et al., 2025). These are relevant context, but their inputs and computational paths differ from a single-pair attribution of a frozen score. Appendix E summarizes the comparison.

3 Methods

3.1 Preliminaries

CLIP setup and notation. We follow the standard CLIP (Radford et al., 2021) setup: an image encoder $f_{\mathcal{I}} : \mathbb{R}^n \to \mathbb{R}^k$ and a text encoder $f_{\mathcal{T}} : \mathbb{R}^m \to \mathbb{R}^k$ produce k-dimensional embeddings whose cross-modal similarity is the cosine $\cos\langle f_{\mathcal{I}}(x_{\mathcal{I}}), f_{\mathcal{T}}(x_{\mathcal{T}})\rangle$. We use f to denote either encoder when the modality is clear from context.

Layer decomposition and target features. For a network with L layers we decompose $f(x) = f^{l-L} \circ f^{1-l}(x)$, where f^{1-l} maps the input to the intermediate layer l and f^{l-L} maps that representation to the final embedding. We write $z = f^{1-l}(x)$ for the target-layer activation and $\tilde{z}$ for the bottlenecked version of z used by the explanation method.

Goal. Given one fixed image-text pair, the goal is an explanation method A that returns $A(x) \in \mathbb{R}^{|x|}$, intended to rank candidate supporting coordinates; whether this ranking preserves or removes paired similarity must be assessed by a specified perturbation protocol.

3.2 Adversarial bottleneck method

We distinguish a gate-space analytical reference from the submitted stochastic implementation (Section 3.3). In this section, we first introduce the Information Bottleneck Principle (IBP) and the path-based view behind adversarial attribution. We then identify the issues that arise when these ideas

are applied directly to CLIP-like dual encoders and derive ABM as a constrained gate update over an alignment surrogate. Detailed derivations and pseudocode are provided in the Appendix.

3.2.1 INFORMATION BOTTLENECK PRINCIPLE (IBP) AND ADVERSARIAL ATTRIBUTION THEORY (AAT)

The Information Bottleneck Principle (IBP) (Tishby et al., 2000) retains task-relevant information by compressing irrelevant components, and several attribution methods build on this view (Schulz et al., 2020; Wang et al., 2023). Among these, M2IB (Wang et al., 2023) is one of the few that targets a *single* image-text pair without auxiliary samples or retraining, and is therefore one of the close comparators for this setting; classical methods such as Integrated Gradients (Sundararajan et al., 2017) or Grad-CAM (Selvaraju et al., 2017) were designed for unimodal classifiers and need to be adapted to the CLIP cosine-similarity objective and intermediate layers. The M2IB objective is the Lagrangian

$$\alpha_m^* = \arg\max_{\alpha_m} I(\tilde{z}_m, e_{m'}; \alpha_m) - \beta I(\tilde{z}_m, x_m; \alpha_m), \tag{1}$$

where $\tilde{z}_m \in \mathbb{R}^{D_m}$ is the bottlenecked encoding of modality m, $e_{m'}$ is the paired-modality embedding, x_m is the input, α_m parameterizes the bottleneck, and β trades off cross-modal relevance against input compression. M2IB realizes the bottleneck through $\tilde{z}_m = \sigma(\alpha_m) \cdot z + (1 - \sigma(\alpha_m)) \cdot \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1)$, and as $\sigma(\alpha_m) \to 0$ the latent collapses to noise and decouples from the input.

Equation 1 is a population/variational IB motivation; mutual information is not estimated from a single fixed pair.

Two practical issues motivate ABM. First, the reported CC sweep shows sensitivity to β in this setting (Appendix H.4); it does not establish sample- or checkpoint-level variation. Second, M2IB initializes near full retention ($\sigma(5) = 0.9933$), which multiplies gate gradients by $\sigma'(5) \approx 0.00665$; its optimization effect also depends on the optimizer and feature gradients. Direct gate updates remove the explicit β coefficient and sigmoid chain factor, but retain their own step-size, noise, and readout choices. Before introducing ABM, we summarize the path-based first-order view shared by adversarial attribution methods in Equation 2 (Pan et al., 2021; Zhu et al., 2024b;a).

$$\sum_{i=1}^{D} \int_0^1 \frac{\partial L(z(\tau))}{\partial z_i} \dot{z}_i(\tau)\, \mathrm{d}\tau = L(z(1)) - L(z(0)) \tag{2}$$

For a piecewise continuously differentiable path $z(\tau)$ and an objective L continuously differentiable on a neighborhood of that path, Equation 2 is the chain-rule identity. Here D is the number of flattened feature coordinates and $g = \nabla L$. Its discrete approximation has a Taylor remainder; Appendix F gives the distinction.

The corresponding finite-step update used by AGI/MFABA-style methods is

$$z^{t+1} = z^t + \Delta z^t = z^t + \eta \cdot \operatorname{sign}\left(\frac{\partial L(z^t)}{\partial z^t}\right). \tag{3}$$

The objective for a CLIP match is cosine similarity rather than a classifier loss. Moreover, finite-step gradient products approximate a line integral and can accumulate quadrature error. ABM uses a final-state readout instead. This is a different scoring design, not evidence that exact Integrated Gradients or every adversarial-path attribution method double-counts reversals.

3.2.2 REASONING FOR THE ADVERSARIAL BOTTLENECK METHOD (ABM)

Let $z \in \mathbb{R}^D$ be a flattened target activation and $e_{m'}$ the fixed paired embedding. With one noise draw $\varepsilon \sim \mathcal{N}(0, I)$ held fixed during search, define

$$\tilde{z}(\gamma) = \gamma \odot z + (1 - \gamma) \odot \varepsilon, \qquad \gamma \in [0, 1]^D. \tag{4}$$

Write $S(\gamma) = S_m(\tilde{z}(\gamma), e_{m'})$ for the resulting paired cosine. Starting from $\gamma^0 = \mathbf{1}$, the gate formulation is

$$g^t = \nabla_\gamma S(\gamma^t), \qquad \gamma^{t+1} = \Pi_{[0,1]^D}(\gamma^t + \eta_t \operatorname{sign}(g^t)). \tag{5}$$

Projection enforces feasibility without a sigmoid. Its effect on progress is explicit in the following standard smoothness-based bound.

Proposition 3.1 (Projected displacement bound). *Let $\eta_t \geq 0$, $\operatorname{sign}(0) = 0$, and $d^t = \gamma^{t+1} - \gamma^t$, and suppose ∇S is L-Lipschitz on a neighborhood of the step segment. For $g_i^t > 0$, set $r_i^t = 1 - \gamma_i^t$; for $g_i^t < 0$, set $r_i^t = \gamma_i^t$; otherwise set $r_i^t = 0$. Then*

$$S(\gamma^{t+1}) - S(\gamma^t) \geq \sum_i |g_i^t| \min(\eta_t, r_i^t) - \frac{L}{2}\|d^t\|_2^2. \tag{6}$$

In particular, the step is nondecreasing if the right-hand side is nonnegative. Every iterate remains on the coordinate-wise feature–noise interpolation.

Differentiability alone does not imply the Lipschitz-gradient assumption; cosine also requires nonzero embedding norms. The proposition gives a conditional one-step statement, not global convergence. Neither a unit step nor $1/T$ is automatically small enough. If all gradients point outwards at $\gamma = \mathbf{1}$, projection leaves the gate unchanged: removing an explicit compression penalty does not rule out full retention or prove a minimal sufficient explanation.

Proof of Proposition 3.1. For each coordinate, projection clips the step at its directional distance to the boundary, so

$$d_i^t = \operatorname{sign}(g_i^t) \min(\eta_t, r_i^t), \qquad (g^t)^\top d^t = \sum_i |g_i^t| \min(\eta_t, r_i^t) \geq 0.$$

The Lipschitz-gradient bound along the segment gives

$$S(\gamma^t + d^t) \geq S(\gamma^t) + (g^t)^\top d^t - \tfrac{L}{2}\|d^t\|_2^2.$$

Substitution yields Equation 6. Projection keeps γ^{t+1} in the box, so Equation 4 remains a coordinate-wise interpolation. These conclusions require no assumption about semantic importance or optimal compression. □

The coarse remainder bound $\|d^t\|_2^2 \leq D\eta_t^2$ can obscure clipping; retaining the actual displacement is tighter. The nonnegative linear term alone is not a monotonicity guarantee. A fixed stepsize need not satisfy the sufficient condition, and full retention is possible if every gradient at $\mathbf{1}$ is nonnegative. Direct projection prevents sigmoid chain-rule attenuation, but it does not eliminate boundary stationarity.

Final-state readout. Freeze the resulting gate and define a regularized reference channel with mean $\mu_i^T = \gamma_i^T z_i$ and variance $\nu_i^T = (1 - \gamma_i^T)^2 + \delta$, $\delta > 0$. We score

$$A_\delta(z_i) = \mathrm{KL}\big(\mathcal{N}(\mu_i^T, \nu_i^T) \,\big\|\, \mathcal{N}(0,1)\big) = \tfrac{1}{2}\big((\mu_i^T)^2 + \nu_i^T - 1 - \log \nu_i^T\big). \tag{7}$$

This is a plug-in score for a channel defined after gate selection, not the conditional law of the noise-adapted search trajectory and not an optimized mutual information. The floor δ is a readout hyperparameter, not part of S. Its influence is explicit: at $\gamma_i = 1$, $A_\delta = \frac{1}{2}(z_i^2 + \delta - 1 - \log\delta)$. Appendix A derives the score and its endpoint behavior.

At the endpoints,

$$A_\delta(1, z) = \tfrac{1}{2}(z^2 + \delta - 1 - \log\delta), \quad A_\delta(0, z) = \tfrac{1}{2}(\delta - \log(1 + \delta)).$$

Thus the floor changes the offset between retained and replaced coordinates. After channel summation, a nearly binary gate can mix retained-coordinate counts with activation magnitudes; min–max normalization does not generally remove the effect. Removing β is therefore distinct from being hyperparameter-free. The search retains choices of layer, stepsize, iteration count, initialization, and noise seed, and the readout retains δ.

From latent score to input-space attribution. For the analytical reference, the score $A(z_i)$ is computed at the target layer and must be projected back to image pixels or text tokens. Let the vision target activation have shape $H_l \times W_l \times C_l$ (so $D = H_l W_l C_l$). We aggregate the per-coordinate KL scores over the channel dimension by summation, $A_{h,w}^{\mathrm{img}} = \sum_{c=1}^{C_l} A(z_{h,w,c})$, which yields a non-negative spatial map of size $H_l \times W_l$; no ReLU is applied because Equation 7 is already non-negative, and we min-max normalize the spatial map to $[0, 1]$ before visualization. The map is then resized

to the input resolution by bilinear interpolation, following the post-processing convention used by Grad-CAM. For the text encoder, the target activation has shape $N_{\text{tok}} \times C_l$ over BPE sub-word tokens; we aggregate by summation over the channel dimension to obtain one score per sub-word token, $A_n^{\text{txt}} = \sum_{c=1}^{C_l} A(z_{n,c})$, and inherit the M2IB tokenization so that scores can be reported per sub-word or aggregated to whole-word level by mean pooling within a word. Special tokens (`[BOS]`, `[EOS]`, padding) are masked to zero before normalization. For baselines that operate at an intermediate layer, we use their native projection where prescribed and match the evaluation-space normalization where applicable; Appendix I summarizes available baseline descriptions and remaining implementation limits.

Submitted ViT mapping. The inspected vision source gates the full target-layer token tensor, including CLS, sums readout channels, discards CLS from the resulting saliency vector, reshapes the 49 patch scores to 7×7, and bilinearly resizes to 224×224 before min–max normalization. Thus its resize/normalize order differs from the analytical presentation above. A constant map gives a zero denominator in that source; no unreported fallback is presumed. These source details do not bind the historical tables to an exact run version.

What the score explains. The search objective and readout answer different questions: S selects a trajectory toward paired alignment, whereas A_δ describes the selected endpoint relative to a noise reference. A retained coordinate can reflect the chosen noise realization and local path; a large KL value is not a certificate of causal necessity. The complete procedure is: replace the target features, search the bounded gates, evaluate the endpoint score, and aggregate to spatial locations or tokens. Algorithm 1 specifies the gate formulation; its relation to feature-space updates is stated separately in Section 3.3.

3.3 Parameterization and implementation contract

For $z_i \neq 0$, suppose an implementation uses $x_i = \gamma_i z_i$ and optimizes the exact composite objective

$$\bar{S}(x) = S_m\big(x + (1 - x/z) \odot \varepsilon, e_{m'}\big),$$

where division is coordinate-wise. The chain rule gives $\partial S/\partial \gamma_i = z_i \partial \bar{S}/\partial x_i$. Consequently an unprojected update

$$x_i^+ = x_i + |z_i| h \operatorname{sign}(\partial \bar{S}/\partial x_i)$$

induces $\gamma_i^+ = \gamma_i + h \operatorname{sign}(\partial S/\partial \gamma_i)$. Projection must be onto $[\min(0, z_i), \max(0, z_i)]$ in x to correspond to a box projection in γ. For the stated scaling $h = 1/T$, this is a $1/T$ gate step, not a unit gate step.

The submitted source instead computes $\gamma^t = x^t/z$, holds its noise amplitude fixed while differentiating a fresh leaf x, averages $R = 5$ feature gradients, and uses $h = 1/T$. For nonzero z_i, define the step-specific surrogate

$$Q_t(\gamma) = \frac{1}{R} \sum_{r=1}^{R} S_m\big(z \odot \gamma + (1 - \gamma^t) \odot \varepsilon_{tr}, e_{m'}\big)\,.$$

With those draws and γ^t frozen, $\nabla_\gamma Q_t = z \odot \bar{h}_t$, where $\bar{h}_t$ is the averaged feature gradient. The sign and interval-projection mapping therefore yields

$$\gamma^{t+1} = \Pi_{[0,1]^D}\big(\gamma^t + T^{-1} \operatorname{sign}(\nabla Q_t(\gamma^t))\big)\,.$$

Proposition 3.1 applies to this single Q_t under its smoothness condition. The functions Q_t change with the amplitude and noise draws, so the statement does not imply monotonicity of one fixed objective or its expectation. In particular, the full derivative of the reference interpolation has a factor $z - \varepsilon$, whereas this surrogate has a factor z. We do not identify the two. At $z_i = 0$, the code's division x_i/z_i is undefined and subsequent NaN summation is not an equivalence proof.

The submitted text wrapper passes a text embedding to an argument intended to hold the image embedding. In addition, the statements `lamb[0]=1` and `lamb[-1]=1` index the batch axis; for a batch of one they overwrite the entire gate, rather than selecting boundary tokens. These facts preclude treating this wrapper as a verified implementation of pair-conditioned token attribution.

Run-bound artifacts showing that a different version generated the historical text tables have not been established here.

The submitted readout uses $v = (1 - \gamma)^2$ and

$$\widehat{A}_\delta = \tfrac{1}{2}\big(\mu^2 + v - 1 - \log(v + \delta)\big), \qquad \delta = 10^{-9}.$$

For finite coordinates this equals Equation 7 minus $\delta/2$ at the same δ. A spatially constant channel-sum offset cancels under min–max normalization only when all locations have the same number of valid channels and a nonzero range. This does not justify replacing 10^{-9} by 10^{-6}, nor does it resolve zero-feature or text-wrapper behavior. Algorithm 1 is explicitly the analytical reference; the submitted stochastic implementation is characterized above.

4 Experiments

4.1 Experimental setup

We evaluate ABM mainly on CLIP ViT-B/32 (Dosovitskiy et al., 2020; Radford et al., 2021), with additional AltCLIP (Chen et al., 2023) results in Appendix H.1. Experiments span three datasets: Conceptual Captions (Sharma et al., 2018), ImageNet (Deng et al., 2009), and Flickr8k (Hodosh et al., 2013). The tables were reported under an inherited M2IB framework; Appendix C records recoverable source details and unresolved protocol/run binding.

Baselines include M2IB (Wang et al., 2023), RISE (Petsiuk et al., 2018), Grad-ECLIP (Zhao et al., 2024), Grad-CAM (Selvaraju et al., 2017), Chefer et al. (Chefer et al., 2021), Saliency Map (Simonyan et al., 2013), MFABA (Zhu et al., 2024b), and FastIG (Hesse et al., 2021). For unimodal baselines adapted to CLIP, we use the same image-text cosine-similarity objective whenever applicable, while CLIP-specific methods use their native objectives. Baselines use native or recommended settings where available, shared evaluation components are described by the archive but exact per-table configuration matching is unverified, and ABM hyperparameters are fixed globally rather than selected separately per dataset or metric. All experiments are run on Linux with CUDA 12.4 and two NVIDIA A100 GPUs.

Reported settings and execution correspondence. The reported target layer is $l = 9$ with $T = 10$ and full-retention initialization. Layer and iteration defaults were selected from the reported evaluation sweep, not an independent held-out tuning set. The image experiments use CLIP ViT-B/32; AltCLIP is reported separately. The report uses a fixed 1,000-pair Conceptual Captions subset, the inherited ImageNet prompt protocol, and Flickr8k with one caption per image. Component results average seeds 13, 17, 23, 29, and 31 across the datasets.

Inspection of the submitted supplementary code establishes that its vision branch updates features by $|z|/T$ times the sign of an average of five fresh-noise gradients. This induces a $1/T$ gate step for nonzero features, but the noise amplitude is held fixed during differentiation. It is not the full gradient of Equation 4. Section 3.3 derives the corresponding per-step surrogate and identifies the exact scope of the projected bound. The code also uses 10^{-9} inside the logarithm of its readout, whereas the earlier manuscript specified a 10^{-6} variance floor. The original tables lack a run-bound version record resolving this discrepancy.

The submitted text wrapper has separate paired-embedding and batch-indexing defects (Section 3.3). Accordingly, historical text entries and qualitative text maps are retained only as archival records, not evidence of a valid cross-modal text attribution or superiority. No corrected implementation is silently substituted for the one associated with those results. The image values are archived point estimates without an exact generating revision; they do not yet validate either algorithmic variant. Their provenance must be resolved before an empirical method claim is established.

Protocol and uncertainty. The main confidence and ROAD tables contain single-run point estimates; the component table uses five seeds and a different aggregation. They cannot establish paired statistical superiority by cross-table comparison. M2IB uses its recommended default rather than matched tuning, and the ROAD, transfer, and throughput tables contain smaller comparator sets than the main table. The historical text protocol rescales embedding rows; a shared weighting rule need not produce equal effective attenuation.

Table 1: Baseline audit and claim-to-evidence map. "Matched" = same evaluation control unless noted; per-baseline source, adapted objective, and target-layer/default settings are listed in Appendix I.

Item / claim	ABM setting	Comparator / evidence	Audit status
Backbone and data	CLIP ViT-B/32; AltCLIP appendix	matched CLIP-family checkpoints	evaluated scope only
Preprocess and prompts	inherited M2IB protocol	shared evaluation framework	realized mask sizes may differ
Baselines	target layer 9, $T{=}10$	M2IB, RISE, Grad-ECLIP, Grad-CAM, Chefer et al., Saliency, MFABA, FastIG	chosen from reported sweep
Primary metrics	Confidence Drop ↓; Confidence Increase ↑	Table 2	point estimates
Image-side ROAD	aggregate ROAD ↑	Table 3	aggregate only
Cross-model behavior	AltCLIP	Appendix H.1	archived point estimates
Efficiency	CLIP on A100	Appendix H.2	descriptive FPS

4.2 Evaluation metrics

We retain Confidence Drop and Confidence Increase for comparability with M2IB (Wang et al., 2023; Chattopadhay et al., 2018). If Δs is the returned text confidence change, text Drop is $\max(-\Delta s, 0)$, whereas text Increase is $\mathbf{1}[\Delta s > 0]$ (Appendix D). These are unscaled per-pair statistics; the archived evaluator reports their means multiplied by 100. Thus Increase is based on a Boolean per-pair indicator. The inherited Drop units are not calibrated across modalities. Small text Drop can reflect weak effective embedding attenuation and is not evidence from a matched token-removal experiment.

The internal ablation additionally reports token deletion AUC (↓), insertion AUC (↑), comprehensiveness@20 (↑), and sufficiency@20 (↓), with @20 referring to the top 20% tokens. These historical entries were intended as component diagnostics; unresolved text execution and variant definitions prevent validated component conclusions. They are not ABM–M2IB equal-budget comparisons. Image-side ROAD (Rong et al., 2022) provides a removal-based diagnostic. Gate optimization perturbs an intermediate representation, whereas input masking re-encodes the entire model; these are different functions sharing the same final cosine score. Consequently the evaluation is not an algebraic restatement of S, but it is also not independent evidence of human or causal validity.

Recovered evaluator operations. The submitted evaluator supplies cosine-similarity targets to image-confidence functions imported from `pytorch_grad_cam` and multiplies their outputs by 100. The dependency revision is not pinned by the inspected record, so an absolute-versus-relative Drop formula cannot be assigned to the original table without recovering that version. The supplied text evaluator explicitly returns $100\max(s - s', 0)$ and $100\mathbf{1}[s' > s]$. It scales token-embedding rows in a model copy by the saliency weights after stripping boundary IDs, rather than physically deleting a rank-budgeted token subset; repeated token IDs cause repeated updates to the same embedding row. This historical protocol must not be equated with the separate continuous deletion study. Neither its exact run binding nor the ROAD schedule is established here.

4.3 Experimental results

Table 2 presents image-side point estimates across three datasets. ABM has the lowest reported Confidence Drop on each dataset and the largest Confidence Increase on CC and ImageNet. On Flickr8k Confidence Increase, M2IB obtains 28.1 and ABM 26.8. We interpret these as protocol-specific point estimates. Historical text entries are not used to support a performance claim because their implementation provenance is unresolved.

Compared with Grad-ECLIP, ABM substantially reduces the image-level Confidence Drop on ImageNet (2.3199 to 0.746) while increasing image Confidence Increase (29.2 to 57.1), a favorable point-estimate comparison in this particular masking protocol. Compared with M2IB, ABM has lower reported image Confidence Drop on all main-table datasets and higher image Confidence Increase on two of three datasets, with the Flickr8k image Confidence Increase exception noted above. These archived numerical comparisons motivate a reproducible evaluation of ABM; they do not establish a validated method-level advantage.

To further complement our evaluation, we include image-side results under the **ROAD** (Rong et al., 2022) metric, which assesses the ability of an XAI method to discriminate between the most

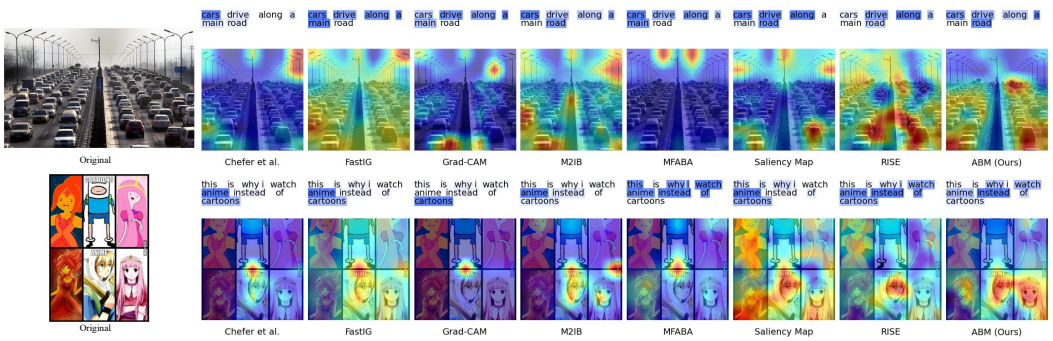

Figure 1: Qualitative comparison on Conceptual Captions. Rows are image-text pairs; columns are methods (ABM rightmost). Within each pair, the top sub-row is per-token text attribution (darker blue = higher), and the bottom sub-row is the image attribution heatmap (warmer = higher); maps are min-max normalized. In the row 1 "cars drive along a main road" example, ABM concentrates image evidence on the cars/road; the archived text map highlights matching nouns but is not validated pair-conditioned attribution; some path- or gradient-based baselines yield more diffuse maps in this example (per-example illustration, not a general claim).

Table 2: Image-side Confidence Drop and Confidence Increase point estimates under the inherited protocol. No matched tuning or paired uncertainty is established. Historical text entries with unresolved implementation provenance are separated in Appendix K.

Method	CC		ImageNet		Flickr8k	
	Drop ↓	Incr. ↑	Drop ↓	Incr. ↑	Drop ↓	Incr. ↑
M2IB	1.1171	39.3	1.1615	49.4	1.4731	28.1
RISE	1.4197	28.8	1.001	54	3.01	5.7
Grad-ECLIP	2.3956	26.5	2.3199	29.2	7.7444	2.6
Grad-CAM	4.1064	20.2	2.5483	33.9	5.1869	13.6
Chefer et al.	2.0138	33.65	1.6636	44	2.6214	26.8
Saliency Map	10.4351	2.95	4.7331	16.4	12.154	0.1
MFABA	10.1878	2.6	5.0242	12.7	12.07	0.1
FastIG	10.5117	2.9	4.7905	16.9	12.2244	0.1
ABM	0.7878	43	0.746	57.1	1.1169	26.8

and least important pixels. Higher ROAD values indicate greater separation under a particular removal/imputation protocol. As shown in Table 3, ABM obtains a slightly higher reported aggregate ROAD point estimate (1.8867) than close baselines such as Chefer et al. (1.8631) and M2IB (1.7673). The absent percentile schedule, imputation settings, and dataset-level aggregation prevent a fully specified comparison; the small aggregate difference is not an established ranking.

Figure 1 illustrates reported spatial responses on selected pairs. The displayed text maps have unresolved implementation validity and are not evidence of pair-conditioned token attribution. These examples are not a representative human-grounding study. Additional archived qualitative examples appear in Appendix G.

We also evaluated ABM and other baselines on AltCLIP (Chen et al., 2023) in Appendix H.1. The archived image-side point estimates are favorable in this additional CLIP-family setting, although ABM is not highest on every image metric (for example, M2IB is slightly higher on ImageNet image

Table 3: Aggregate image-side ROAD point estimate (↑) under the archived masking protocol (schedule and aggregation not fully specified); no uncertainty intervals or Grad-ECLIP result are reported in this table.

Metric	ABM	M2IB	Chefer et al.	Grad-CAM	MFABA	RISE	Saliency Map	FastIG
ROAD	**1.8867**	1.7673	1.8631	1.5427	-0.2851	0.0528	-0.4216	1.3211

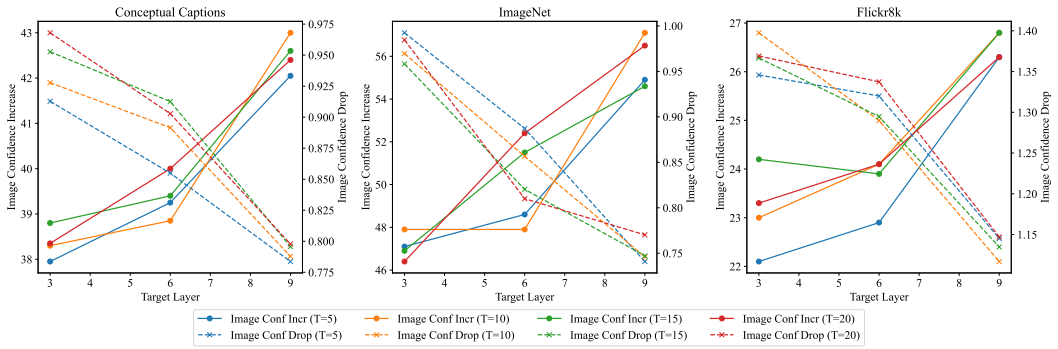

Figure 2: Impact of target layer and iteration count T on ABM image interpretability across Conceptual Captions, ImageNet, and Flickr8k. Shown are Confidence Drop and Confidence Increase for target layers 3, 6, and 9, and $T \in \{5, 10, 15, 20\}$. These scores quantify the reported masking response, not human interpretation quality. The historical report selected target layer 9 and $T = 10$ from this sweep; no held-out or single scalar selection rule is established.

Table 4: Component ablation: macro averages over CC/ImageNet/Flickr8k with five fixed seeds (13, 17, 23, 29, 31). Historical values; text implementation provenance is unresolved. Bold marks numerical extrema only, not validated text-attribution performance or paired significance.

Variant	Image Conf Drop ↓	Image Conf Incr ↑	Text Deletion AUC ↓	Text Insertion AUC ↑	Text Comp.@20 ↑	Text Suff.@20 ↓
Full ABM	**1.078**	**37.31**	**0.2299**	**0.2616**	**0.0411**	**0.0579**
w/o Adv	1.841	25.68	0.2641	0.2262	0.0122	0.0983
w/o Noise	1.260	30.45	0.2507	0.2420	0.0176	0.0882
w/o Projection	1.948	22.29	0.2533	0.2393	0.0149	0.0892
w/o KL	1.584	25.95	0.2543	0.2385	0.0170	0.0886
Random sparsity	2.732	24.83	0.2468	0.2465	0.0201	0.0822

Confidence Increase). For computational efficiency analysis, we report the frames per second (FPS) of each method under the image modality in Appendix H.2. ABM is faster than M2IB in this setting, but several simpler baselines have higher FPS.

4.4 Ablation Studies

We evaluate the impact of target layer l and iteration count T on ABM's image interpretability. As shown in Figure 2, the historical defaults are layer 9 and T=10; the record does not define a scalar rule for balancing its two metrics. Appendix H.3 further evaluates early iterations (T=1, 2, 3, 4) for both image and text, retaining historical values whose text branch is not validated.

The full method has the favorable numerical extrema in the historical Table 4; the unresolved text implementation prevents interpreting these as validated component effects. Moreover, each component removal is also worse than random sparsity on the four continuous text criteria. Even with verified provenance, this pattern would not isolate additive or independently sufficient component effects. It does not resolve the historical confidence table's unverified text comparison and unequal effective perturbations, because M2IB is absent from this ablation.

5 Conclusion and Limitations

ABM separates a pair-conditioned gate search from a final-state bottleneck readout. The projected-displacement analysis makes the update's feasible direction and finite-step limitations explicit, and the reported image-side results motivate further examination of this mechanism. The main unresolved issues are execution correspondence, matched-budget text evaluation, close-comparator coverage, and statistical uncertainty. Retention and removal probe different properties of redundant evidence; neither establishes human-semantic or causal correctness. Attribution maps should therefore be used as model diagnostics rather than as a justification of safety-critical decisions.

AI USE STATEMENT

Generative AI tools assisted with manuscript refinement, language clarity and readability, and LaTeX formatting. The authors are responsible for the final manuscript.

REFERENCES

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8666–8682, 2023. doi: 10.18653/v1/2023.findings-acl.552.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.

Shizhan Gong, Haoyu Lei, Qi Dou, and Farzan Farnia. Boosting the visual interpretability of clip via adversarial fine-tuning. In *International Conference on Learning Representations*, 2025.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Fast axiomatic attribution for neural networks. *Advances in Neural Information Processing Systems*, 34:19513–19524, 2021.

Nhat Hoang-Xuan, Xiyuan Wei, Wanli Xing, Tianbao Yang, and My T. Thai. Advancing interpretability of clip representations with concept surrogate model. In *Advances in Neural Information Processing Systems*, 2025.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 143–160. Springer, 2024.

Yiming Lei, Zilong Li, Yangyang Li, Junping Zhang, and Hongming Shan. Lico: explainable models with language-image consistency. *Advances in Neural Information Processing Systems*, 36, 2023.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025. doi: 10.1016/j.patcog.2025.111409.

Chris Lin, Hugh Chen, Chanwoo Kim, and Su-In Lee. Contrastive corpus attribution for explaining representations. In *The Eleventh International Conference on Learning Representations*, 2023.

Deng Pan, Xin Li, and Dongxiao Zhu. Explaining deep neural network models with adversarial gradient integration. In *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Ying Wang, Tim GJ Rudner, and Andrew G Wilson. Visual explanations of image-text representations via multi-modal information bottleneck attribution. *Advances in Neural Information Processing Systems*, 36:16009–16027, 2023.

Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and Antoni B Chan. Gradient-based visual explanation for transformer-based clip. In *International Conference on Machine Learning*, pp. 61072–61091. PMLR, 2024.

Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Jason Xue, and Flora D Salim. Attexplore: Attribution for explanation with model parameters exploration. In *The Twelfth International Conference on Learning Representations*, 2024a.

Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Minhui Xue, Dongxiao Zhu, and Kim-Kwang Raymond Choo. Mfaba: A more faithful and accelerated boundary-based attribution method for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17228–17236, 2024b.

Zhiyu Zhu, Zhibo Jin, Jiayu Zhang, Nan Yang, Jiahao Huang, Jianlong Zhou, and Fang Chen. Narrowing information bottleneck theory for multimodal image-text representations interpretability. *arXiv preprint arXiv:2502.14889*, 2025. URL `https://arxiv.org/abs/2502.14889`.

A Derivation and scope of the KL readout

After selection, condition on z and the realized gate γ^T. Define a fresh reference-channel variable

$$Y_i = \gamma_i^T z_i + (1 - \gamma_i^T)\xi_i + \sqrt{\delta}\,\zeta_i, \qquad \xi_i, \zeta_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

These fresh variables define the score; no additional sampling is needed to evaluate it. Their independence gives the stated Gaussian mean and variance and hence Equation 7. By contrast, γ^T depends on the search noise ε, so the optimized search activation conditioned only on the input need not have this Gaussian law. The score is a regularized plug-in readout, not that trajectory's exact conditional KL or an MI optimum.

B Pseudocode of the Adversarial Bottleneck Method (ABM)

Algorithm 1 ABM gate-space reference formulation

Require: Frozen CLIP-like model; target-layer feature z; paired embedding $e_{m'}$; steps T; step sizes $\{\eta_t\}_{t=0}^{T-1}$; noise $\varepsilon \sim \mathcal{N}(0, I)$; variance floor δ

Init: $\gamma^0 = \mathbf{1}$
for $t = 0, \ldots, T-1$ **do**
  $\tilde{z}^t = \gamma^t \odot z + (1 - \gamma^t) \odot \varepsilon$
  Replace the target-layer activation by $\tilde{z}^t$ and compute the alignment surrogate $S_m(\tilde{z}^t, e_{m'})$
  $g^t = \nabla_{\gamma^t} S_m(\tilde{z}^t, e_{m'})$
  $\gamma^{t+1} = \Pi_{[0,1]^D}(\gamma^t + \eta_t\ \mathrm{sign}(g^t))$
end for
for $i = 1, \ldots, D$ **do**
  $\mu_i^T = \gamma_i^T z_i$
  $\nu_i^T = (1 - \gamma_i^T)^2 + \delta$
  $A(z_i) = \frac{1}{2}\left((\mu_i^T)^2 + \nu_i^T - 1 - \log \nu_i^T\right)$
end for
Aggregate $A(z)$ along the channel dimension by summation, min-max normalize, and resize to the input resolution via bilinear interpolation for image attribution; for text, aggregate along the channel dimension per sub-word token and mask special tokens.
return A

C Archived evaluator and component definitions

The evaluator implementation is described in Section 4.2.

The component labels alone do not define the alternative update when adversarial search is removed, the replacement readout without KL, the domain without projection, or the support law/budget for random sparsity. Table 4 therefore records historical aggregates rather than a fully specified causal isolation of the components. Their original configurations are required before attributing the changes to a particular mechanism.

The submitted implementation labeled Grad-CAM computes $\mathrm{ReLU}(\sum_c z_{pc}\,\partial s/\partial z_{pc})$ at vision encoder block index 9, followed by CLS removal, patch reshaping, resizing, and guarded min–max scaling. This is an activation-times-local-gradient variant, not the standard spatially averaged Grad-CAM formula. Historical tables retain the source label; no standard-Grad-CAM comparison is certified by that label alone.

D Text confidence metrics

For the text modality, we follow the evaluation method used in M2IB. The Confidence Increase implementation is Boolean: it records whether the confidence change after perturbation is greater than zero rather than preserving the magnitude of that change. As a result, changes in the number of iterations T can be hidden by the metric even when the underlying attribution scores differ. We

retain these values only as historical records. Continuous token metrics do not remedy the unresolved text-wrapper defects. The evaluation code is as follows:

```
class DropInConfidenceText(CamMultImageConfidenceChange):
   def __init__(self):
      super(DropInConfidenceText, self).__init__()

   def __call__(self, *args, **kwargs):
      scores = super(DropInConfidenceText, self).__call__(*args,
                                                **kwargs)
      scores = -scores
      return np.maximum(scores, 0)

class IncreaseInConfidenceText(CamMultImageConfidenceChange):
   def __init__(self):
      super(IncreaseInConfidenceText, self).__init__()

   def __call__(self, *args, **kwargs):
      scores = super(IncreaseInConfidenceText, self).__call__(*args,
                                                 **kwargs)
      return np.float32(scores > 0)
```

E Positioning of ABM against representative comparators

Table 5 consolidates how ABM differs from the closest CLIP-oriented and adversarial-attribution methods along four axes that matter for the single-pair frozen-model setting. Note that AGI and MFABA are not native CLIP image-text single-pair attribution methods: they were developed for unimodal classifier attribution and would require redefining the loss, target representation, and modality-specific perturbation policy to apply to CLIP. We list them here as path-attribution ancestors for context, not as direct CLIP baselines.

F Exact integration versus finite-step approximation

The chain rule along a piecewise differentiable path gives the exact identity in Equation 2. For sampled path points $z^0, \ldots, z^T$, define $\Delta z^t = z^{t+1} - z^t$. If ∇L is K-Lipschitz on the step segments, Taylor expansion instead gives

$$\left| L(z^T) - L(z^0) - \sum_{t=0}^{T-1} \nabla L(z^t)^\top \Delta z^t \right| \leq \frac{K}{2} \sum_{t=0}^{T-1} \|\Delta z^t\|_2^2. \tag{8}$$

For a concrete one-dimensional example, let $L(z) = \frac{1}{2}z - \frac{1}{2}z^2$ and use unit sign-ascent steps from $z^0 = 1$. Its derivative is $1/2 - z$, so the path is $1 \to 0 \to 1$. The left-endpoint gradient products sum to $(-1/2)(-1) + (1/2)(1) = 1$, although $L(1) - L(1) = 0$. The exact signed line integral on the same closed path is zero; the discrepancy is quadrature error, not a failure of the fundamental theorem of calculus. ABM's endpoint readout avoids this particular accumulation, but relinquishes path-completeness semantics. This example neither establishes a generic defect in Integrated Gradients nor proves superiority to the implementations of AGI or MFABA.

G Additional qualitative visualization explanations

H Additional experiments

H.1 Experiment results on AltCLIP model

We conducted additional experiments on AltCLIP to test whether the CLIP results transfer to another CLIP-family model. Table 6 records favorable image-side point estimates in this setting. Its text values are excluded from comparative conclusions.

Table 5: Positioning of ABM against representative methods. "Single-pair" means explaining one fixed image-text pair without auxiliary contrastive samples or labels. "Frozen" means no extra training or architectural change. AGI / MFABA are unimodal-only path-attribution ancestors and require adaptation to CLIP; we evaluate adapted MFABA in Tables 2 and 3.

Method	Single-pair (CLIP-native)	Frozen	Objective optimized	Attribution readout
M2IB (Wang et al., 2023)	yes	yes	MI-style bottleneck w/ tunable β	bottleneck-derived attribution
NIB (Zhu et al., 2025)	yes	yes	deterministic narrowing construction	information-change attribution
Grad-ECLIP (Zhao et al., 2024)	yes	yes	cosine-similarity gradient	gradient $\times$ activation map
AGI / MFABA (Pan et al., 2021; Zhu et al., 2024b)	no (unimodal only; not CLIP-native)	yes	classifier-loss adversarial path	accumulated path contribution
CLIP Surgery / ClearCLIP (Li et al., 2025; Lan et al., 2024)	no (dense)	modified	dense-inference attention rewrite	attention/feature map
COCOA / LICO (Lin et al., 2023; Lei et al., 2023)	no	no	contrastive samples or extra model	learned alignment
ABM (ours)	yes	yes	projected gate ascent on cosine surrogate	final-state KL of Gaussian channel to prior

Table 6: AltCLIP image-side point estimates, with unvalidated historical text entries retained for traceability. Text columns are excluded from comparative conclusions.

Dataset	Method	Image		Text	
		Conf Drop (↓)	**Conf Incr (↑)**	**Conf Drop (↓)**	**Conf Incr (↑)**
Conceptual Captions	M2IB	1.3324	32.0	4.8767	22.25
	RISE	1.3115	27.6	1.6627	32.95
	Grad-CAM	6.6955	9.6	8.0304	8.9
	Chefer et al.	5.3088	13.2	6.2125	11.30
	SaliencyMap	1.2415	32.5	2.6278	24.35
	FastIG	1.8274	26.8	2.6237	30.65
	ABM(Ours)	**0.9183**	**36.8**	0.0020	47.65
ImageNet	M2IB	1.6074	**45.5**	5.7696	4.4
	RISE	2.0764	35.2	4.6092	4.2
	Grad-CAM	8.1205	5.8	7.0187	3.7
	Chefer et al.	7.0399	6.8	7.2695	3.4
	SaliencyMap	1.7094	40.0	3.5131	11.1
	FastIG	1.8590	39.0	4.1739	11.8
	ABM(Ours)	**1.2385**	45.0	0.0076	21.9
Flickr8k	M2IB	1.2767	28.4	3.8204	13.0
	RISE	1.5285	19.0	2.2361	20.5
	Grad-CAM	7.5028	6.2	9.0013	2.5
	Chefer et al.	5.5111	11.1	9.3455	1.3
	SaliencyMap	1.4095	24.7	2.6532	19.8
	FastIG	2.5895	12.6	2.6520	25.1
	ABM(Ours)	**0.8920**	**33.2**	0.0042	40.7

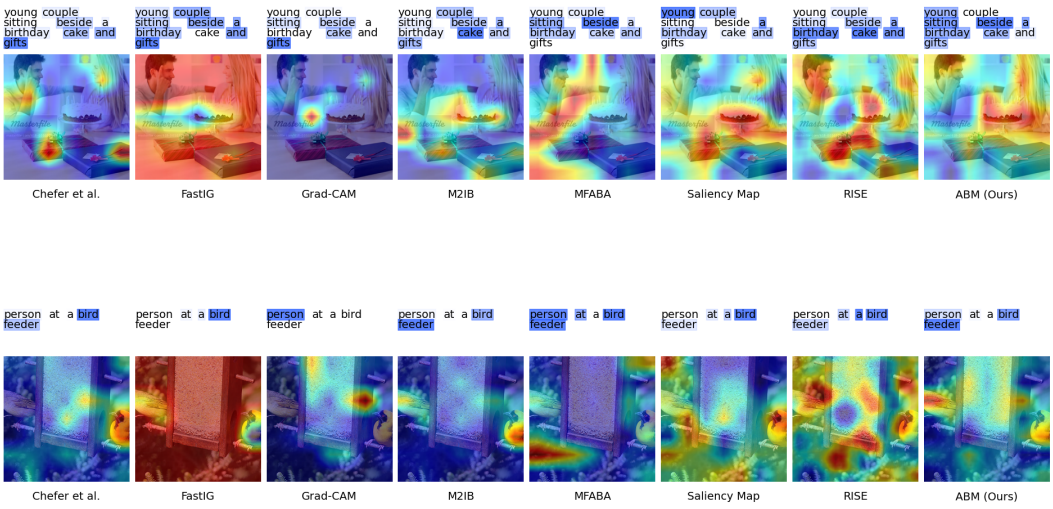

Figure 3: Additional qualitative visualization on Conceptual Captions, following the same column-wise method layout and per-pair top-token / bottom-heatmap reading convention as Figure 1.

Table 7: Archived image-modality FPS on Conceptual Captions (reported NVIDIA A100, FP32, batch size 1). Warmup count and inclusion of gradient/noise/readout/preprocessing costs are not established by a run-bound timing script.

Method	M2IB	ABM	Chefer et al.	FastIG	MFABA	RISE	Grad-CAM	SaliencyMap
FPS	1.02	1.67	5.83	6.05	0.96	0.02	4.79	6.36

H.2 Computational efficiency

We conducted experiments measuring the inference speed in terms of frames per second (FPS) on the CLIP model using the Conceptual Captions dataset, all tested on an Nvidia A100 GPU.

The archived image-side throughput is 1.67 FPS for ABM and 1.02 FPS for M2IB under the reported setup. No timing uncertainty or general speed advantage is established, and several simpler baselines have higher FPS. The timing scope must be recovered before these values can be interpreted as end-to-end explanation cost.

H.3 Additional ablation study on hyperparameter T before 5

We have conducted additional ablation experiments to provide a more comprehensive analysis of the model's behavior across different numbers of iterations (T) for both image and text modalities. These experiments include evaluations for $T = 1, 2, 3, 4$ across various datasets (Conceptual Captions, Flickr8k, ImageNet) and target layers (3, 6, 9).

As shown in Table 8, for the text modality, Confidence Drop and Confidence Increase remain nearly constant across T settings. Only text Increase is based on a Boolean per-pair indicator; Drop is clipped-continuous. Equal aggregate Drop values do not identify their cause: clipping, unchanged rankings, or realized masks can all matter. The table alone cannot establish gate binarization or insensitivity of the underlying scores. The continuous token-level entries also have unresolved implementation provenance and support no validated text-side component conclusion.

H.4 Sensitivity to the hyperparameter β

To investigate the sensitivity of M2IB to the hyperparameter β, we conducted a controlled experiment on the Conceptual Captions (CC) dataset using the CLIP model. As shown in Figure 4, M2IB

Table 8: Additional ablation study results on hyperparameter T before 5

Dataset	Target layer	steps T	Image		Text	
			Image Conf Drop	Image Conf Incr	Text Conf Drop	Text Conf Incr
Conceptual Captions	3	1	0.864	39.6	0.000983	44.5
	3	2	0.922	37.55	0.000983	44.5
	3	3	0.913	38.25	0.000983	44.5
	3	4	0.895	37.85	0.000983	44.5
	6	1	0.784	41.6	0.002515	44.55
	6	2	0.880	37.9	0.002515	44.55
	6	3	0.853	39.85	0.002515	44.55
	6	4	0.879	38.75	0.002515	44.55
	9	1	0.770	43.7	0.005019	44.4
	9	2	0.785	41.5	0.005019	44.4
	9	3	0.785	43.45	0.005019	44.4
	9	4	0.770	41.35	0.005019	44.4
Flickr8k	3	1	1.398	24	0.000947	57.8
	3	2	1.442	21	0.000947	57.8
	3	3	1.395	21.8	0.000947	57.8
	3	4	1.381	22.5	0.000947	57.8
	6	1	1.243	26.6	0.002199	60.4
	6	2	1.381	20.9	0.002199	60.4
	6	3	1.303	25.6	0.002199	60.4
	6	4	1.299	23.4	0.002199	60.4
	9	1	1.260	25.3	0.004142	60.8
	9	2	1.200	25.9	0.004142	60.8
	9	3	1.168	26.3	0.004142	60.8
	9	4	1.153	26.3	0.004142	60.8
ImageNet	3	1	0.835	47.1	0.000720	62.9
	3	2	1.035	42.4	0.000720	62.9
	3	3	1.030	45.1	0.000720	62.9
	3	4	0.999	44	0.000720	62.9
	6	1	0.878	48.8	0.001824	60.3
	6	2	0.931	46.5	0.001824	60.3
	6	3	0.957	47.1	0.001824	60.3
	6	4	0.848	50.2	0.001824	60.3
	9	1	0.873	49.7	0.004906	60.4
	9	2	0.752	53.8	0.004906	60.4
	9	3	0.754	55.5	0.004906	60.4
	9	4	0.721	56.4	0.004906	60.4

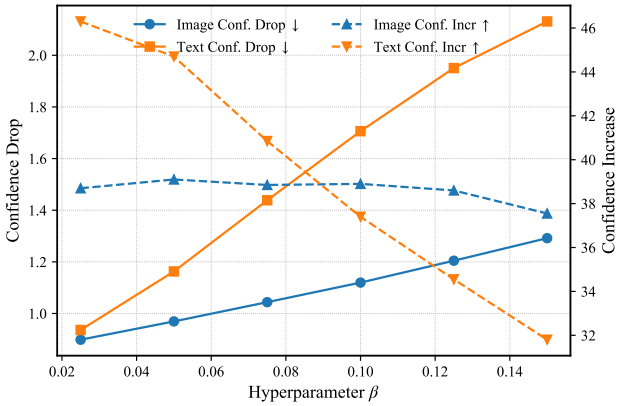

Figure 4: Performance sensitivity of M2IB with respect to the hyperparameter β on the Conceptual Captions dataset using the CLIP model. The plot illustrates the trends of four evaluation metrics: Image/Text Confidence Drop (↓) and Image/Text Confidence Increase (↑), as β varies. The results show that M2IB exhibits significant performance fluctuations across modalities, indicating that the choice of β has a substantial impact on model behavior.

exhibits considerable performance fluctuation with respect to different β values. Although the default setting in the original paper is $\beta = 0.1$, our results indicate that this may not be the optimal choice. Specifically, we observe that the confidence drop and confidence increase metrics for both image and text modalities vary significantly as β changes. For instance, smaller values of β tend to achieve lower confidence drop scores for image features, whereas intermediate values (around $\beta = 0.08$) offer more balanced performance across modalities. The inconsistency and sharp variations across all four metrics suggest that M2IB is highly sensitive to the selection of β, which raises concerns about transferability under different hyperparameter settings.

Table 9: Baseline configurations used in our experiments. "Adapted objective" refers to the score actually optimized for CLIP image-text matching after adaptation; default settings reuse the values recommended by each method's original implementation unless noted.

Baseline	Source	Adapted objective for CLIP	Target layer / key default
M2IB (Wang et al., 2023)	official release	MI-style cross-modal bottleneck w/ Gaussian channel	layer 9 (matched); $\beta = 0.1$ default
RISE (Petsiuk et al., 2018)	official release	cosine similarity over masked images	8000 random masks, mask resolution as published
Grad-ECLIP (Zhao et al., 2024)	official release	gradient of cosine similarity at chosen layer	native target layer; native softmax-style attention
Grad-CAM (Selvaraju et al., 2017)	re-implementation	gradient of cosine similarity at chosen layer	matched layer 9; channel-summed activation $\times$ gradient
Chefer et al. (Chefer et al., 2021)	official release	attention-based relevance for ViT	native ViT relevance propagation
Saliency Map (Simonyan et al., 2013)	re-implementation	input gradient of cosine similarity	no smoothing, native abs aggregation
MFABA (Zhu et al., 2024b)	official release	adversarial path w/ classifier-loss objective adapted to cosine similarity	native step count and step size
FastIG (Hesse et al., 2021)	official release	integrated gradient on cosine similarity	native step schedule
ABM (ours)	this paper (Sec. 4.1)	projected gate ascent on cosine surrogate; final-state KL readout	layer 9; $T = 10$; submitted code: five fresh-noise gradients, induced step $1/T$, log floor 10^{-9}; run provenance unresolved

Table 10: Assets used in this work. We do not redistribute restricted assets; exact license records and version identifiers remain to be verified.

Asset	Type	License / terms (best known)	Reference / notes
CLIP ViT-B/32	model	MIT (per the OpenAI CLIP release)	Radford et al. (2021); weights via official release.
AltCLIP	model	FlagAI release terms	Chen et al. (2023); weights via official release.
Conceptual Captions	dataset	Google research terms; non-redistribution	Sharma et al. (2018); we use a fixed 1000-pair subset following the M2IB protocol.
ImageNet (validation)	dataset	ImageNet research license	Deng et al. (2009); not redistributed.
Flickr8k	dataset	Flickr8k research-use terms	Hodosh et al. (2013); not redistributed.
Baseline codebases	code	Original-repository licenses (typically MIT/Apache-2.0/BSD)	M2IB / RISE / Grad-ECLIP / Chefer et al. / MFABA / FastIG; cited in Table 9.

I BASELINE CONFIGURATION TABLE

Table 9 summarizes the source, objective adaptation, and target-layer/default-setting choices used for each baseline. All baselines share the inherited M2IB masking and evaluation settings; method-specific defaults follow the original implementations. Exact per-table code/dependency binding is not established by the supplied archive.

J ASSET AND LICENSE DOCUMENTATION

We use only public datasets, model checkpoints, and baseline implementations; no new data or models are released in this submission. Table 10 summarizes the assets we rely on. The supplied record does not establish a public artifact containing all exact license files, version identifiers, and download URLs; these fields remain to be completed before a reproducible release.

We list the license/terms above to our best knowledge at submission time; exact license files, version identifiers, and download URLs still require a complete artifact manifest.

K HISTORICAL TEXT RECORDS

These values reproduce the earlier main table without changing measurements. The submitted wrapper defects and missing run-version binding in Section 3.3 prevent interpreting them as validated

Table 11: Historical text values retained for traceability only; no validated cross-method ranking is claimed. Drop is clipped-continuous; Increase averages a Boolean indicator, with both displayed after multiplication by 100.

Method	CC		ImageNet		Flickr8k	
	Drop ↓	Incr. ↑	Drop ↓	Incr. ↑	Drop ↓	Incr. ↑
M2IB	1.706	37.4	2.6018	25.4	2.0783	34.7
RISE	0.8002	43.95	0.9928	46.8	0.8914	46.4
Grad-ECLIP	1.2894	43.2	2.0104	28.3	2.1114	37.5
Grad-CAM	1.7994	34.4	2.6424	25.7	2.1823	34.2
Chefer et al.	0.9333	45.3	1.6732	29.9	1.362	42.6
Saliency Map	1.0723	40.05	1.7631	33.1	1.0797	45.9
MFABA	1.0503	36.25	1.7437	28.5	1.1551	42.6
FastIG	0.9718	41.25	1.6486	34.8	1.3098	43.9
ABM	0.005019	44.5	0.0049	60.4	0.0039	59.3

text-attribution comparisons. Numerical ordering does not establish method superiority. The same qualification applies to archived text maps, AltCLIP text columns, and token-level component/iteration tables.

L Interpretation and reproducibility boundaries

Search, readout, and faithfulness. The cosine objective chooses a gate trajectory and the regularized KL describes its endpoint. Neither establishes necessary evidence, a minimal sufficient subset, or human-semantic alignment. Retention and deletion need not rank methods identically when evidence is redundant. Both use the model's score and are sensitive to masking policy.

Text and uncertainty. The supplied text evaluator clips negative confidence changes for Drop and thresholds the sign for Increase. The historical evaluator continuously rescales token-embedding rows; a shared weighting rule does not imply equal effective perturbation strength. Its very small Drop is not evidence from a matched token-removal experiment. The continuous component study lacks M2IB and cannot resolve this cross-method issue. Paired uncertainty should use the image–text pair as the resampling unit; five noise seeds on the same pairs do not create five times as many independent examples. No unavailable interval is inferred from the aggregate tables.

Execution and comparator coverage. The submitted feature update and the analytical gate reference are different stochastic-gradient constructions, as established in Section 3.3. Exact run configurations are still required to link the historical tables to a particular version. The text-wrapper defects cannot be repaired retrospectively by editing a theorem. The baseline tables record different comparator sets; absence from ROAD, transfer, or throughput is not evidence of inferiority. NIB is a close prior method but lacks a verified matched result here. The scope remains the reported CLIP-family settings rather than universal VLM attribution.

Qualitative evidence. The figures illustrate particular model responses, not a representative human-grounding study. Their names and maps should not be used as causal explanations of the input world or as a standalone justification for content-moderation or other high-stakes decisions.