TIB: Sample-wise Tempered Information Bottleneck for Multimodal Attribution beyond Alignment Assumption

Qiming Huang **Peixi Liu** **Jianbo Jiao**
The MIx group, School of Computer Science, University of Birmingham
{qxh366, pxl416}@student.bham.ac.uk, j.jiao@bham.ac.uk

Abstract

Multimodal (*e.g.* vision–language, in this work) attribution aims to interpret the models without access to ground-truth explanatory supervision. Existing multimodal attribution methods typically rely on paired modalities as proxy supervision, implicitly assuming that image–text pairs are semantically aligned. Information bottleneck (IB)–based attribution methods follow this general paradigm and instantiate it through a global sufficiency–compression objective. In realistic multimodal datasets, however, this assumption is frequently violated due to partial semantic misalignment and unreliable cross-modal correspondences. Through theoretical analysis, we show that, under such misalignment, information-bottleneck-based attribution objectives exhibit a structural limitation: by assuming equal reliability across all proxy supervision signals, they induce posterior over-contraction and forced explanations. To address this issue, we propose TIB, a sample-wise **T**empered **I**nformation **B**ottleneck framework that adaptively modulates attribution strength according to cross-modal reliability, without assuming semantic alignment. Extensive experiments on controlled benchmarks and large-scale datasets show the effectiveness of the proposed approach, while also validating the identified failure mode. Project page: https://qiming-huang.github.io/TIB/.

"Love all, trust a few, do wrong to none."

William Shakespeare

1 Introduction

Recent advances in multimodal learning, exemplified by large-scale vision–language models such as CLIP [1], have enabled powerful joint representations across diverse modalities and achieved strong performance on tasks such as image–text retrieval and visual question answering. As these models increasingly serve as general-purpose foundations, it becomes essential to understand which modalities and components contribute to their predictions.

This growing demand has motivated increasing interest in multimodal attribution, which aims to explain model predictions by identifying salient elements across different modalities. Unlike unimodal settings—where unimodal attribution can be evaluated against well-defined supervision signals such as class labels for images or next-token prediction for language models—multimodal attribution lacks explicit ground-truth explanations. Consequently, existing approaches typically operate on paired multimodal samples and implicitly assume cross-modal consistency, treating one modality as a proxy ground truth to attribute or explain the other [2–6].

However, this assumption of cross-modal consistency is inherently fragile. As shown in Fig. 1, in most real scenarios, multimodal data, paired modalities may be missed, replaced, or only weakly

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

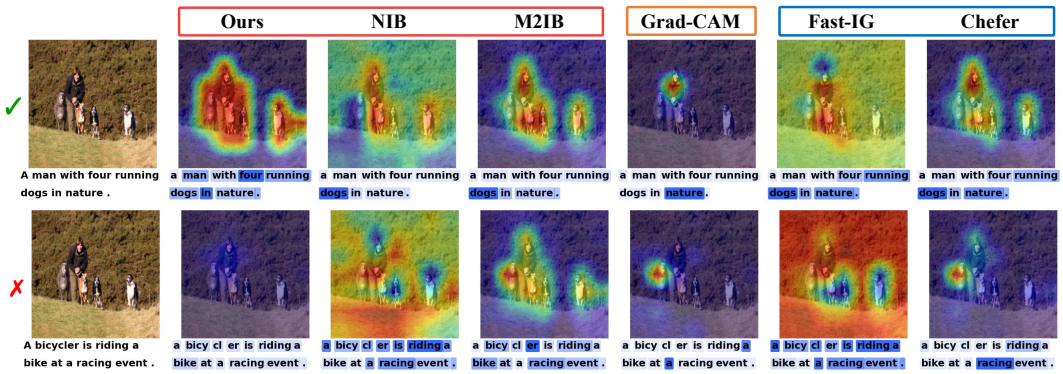

Figure 1: Qualitative comparison of multimodal attribution under completely aligned (top: ✓) and misaligned (bottom: ✗) image–text pairs. Our method suppresses attribution on completely mismatched image–text pairs.

correlated. Empirical studies on large-scale video–text datasets report that over 50% of purportedly aligned clip–caption pairs are in fact misaligned [7]. Such conditions are widely studied in multimodal learning and formally defined as *multimodal misalignment* [8, 9]. While prior work has extensively examined how misalignment affects representation learning, either as detrimental noise [10, 11] or as a form of regularization [12–14], its implications for multimodal attribution remain largely under-explored. From the perspective of attribution, this raises a fundamental question: *Does multimodal misalignment systematically affect the stability and reliability of multimodal attribution?*

In light of this, we aim to study multimodal attribution through an information-theoretic lens. In detail, we analyse the attribution objectives derived from the information bottleneck (IB) principle under a unified misalignment data generation assumption (see Assump. 3.1). Under this assumption, multimodal pairs are drawn from a mixture of aligned and misaligned components. We then characterise IB-style attribution objectives as enforcing representational sufficiency with respect to the entire data mixture, without distinguishing aligned from misaligned samples. However, the mutual information between modalities conditioned on misalignment is generally non-zero, despite the absence of genuine semantic correspondence. Consequently, the sufficiency constraint inevitably allocates representation capacity to such spurious cross-modal dependencies, diverting explanatory capacity toward noise and leading to increasingly concentrated and unstable attribution maps.

Through experimental analysis, we reveal that this failure mode is not an artefact of optimisation or hyperparameter choices, but a structural limitation of standard homoscedastic bottleneck formulations. In particular, homoscedastic information bottlenecks impose a trade-off between sufficiency and compression across all samples, treating aligned and misaligned pairs as equally reliable under the unified data generation assumption. As a result, attribution objectives cannot adapt to heterogeneity in cross-modal reliability. Besides, explanatory learning remains driven by uniform fitting pressure even when semantic sufficiency becomes statistically degenerate. In contrast, we show that stable multimodal attribution under misalignment becomes attainable by adopting heteroscedastic information bottlenecks, named as Tempered Information Bottleneck (TIB), which allows effective information capacity to vary across samples. This flexibility enables the attribution objective to suppress contributions from misaligned components within samples, while preserving explanatory capacity for their reliably aligned semantic content.

Beyond theoretical analysis, we empirically validate our approach on a synthetic benchmark that allows precise control over the structure and severity of misalignment, enabling systematic evaluation of attribution behaviour. Our contributions are summarised as follows:

- We study multimodal attribution from an information bottleneck (IB) perspective under a unified misalignment data generation assumption, and show that classical IB objectives are contaminated by multimodal misalignment pressure.
- We propose a heteroscedastic bottleneck, TIB, that enables more robust multimodal attribution by adaptively suppressing misaligned components while preserving reliably aligned semantic content.

- We validate the proposed approach on both synthetic and real-world multimodal datasets, including a controllable misalignment benchmark for systematic evaluation.

2 Related Work

Unimodal Interpretability Methods. Traditional interpretability methods were primarily developed for unimodal deep learning models and face significant limitations when applied to multimodal architectures. Early gradient-based approaches, such as Saliency Maps, compute input-output gradients to highlight important regions in the input. However, they are highly sensitive to noise and often produce low-resolution or unstable explanations. Grad-CAM [2] improves upon this by using gradients of class-specific activation maps in convolutional layers, resulting in more focused and human-interpretable heatmaps. LIME [15] adopts a black-box approach by perturbing the input and fitting a local surrogate model, offering model-agnostic explanations. Despite its generality, LIME may fail on complex models due to its reliance on local linearity assumptions. More recent attribution methods, such as AGI [16] and MFABA [17], utilise adversarial perturbations to generate more robust saliency maps. These approaches satisfy formal interpretability criteria like Sensitivity and Implementation Invariance [18], offering theoretical guarantees.

Nevertheless, most of these methods are designed for unimodal tasks and require access to downstream task labels or internal gradients, making them less suitable for large-scale vision-language models. Attempts to extend such methods to the multimodal setting—such as CLIP—face challenges due to architectural differences and a lack of task-specific supervision. As shown in our experiments, these methods often fail to provide meaningful explanations when applied directly to multimodal models.

Multimodal Interpretability Methods CLIP [1] learns joint image-text representations by training separate encoders on large-scale image-text pairs, aligning the two modalities in a shared embedding space. This design enables zero-shot transfer, allowing the model to perform various tasks based solely on natural language prompts, without requiring task-specific annotations. However, the complexity of multimodal reasoning raises the need for interpretability to ensure that the model's predictions are based on semantically meaningful features. Investigating CLIP's interpretability is thus crucial to determine whether it captures genuine vision-language associations or exploits spurious correlations.

A range of methods have been proposed to interpret vision-language pre-trained models, yet many introduce limitations in fidelity, scalability, or practicality. M2IB [5] applies a multimodal information bottleneck to filter irrelevant features, but increases architectural complexity. COCOA [19] modifies Integrated Gradients with contrastive learning, requiring additional positive and negative samples that may introduce irrelevant context. TEXTSPAN [20] relies on constructing sample-specific sets or selecting neighbours in embedding space, making it dependent on external data and less generalizable. LICO [21] retrains models to maintain cross-modal alignment, but its explanations apply to the altered model, not the original one, and are affected by training randomness. FALCON [22] explains individual features via highly activating examples, but lacks per-instance interpretability. PIDReg [23], a PID-based method, seeks to improve interpretability by decomposing information across modalities on a multimodal regression task. FIxLIP [24] introduces faithful interaction explanations of CLIP and SigLIP models, offering a game-theoretic perspective on interpreting image–text similarity.

Recent work [25] provides a comprehensive review of multimodal attribution methods, but largely focuses on fixed paired settings and does not account for multimodal misalignment arising from the data generation process itself. In practice, many existing interpretability methods for vision–language models rely on strong assumptions, such as clean image–text alignment, access to contrastive examples, or task-specific supervision, and often require additional sampling, surrogate retraining, or architectural modifications, which limit their robustness and practicality. As a result, when image–text pairs are noisy or mismatched, these methods may still produce forced explanations, undermining the reliability of the interpretability results, as illustrated in Fig. 1. By contrast, our method does not assume reliable modality alignment or strict pairwise correspondence; instead, it employs a soft alignment mechanism that adapts to varying degrees of semantic consistency, yielding more effective and credible multimodal attribution under misalignment.

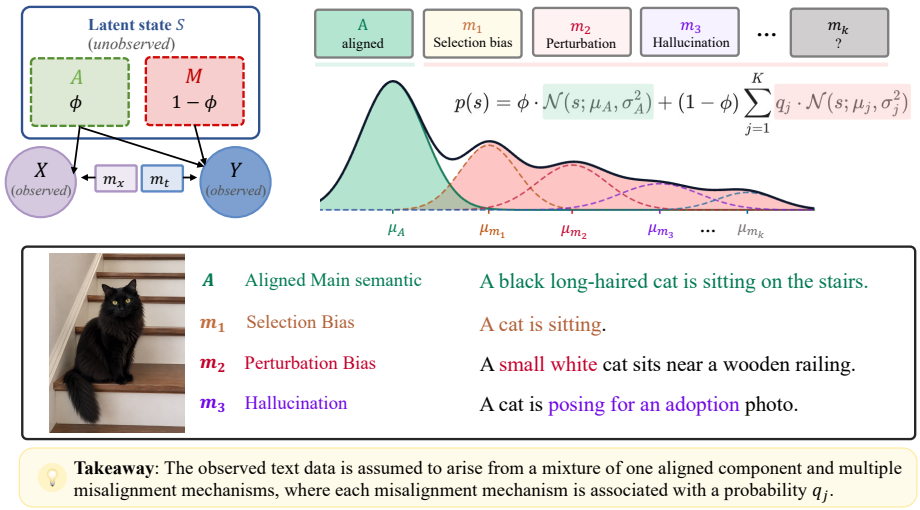

Figure 2: Conceptual visualisation of our multimodal data generation assumption. We assume that each multimodal pair contains both semantically aligned and misaligned information.

3 Contamination in the Multimodal Attribution Objective

The Information Bottleneck (IB). The IB seeks to learn a representation that preserves the information in the input that is essential for a downstream task, while discarding irrelevant or redundant content [26]. Given an input variable X, a target variable Y, and a representation Z, the IB objective is formulated as

$$\max_{Z} I(Z;Y) - \beta I(Z;X) \tag{1}$$

where $I(\cdot;\cdot)$ denotes mutual information and β is a Lagrange multiplier that governs the trade-off between task-relevant sufficiency and information compression. To analyse the objective contamination arising in multimodal attribution, we first provide the following definition.

Assumption 3.1 (Multimodal Mixture Data Generative Model)**.** Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ denote two modalities, and let (x, y) be a paired sample. We assume that the multimodal pairs are generated from a mixture of K latent components determined by the latent semantic space, indexed by an indicator $S \in \{1, \ldots, K\}$, with mixture weights $w_k = P(S = k)$ satisfying $\sum_{j=1}^{K} q_j = 1$. The joint distribution is given by:

$$p(s) = \phi p_A(s) + (1-\phi) \sum_{j=1}^{K} q_j p_{M_j}(s), \qquad \text{with } \sum_{j=1}^{K} q_j = 1 \tag{2}$$

where $p_A(s)$ denotes the aligned semantic component induced solely by the shared state A, with mixture weight $\phi \in (0, 1)$. The distributions $\{p_{M_j}(s)\}_{j=1}^{K}$ denote heterogeneous misaligned semantic components associated with different subsets or modes of the misalignment factors M, and $\{q_j\}_{j=1}^{K}$ are their normalised mixture weights satisfying $\sum_{j=1}^{K} q_j = 1$. Consequently, the total proportion of cross-modal misalignment is $1 - \phi$. A conceptual visualisation is shown in Fig. 2.

Theorem 3.2 (Decomposition of the Contamination IB Fitting Objective)**.** Under Assumption 3.1, the fitting term in the standard Information Bottleneck objective in Eq. 1 admits the following exact decomposition, which explicitly characterises the contributions of aligned and misaligned samples to the information captured by the representation:

$$\begin{aligned} I_p(Z;Y) &= I_p(Z;Y \mid S) + I_p(Z;Y;S) \\ &= \underbrace{\phi I_{p_A}(Z;Y)}_{\text{Aligned Semantic Fitting}} + \underbrace{(1-\phi)\bar{I}_{p_M}(Z;Y)}_{\text{Direct Contamination}} + \underbrace{I_p(Z;Y;S)}_{\text{Latent-State Dependence}} . \end{aligned} \tag{3}$$

Here, $\phi I_{p_A}(Z;Y)$ captures the useful semantic fitting pressure induced by aligned multimodal pairs, $(1-\phi)\bar{I}_{p_M}(Z;Y)$ represents the direct fitting contribution arising from misaligned samples, and $I_p(Z;Y;S)$ characterises the Latent-State dependence introduced by the latent mixture state S. Therefore, although the standard IB objective treats all samples uniformly, its fitting term is in fact jointly shaped by aligned supervision, contamination from misaligned pairs, and the dependence structure induced by the hidden alignment state. See appendix E.1 for full derivation.

Proposition 3.3 (Contaminated Fitting Pressure in the Global IB Objective)**.** Maximising the global objective $\max_Z \; I_p(Z;Y) - \beta I_p(Z;X)$ does not optimise the aligned semantic term in isolation. Instead, up to the compression penalty $-\beta I_p(Z;X)$, the objective assigns fitting pressure to all components appearing in this decomposition: the aligned semantic fitting term $\phi I_{p_A}(Z;Y)$, the direct contamination term $(1-\phi)\bar{I}_{p_M}(Z;Y)$, and the latent-state dependence term $I_p(Z;Y;S)$.

In particular, as conceptually illustrated in Fig. 3, when $\phi > 0$, the objective rewards representations that preserve dependencies under the aligned component p_A. When $\phi < 1$ and $\bar{I}_{p_M}(Z;Y) > 0$, it also rewards representations that preserve statistical dependencies within the misaligned component p_M, even when these dependencies arise from sampling bias, annotation noise, or accidental correlations rather than genuine cross-modal semantics. Moreover, when $I_p(Z;Y;S) \neq 0$, the objective is additionally affected by mixture-induced coupling between the representation Z, the target modality Y, and the unobserved assignment variable S.

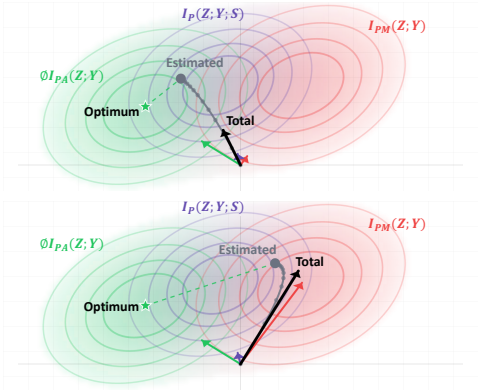

Figure 3: Illustration of the IB fitting decomposition under misalignment. Top: mild contamination. Bottom: severe contamination. As contamination increases, the estimated optimisation direction deviates further from the semantic optimum.

Remark 3.4 (Implicit aligned-regime assumption in existing multimodal attribution methods)**.** When $\phi = 1$, the data distribution reduces to the fully aligned regime. In this case, the fitting objective reduces to the aligned semantic term alone, without direct contamination or latent-state dependence. **Most existing multimodal attribution methods are implicitly developed under this ideal setting**, as they assume that paired modalities provide reliable proxy supervision without explicitly accounting for cross-modal misalignment. We further provide a variational-inference perspective on objective contamination in Section D.3, offering additional theoretical insights into how cross-modal misalignment affects the optimisation of the Information Bottleneck objective.

4 Contamination Mitigation through Uncertainty Modeling

Having identified that the failure of the classical IB objective originates from the decomposition above, we next seek a surrogate mechanism that can play a role analogous to the unobserved latent state S. Since S is not directly accessible, such a mechanism should modulate the fitting pressure in a sample-adaptive manner, so as to avoid indiscriminate fitting under severe multimodal misalignment.

To this end, we first reformulate the objective itself rather than the architecture, replacing uniform fitting pressure with a sample-wise tempered objective in which each sample is weighted according to its estimated reliability. We realise this adaptive weighting through input-dependent predictive variance. We describe its parameterisation, including the variance-prediction network, in Sec. 4.2.

4.1 Practical realisation via heteroscedastic variational inference

To realise the sample-adaptive objective introduced above, we adopt a heteroscedastic Gaussian predictive model

$$r(y \mid z, x) = \mathcal{N}\big(y; \mu(z), \sigma^2(x) I_d\big), \tag{4}$$

together with an IB posterior $q(z \mid x)$ and prior $p(z)$. Here, $r(y \mid z, x)$ denotes the variational predictive distribution, $\mu(z)$ is the prediction mean produced from the bottleneck representation z,

and $\sigma^2(x_i)$ is an input-dependent predictive variance that quantifies the estimated uncertainty of sample x_i, see Sec. 4.2 for more details. The matrix I_d denotes the d-dimensional identity matrix, where d is the dimension of the target y. Under this formulation, samples with higher predicted variance are considered less reliable and exert weaker fitting pressure during optimisation.

Using a Monte Carlo estimator with K samples $z_{ik} \sim q(z \mid x_i)$, and combining the likelihood term with the KL regularizer, the resulting empirical objective to minimise is

$$\mathcal{J} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{d}{2}\log\sigma_i^2(x_i) - \frac{1}{\sigma_i^2(x_i)}\cdot\frac{1}{K}\sum_{k=1}^{K}\ell\big(y_i,\mu(z_{ik})\big) + \beta\,\mathrm{KL}(q(z\mid x_i)\,\|\,p(z))\right]. \quad (5)$$

The detailed derivation of Eq. (5) from the heteroscedastic Gaussian likelihood in Eq. (4) is provided in Appendix E.4. Equivalently, this objective can be interpreted as applying a sample-wise temperature to the log-likelihood term, where the effective temperature is controlled by $1/\sigma^2(x_i)$, with an additional regularisation term that prevents the predicted variance from being inflated without bound. Under this view, the likelihood contribution of each sample is adaptively tempered according to its predicted uncertainty, allowing less reliable samples to exert weaker influence on the optimisation process. To emphasise this uncertainty-aware adjustment of fitting strength, we refer to the resulting formulation as a **Tempered Information Bottleneck (TIB)**.

Stabilised weighting. However, heteroscedastic likelihoods scale gradients by the inverse predicted variance. Following the variance-weighted objective of α-NLL [27], we introduce an exponent $\alpha \in [0,1]$ and denote the stop-gradient operator with the brackets $\lfloor\cdot\rfloor$. For each sample, we draw $z_{ik} \sim q(z \mid x_i)$ and weight the objective by $\lfloor\sigma_i^{2\alpha}\rfloor$, so that the predicted variance acts as a multiplicative factor on the gradient. Therefore, our final objective function is formulated as follows:

$$\mathcal{J}_{\mathrm{emp}} = \frac{1}{N}\sum_{i=1}^{N}\left\{\lfloor\sigma_i^{2\alpha}(x_i)\rfloor\Big[\frac{d}{2}\log\sigma_i^2(x_i) - \frac{1}{\sigma_i^2}\cdot\frac{1}{K}\sum_{k=1}^{K}\ell\big(y_i,\mu(z_{ik})\big)\Big] + \beta\,\mathrm{KL}\big(q(z|x_i)\,\|\,p(z)\big)\right\}. \quad (6)$$

4.2 Implementation

Input-dependent predictive variance Let $f^l(x_i) \in \mathbb{R}^J$ denote the intermediate feature representation at layer l. Following the heteroscedastic construction in [28], we introduce a variance-prediction branch to model input-dependent uncertainty. Consistent with the objective in Sec. 4.1, we use a scalar predictive variance for each sample, rather than a feature-wise variance vector. We therefore define a variance prediction network g_{var} that outputs a positive variance $\sigma^2(x_i) \in \mathbb{R}_+$:

$$\sigma^2(x_i) = \exp\big(w_{\mathrm{var}}^\top \mathbf{h}^{\mathrm{var}}(x_i) + b_{\mathrm{var}}\big)\,. \quad (7)$$

Here, $\mathbf{h}^{\mathrm{var}}(x_i)$ denotes the hidden activation of the variance-prediction branch, and w_{var} and b_{var} are learnable parameters. The exponential activation ensures that the predicted variance remains strictly positive. This variance quantifies the overall data-dependent uncertainty associated with the input sample x_i and serves as the implementation-level counterpart of the predictive variance.

Multimodal Attribution Mask Modeling To learn the attribution map from modality X to modality Y, we follow the multimodal information bottleneck framework [5] and introduce a parameterised mask mapping $h(x_i;\lambda) \in [0,1]^J$ to control the information retained in the latent representation. Specifically, the mask acts on the intermediate representation $f^l(x_i)$, and the resulting stochastic latent variable is constructed by injecting the predictive variance introduced above into the reparameterisation of z:

$$z_i \mid x_i;\theta = h(x_i;\lambda)\odot f^l(x_i) + \sigma(x_i)\big(1_J - h(x_i;\lambda)\big)\odot\epsilon, \quad (8)$$

where $\epsilon \sim \mathcal{N}(0, I_J)$ is isotropic Gaussian noise, $\odot$ denotes the Hadamard product, and $\sigma(x_i)$ is the sample-dependent predictive standard deviation induced by the variance prediction network. In this design, a mask value of 1 preserves the original feature, whereas a mask value of 0 replaces that dimension with uncertainty-aware noise. For each input x_i, the model learns which features to mask (via h) and how much noise to inject (via $\sigma^2(x_i)$). The variance $\sigma^2(x_i)$ quantifies uncertainty in cross-modal reliability for sample x_i, reflecting how reliably modality Y can explain modality X.

5 Experiments

The implementation and model design details are provided in appendix C.1. In our experiments, we investigate the following questions and draw the corresponding conclusions:

Sec. 5.2: Do real-world multimodal datasets conform to our assumption? *Largely yes.* The similarity-distribution evidence and appendix analyses support our heterogeneous-mixture assumption; at the same time, we do not claim that it fully captures the complexity of real-world data, which may involve even richer heterogeneity.

Sec. 5.3: Why does the classical homoscedastic IB objective fail under multimodal misalignment? *Because* uniform fitting pressure forces the model to fit both aligned and misaligned pairs indiscriminately, leading to unstable and over-confident attribution behaviour. **Why can TIB avoid this failure?** TIB may mitigate it by providing a more informative sample-wise reliability signal than the imperfect reliability ordering induced by homoscedastic weighting, allowing it to better identify lower-risk samples and achieve improved selective-risk behaviour.

Sec. 5.4 and Sec. 5.5: Real-world attribution and ablation studies. *TIB achieves better attribution performance on real-world multimodal datasets.* We further investigate how predictive variance distinguishes challenging but aligned samples from genuinely misaligned ones, and examine how different variance parameterisations and modelling choices affect attribution robustness and reliability.

5.1 Datasets

Synthetic Misalignment Data To construct a controlled benchmark for multimodal misalignment, we generate synthetic image–text pairs under Assumption 3.1 from Flickr8k [29] and RefCOCO [30]. For each image, we pair one aligned caption with two misaligned captions, corresponding to *selective bias* and *perturbation bias*. Full construction details are provided in Appendix F.

Large-scale Multimodal Datasets In addition to the carefully controlled synthetic misalignment benchmark, we further evaluate our method on large-scale, publicly available image–caption datasets to demonstrate its effectiveness on real-world multimodal data. Specifically, we use Conceptual Captions 3M (CC3M) [31], Flickr8k [29] and LAION-400M [32], all of which consist of web-scale image–text pairs collected with minimal manual filtering. Due to the large scale and inherent imbalance of these datasets, we aim to study attribution behaviour under realistic data distributions rather than exhaustively attributing on the full scale. We use a fixed random seed to uniformly sample 5,000 image–caption pairs from each dataset for evaluation.

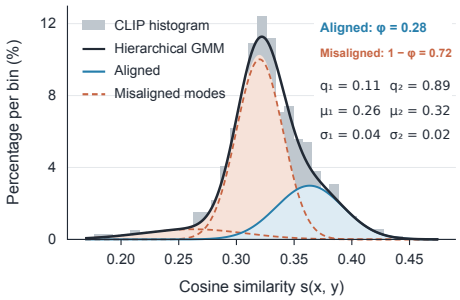

Figure 4: GMM fit on a sampled subset of LAION-400M.

5.2 Empirical validation of the data assumption

We examine whether real-world multimodal datasets exhibit the heterogeneous mixture structure described in Assumption 3.1, rather than a single fully aligned regime. Specifically, we assess whether the observed image–text pairs are better characterised by a dominant aligned component and heterogeneous misaligned components, as illustrated in Fig. 4. As a representative example, we analyse a sampled subset of LAION-400M [32], a large-scale and noisy image–caption dataset. We use image–text cosine similarity as a proxy for the latent alignment state S, where higher similarity indicates a pair is more likely to be aligned. We then fit a Gaussian mixture model to the empirical similarity distribution and use BIC to select the number of components. The component with the highest mean is treated as the aligned mode, and the remaining components as misaligned modes. The fitted mixture structure supports our assumption that real-world multimodal data are better characterised by heterogeneous aligned and misaligned components than by a fully aligned regime.

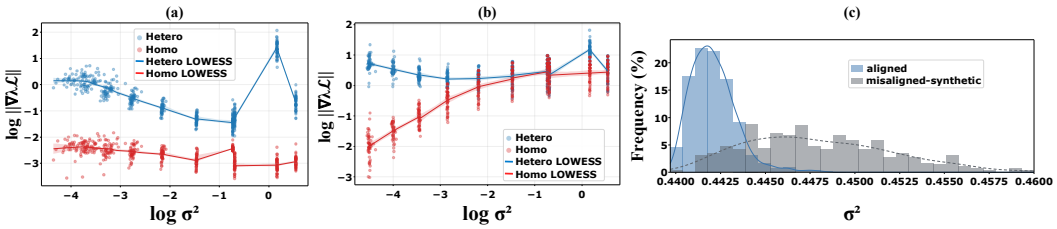

Figure 5: **Mechanistic analysis of heteroscedastic gating ($\sigma^2(x_i)$).** (a,b) Relation between predicted variance σ^2 and attribution-mask gradient magnitude $\|\nabla_\lambda \mathcal{L}\|$ under $V \to T$ and $T \to V$, comparing heteroscedastic and homoscedastic bottlenecks. Heteroscedastic IB (TIB) shows variance-dependent gradient flow, whereas homoscedastic IB (classical IB) does not. (c) Predicted variance is consistently larger for misaligned than aligned pairs, indicating higher uncertainty.

We extend the same analysis to CC3M [31] and Flickr8k [29], with detailed results provided in Appendix D.1. The findings reveal similar heterogeneous mixture structures across these datasets, providing further empirical support for our assumption. This consistency further supports the heterogeneous-mixture assumption across multimodal datasets of varying scales and characteristics.

5.3 Does the Classical IB Objective Fail under Multimodal Misalignment?

We examine the mechanism of heteroscedastic gating in Fig. 5. Panels (a,b) show the relationship between predicted variance $\sigma^2(x_i)$ and attribution-mask gradient magnitude $\|\nabla_\lambda \mathcal{L}\|$ under $V \to T$ and $T \to V$, respectively. The heteroscedastic and homoscedastic bottlenecks exhibit distinct gradient profiles, highlighting how sample-dependent predictive variance influences gradient magnitudes and modulates the optimisation of attribution masks. Panel (c) further shows that the variance distribution of synthetically misaligned pairs shifts towards higher values than that of aligned pairs, indicating greater predictive uncertainty under cross-modal misalignment. Together, these findings highlight how learned variance captures cross-modal unreliability and results in uncertainty-aware attribution.

This failure can be traced to incorrect reliability ordering induced by homoscedastic weighting. As shown in Fig. 6, we evaluate different reliability-ranking strategies using a risk–coverage curve. Samples are first ordered from most to least reliable under each strategy. We then compute the average attribution risk over the retained prefix of this ordering as the coverage increases from 0 to 1. Lower curves indicate better reliability ordering because they retain lower-risk samples at smaller coverage. Thus, at coverage 0.2, the plotted value corresponds to the mean attribution risk of the top 20% of selected samples.

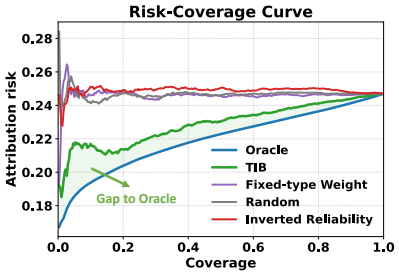

Figure 6: Risk–coverage curves for uncertainty-based reliability ranking. Lower is better.

The compared ranking strategies have the following meanings. *Oracle* uses the ground-truth risk itself as the ranking score and therefore represents the ideal upper bound on selection quality. *TIB* ranks samples by the learned uncertainty score and corresponds to our method. *Fixed-type Weight* ranks samples using coarse-grained group averages defined by caption type and perturbation slot, serving as a structured baseline. *Random* applies a random ordering and therefore represents a baseline with no reliable selection ability. *Inverted Reliability* reverses the reliability ranking induced by TIB and serves as a counterexample or lower-bound reference. We observe that TIB consistently approaches the oracle ordering and clearly outperforms the coarse-grained and random baselines, whereas inverted reliability leads to substantially higher selective risk. This confirms that attribution failure under misalignment is primarily driven by incorrect reliability ordering: once unreliable samples are down-weighted or deferred, attribution quality is substantially improved.

Table 1: **Ground-truth attribution evaluation on the synthetic RefCOCO benchmark.** We apply the same synthetic caption protocol to RefCOCO and report IoU (↑) as mean ± standard deviation across five random seeds. Coloured cells indicate caption types using the corresponding curve colours.

Caption type	Grad-CAM [2]	RISE [4]	M2IB [5]	Fast-IG [3]	Chefer [33]	Grad-ECLIP [34]	FIxLIP [24]	NIB [6]	Ours (TIB)
Aligned	8.89 ± 3.34	10.65 ± 3.12	14.20 ± 2.18	11.92 ± 2.86	12.83 ± 2.68	13.48 ± 2.43	13.82 ± 2.27	11.96 ± 2.05	$\mathbf{16.26 \pm 1.02}$
Selective-bias	7.42 ± 3.28	8.91 ± 3.05	11.76 ± 2.12	9.84 ± 2.79	10.53 ± 2.61	11.18 ± 2.38	11.47 ± 2.21	9.72 ± 2.01	$\mathbf{14.21 \pm 1.01}$
Perturbation-bias	5.36 ± 3.19	6.48 ± 2.96	8.73 ± 2.08	7.12 ± 2.71	7.81 ± 2.55	8.26 ± 2.31	8.55 ± 2.17	7.08 ± 1.98	$\mathbf{11.83 \pm 1.00}$

5.4 How does TIB perform on attribution over real-world multimodal datasets?

Table 1 shows that TIB achieves the highest IoU across aligned, selective-bias, and perturbation-bias captions on the synthetic RefCOCO benchmark, supporting its attribution robustness under caption bias.

As shown in Table 2, we report attribution performance on real-world multimodal datasets under two evaluation settings: *full evaluation* on all samples and *filtered evaluation* on samples selected by the learned predictive variance $\sigma^2(x_i)$. The filtered setting retains samples with lower predicted variance, *e.g.* those considered more reliable, to assess attribution quality under a reliability-aware protocol. We evaluate all methods using three metric families: Drop/Incr. [5], grounding metrics, and ROAR-style [5] similarity metrics (*i2c*; see Appendix C.2 for details). In Table 2, $\mathrm{mean}_{\mathrm{acc}}$ denotes the mean metric over accepted samples. We rank samples by increasing predictive variance $\sigma^2(x_i)$ and retain the lowest-variance 80% at acceptance rate $\kappa = 0.8$, excluding the remaining 20% with higher estimated uncertainty. Each entry reports the full-sample mean as the main value and the accepted-sample mean as the subscript. This evaluates attribution quality on samples considered more reliable by the learned uncertainty.

Table 2: **Quantitative comparison of interpretability methods across vision–language benchmarks.** Results are reported as $\mathrm{mean}_{\mathrm{acc}}$ with acceptance rate $\kappa = 0.8$. Main values report full-sample means; subscripts report accepted-sample means at acceptance rate.

Dataset	Metric	Grad-CAM [2]	M2IB [5]	NIB [6]	Ours (TIB)
Drop / Incr.					
cc3M-I	Drop↓	$8.07_{6.02}$	$1.11_{1.04}$	$1.03_{0.95}$	$\mathbf{1.00}_{0.91}$
cc3M-I	Incr↑	$7.20_{10.21}$	$37.20_{39.43}$	$38.80_{41.90}$	$\mathbf{41.21}_{44.77}$
cc3M-T	Drop↓	$1.79_{1.24}$	$\mathbf{0.90}_{0.85}$	$1.13_{1.04}$	$1.10_{0.96}$
cc3M-T	Incr↑	$33.90_{35.12}$	$37.80_{40.07}$	$40.50_{43.74}$	$\mathbf{41.11}_{44.11}$
Flickr8k-I	Drop↓	$9.19_{7.23}$	$1.59_{1.49}$	$1.49_{1.37}$	$\mathbf{1.19}_{1.13}$
Flickr8k-I	Incr↑	$5.10_{10.22}$	$27.50_{29.15}$	$28.30_{30.56}$	$\mathbf{30.50}_{32.78}$
Grounding					
RefCOCO	vDrop↓	$1.98_{1.55}$	$0.84_{0.79}$	$\mathbf{0.72}_{0.66}$	$0.74_{0.65}$
RefCOCO	vIncr↑	$46.38_{49.22}$	$45.52_{48.25}$	$48.16_{52.01}$	$\mathbf{51.20}_{55.22}$
RefCOCO	tDrop↓	$1.97_{1.33}$	$1.02_{0.96}$	$1.21_{1.11}$	$\mathbf{1.00}_{0.90}$
RefCOCO	tIncr↑	$37.20_{40.65}$	$41.16_{43.63}$	$41.88_{45.23}$	$\mathbf{45.17}_{48.53}$
RefCOCO	mIoU↑	$8.89_{10.32}$	$14.20_{15.05}$	$11.96_{12.92}$	$\mathbf{16.26}_{18.11}$
ROAR					
cc3M-I	i2t-oc↑	$57.01_{60.12}$	$48.96_{51.90}$	$58.98_{63.70}$	$\mathbf{60.76}_{66.95}$
cc3M-I	i2t-co↑	$57.33_{62.42}$	$68.20_{72.29}$	$57.98_{62.62}$	$\mathbf{69.99}_{75.87}$
Flickr8k-T	t2i-oc↑	$65.21_{69.23}$	$68.75_{72.88}$	$64.34_{69.49}$	$\mathbf{71.85}_{79.04}$

5.5 Ablation study

Predictive variance captures cross-modal reliability through conditional fitting along $x \to z \to y$, with $z \sim q(z \mid x)$. Although $\sigma^2(x)$ is parameterised from x, its uncertainty head receives supervision from y through the fitting objective. High predictive variance may therefore indicate either challenging but valid cross-modal correspondence or persistent semantic incompatibility caused by misalignment. Fig. 7 investigates this distinction by examining how predictive variance responds to matching difficulty and how different variance parameterisations affect the identification of misaligned samples.

To isolate matching difficulty from misalignment, we construct a hard-but-aligned benchmark using CC3M images. An LLM rewrites each original caption twice with progressively increasing semantic ambiguity while preserving image–text alignment, yielding three caption levels. For example, "A busy bus station" is rewritten as "A public transport hub crowded with commuters" and then "A transit hub where numerous travellers are waiting, boarding, or moving between buses." The final caption is used as the hard-but-aligned sample. In Fig. 7(a), the variance of hard-but-aligned samples decreases substantially with more optimisation steps, whereas perturbed samples retain high variance over the tested budgets. The panel (b) shows that larger models also reduce variance for hard-but-aligned samples, supporting the distinction between reducible matching difficulty and persistent misalignment. And the panels (c,d) show that $\sigma^2(x)$ separates aligned, selective, and perturbed samples more clearly than $\sigma^2(x, y)$, supporting its use for identifying cross-modal unreliability.

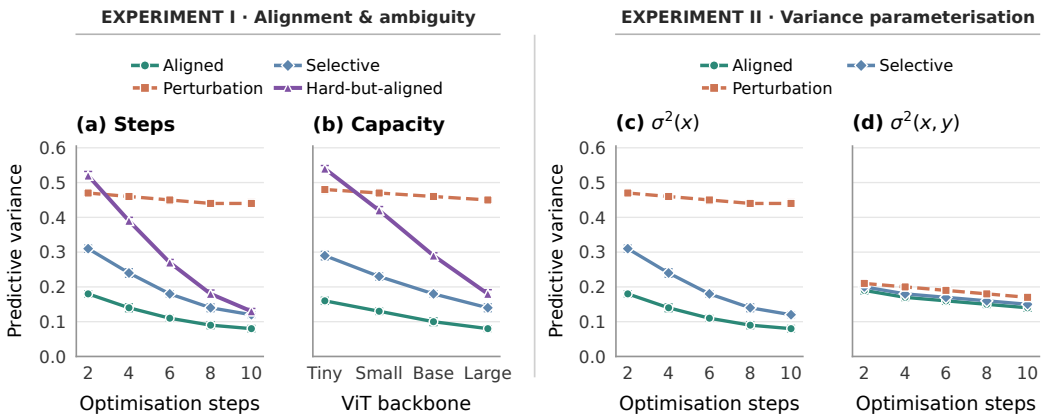

Figure 7: **Predictive variance analysis. (a–b)** Reducible uncertainty in hard-but-aligned pairs versus persistent uncertainty in misaligned pairs. **(c–d)** Discriminability of misalignment under different variance parameterisations.

Table 3: **Ablation of variance modelling in TIB.** Each entry reports $\Delta\text{Incr.}/\Delta\text{Drop}$ relative to the default TIB setting with learned σ^2. Negative $\Delta\text{Incr.}$ and positive ΔDrop indicate performance degradation.

Setting	Aligned	Selective	Perturbation
Random σ^2	$-3.62/+0.15$	$-4.44/+0.44$	$-4.74/+0.54$
Fixed σ^2	$-4.19/+0.12$	$-3.79/+0.38$	$-4.05/+0.47$
Inverted learned σ^2	$-3.41/+0.29$	$-8.17/+0.81$	$-8.40/+0.94$
w/o stop-gradient	$-0.60/+0.06$	$-1.44/+0.12$	$-1.50/+0.16$
w/o variance regularisation	$-1.86/+0.28$	$-4.96/+0.37$	$-3.99/+0.62$

Table 3 reports performance changes relative to the default TIB with learned σ^2. Random or fixed variance consistently degrades performance, while inverting the learned variance is especially harmful for selective and perturbed samples. Removing stop-gradient or variance regularisation also reduces performance, supporting the proposed sample-wise uncertainty modelling.

Additional experiments are provided in the appendix to further examine TIB's sensitivity and attribution behaviour. We analyse the effects of β on attribution performance (Fig. 10) and investigate a β-free variant that reduces reliance on explicit trade-off parameter selection (Appendix D.2). We also provide attribution map visualisations (Figs. 11, 12, and 13) to complement the quantitative results and illustrate attribution behaviour under different image–text alignment conditions.

6 Conclusion

In this paper, we investigated multimodal attribution from an information bottleneck (IB) perspective under cross-modal misalignment, revealing how violations of conventional alignment assumptions introduce spurious cross-modal dependencies into standard IB objectives. Such contamination can distort the optimisation process, leading to over-confident and unstable attribution explanations that fail to reflect genuine image–text correspondence. To address this limitation, we propose a sample-wise heteroscedastic information bottleneck that explicitly models predictive uncertainty and adaptively adjusts fitting strength according to cross-modal reliability. By reducing the influence of unreliable image–text pairs, TIB mitigates the effects of misalignment on attribution optimisation. Our theoretical analysis and extensive experiments on synthetic benchmarks and large-scale real-world datasets showed that the proposed TIB effectively suppresses fitting to misaligned image–text pairs. These findings highlighted the importance of incorporating data-dependent uncertainty into information bottleneck objectives for reliable multimodal interpretability.

Acknowledgements

This project is partially supported by an Amazon Research Award, the Doctoral Studentship from the University of Birmingham College of Engineering and Physical Sciences and the China Scholarship Council (Grant No. 202408060321). The computations in this research were partially performed using the Baskerville Tier 2 HPC service. Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP\T022221\1) and the Digital Research Infrastructure programme (EP\W032244\1) and is operated by Advanced Research Computing at the University of Birmingham.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[3] R. Hesse, S. Schaub-Meyer, and S. Roth, "Fast axiomatic attribution for neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19513–19524, 2021.

[4] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *Proceedings of the British Machine Vision Conference*, 2018.

[5] Y. Wang, T. G. Rudner, and A. G. Wilson, "Visual explanations of image-text representations via multi-modal information bottleneck attribution," *Advances in Neural Information Processing Systems*, vol. 36, pp. 16009–16027, 2023.

[6] Z. Zhu, Z. Jin, J. Zhang, N. Yang, J. Huang, J. Zhou, and F. Chen, "Narrowing information bottleneck theory for multimodal image-text representations interpretability," in *Proceedings of the International Conference on Learning Representations*, 2025.

[7] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[8] C. Chen, J. Qi, X. Liu, K. Bin, R. Fu, X. Hu, and P. Zhong, "Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26836–26845, 2024.

[9] Y. Cai, Y. Liu, E. Gao, T. Jiang, Z. Zhang, A. van den Hengel, and J. Q. Shi, "On the value of cross-modal misalignment in multimodal representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[10] Y. Liu, K. Zhang, X. Ren, Y. Huang, J. Jin, Y. Qin, R. Su, R. Xu, Y. Yu, and W. Zhang, "Alignrec: Aligning and training in multimodal recommendations," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1503–1512, 2024.

[11] T. Tao, G. Wang, Y. Lao, P. Chen, J. Liu, L. Lin, K. Yu, and X. Liang, "Alignmif: Geometry-aligned multimodal implicit field for lidar-camera joint synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21230–21240, June 2024.

[12] C.-E. Wu, Y. Tian, H. Yu, H. Wang, P. Morgado, Y. H. Hu, and L. Yang, "Why is prompt tuning for vision-language models robust to noisy labels?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15488–15497, 2023.

[13] R. Nakada, H. I. Gulluk, Z. Deng, W. Ji, J. Zou, and L. Zhang, "Understanding multimodal contrastive learning and incorporating unpaired data," in *International Conference on Artificial Intelligence and Statistics*, pp. 4348–4380, PMLR, 2023.

[14] S. Gupta, S. Sundaram, C. Wang, S. Jegelka, and P. Isola, "Better together: Leveraging unpaired multimodal data for stronger unimodal models," in *International Conference on Learning Representations*, vol. 2026, pp. 73970–74025, 2026.

[15] M. T. Ribeiro, S. Singh, and C. 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.

[16] D. Pan, X. Li, and D. Zhu, "Explaining deep neural network models with adversarial gradient integration," in *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[17] Z. Zhu, H. Chen, J. Zhang, X. Wang, Z. Jin, M. Xue, D. Zhu, and K.-K. R. Choo, "Mfaba: A more faithful and accelerated boundary-based attribution method for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 17228–17236, 2024.

[18] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, pp. 3319–3328, PMLR, 2017.

[19] C. Lin, H. Chen, C. Kim, and S.-I. Lee, "Contrastive corpus attribution for explaining representations," in *The Eleventh International Conference on Learning Representations*, 2023.

[20] Y. Gandelsman, A. Efros, and J. Steinhardt, "Interpreting clip's image representation via text-based decomposition," in *International conference on learning representations*, vol. 2024, pp. 18395–18416, 2024.

[21] Y. Lei, Z. Li, Y. Li, J. Zhang, and H. Shan, "Lico: Explainable models with language-image consistency," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 61870–61887, Curran Associates, Inc., 2023.

[22] N. Kalibhat, S. Bhardwaj, C. B. Bruss, H. Firooz, M. Sanjabi, and S. Feizi, "Identifying interpretable subspaces in image representations," in *International Conference on Machine Learning*, pp. 15623–15638, PMLR, 2023.

[23] Z. Ma and S. Yu, "Explainable multimodal regression via information decomposition," *arXiv preprint arXiv:2512.22102*, 2025.

[24] H. Baniecki, M. Muschalik, F. Fumagalli, B. Hammer, E. Hüllermeier, and P. Biecek, "Explaining similarity in vision-language encoders with weighted banzhaf interactions," in *Advances in Neural Information Processing Systems* (D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, eds.), vol. 38, Main Conference, pp. 54529–54566, Curran Associates, Inc., 2025.

[25] C. Agarwal, "Rethinking explainability in the era of multimodal ai," *arXiv preprint arXiv:2506.13060*, 2025.

[26] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[27] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," in *International Conference on Learning Representations*, 2022.

[28] D. Nix and A. Weigend, "Estimating the mean and variance of the target probability distribution," in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, vol. 1, pp. 55–60 vol.1, 1994.

[29] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[30] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European conference on computer vision*, pp. 69–85, Springer, 2016.

[31] P. Sharma, N. Ding, S. Goodman, and R. 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.

[32] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[33] H. Chefer, S. Gur, and L. 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.

[34] C. Zhao, K. Wang, J. H. Hsiao, and A. B. Chan, " Grad-ECLIP: Gradient-Based Visual and Textual Explanations for CLIP ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 48, pp. 14002–14019, Nov. 2026.

[35] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *International Conference on Learning Representations*, 2020.

Appendix Contents

A Limitation **14**

B Broader impacts **14**

C Experimental and model details **14**

C.1 Model implementation details 14

C.2 Evaluation metrics for real-world multimodal attribution 15

D Additional Experiments **15**

D.1 Do Real-World Multimodal Data Support Our Assumption? 15

D.2 Analysis of the roles of β and α 16

D.3 Understanding objective contamination from a variational inference perspective . . 17

E Proofs **18**

E.1 Decomposition of the Contamination IB Fitting Objective 18

E.2 Variational Relaxation of the Contaminated IB Objective 19

E.3 Barber–Agakov bound 21

E.4 Derivation of the Empirical Heteroscedastic VIB Objective 23

F Synthetic misalignment data generation **24**

G Visualisation **24**

A Limitation

Our analysis relies on a simplified mixture-based misalignment model that distinguishes only between aligned and misaligned components. While this abstraction is sufficient to reveal the contamination mechanism in IB-style attribution, real-world multimodal misalignment can be more complex (*e.g.* partial, hierarchical, or temporal misalignment), which is not explicitly modelled here. Nevertheless, we believe this simplified formulation captures the essential source of misalignment and is sufficient to highlight the non-negligible impact of misaligned data on multimodal attribution.

B Broader impacts

Many existing multimodal attribution methods implicitly assume that paired modalities in real-world datasets are well aligned and can serve as a reliable proxy supervision for explanation. In practice, however, large-scale multimodal data are often collected with minimal filtering and contain substantial semantic misalignment. Under such conditions, attribution methods may produce over-confident or forced explanations driven by spurious cross-modal correlations, reducing the reliability and stability of interpretability results and limiting their practical usefulness.

This work highlights the potential risks of applying attribution methods under unrealistic alignment assumptions and emphasises the importance of accounting for data quality and uncertainty when interpreting multimodal models. The proposed analysis operates as a post-hoc interpretability framework and does not introduce new predictive capabilities or application domains. We do not anticipate specific societal risks arising directly from this work beyond those already associated with the use of large-scale multimodal models.

C Experimental and model details

C.1 Model implementation details

For all experiments, we follow the identical setup as M2IB [5] and use a pretrained CLIP model with ViT-B/32 [1] as the image encoder and a 12-layer self-attention Transformer as the text encoder. We use the pretrained weights from `openai/clip-vit-base-patch32`. For each image–caption pair, we insert an information bottleneck into a selected layer of each encoder and train the bottleneck using the same setup as the per-sample bottleneck in the original IBA [35]. Specifically, each sample is duplicated five times to stabilise training, and the bottleneck is optimised for 10 iterations using Adam with a learning rate of 0.01. We observe no significant performance differences across different learning rates or larger numbers of training steps. We set β to 0.1 in all experiments. We use the ninth layer for attribution, following M2IB. All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU, equipped with an Intel Core i9-13900K CPU (24 cores, 32 threads) and approximately 64 GB of system memory. The software environment was configured with Python 3.11.14, PyTorch 2.9.0 (CUDA 12.8), and Transformers 5.5.4. We used this hardware and software configuration consistently across experiments to ensure a uniform evaluation environment.

Implementation of the predictive-variance bottleneck. In implementation, the input modality is first passed through its corresponding encoder, and the intermediate representation at layer l is used as the bottleneck input, *i.e.* $f^l(x_i)$. Based on this representation, we jointly learn an attribution mask branch and a predictive-variance branch. The mask branch outputs the feature-wise attribution mask $h(x_i;\lambda)$ with the same shape as $f^l(x_i)$, while the variance branch predicts a sample-wise scalar variance $\sigma^2(x_i)$ and its corresponding standard deviation $\sigma(x_i)$. Since $\sigma(x_i)$ is defined per sample, it is broadcast to match the dimensionality of the intermediate feature tensor before noise injection. The stochastic bottleneck variable is then constructed as

$$z_i = h(x_i;\lambda) \odot f^l(x_i) + \sigma(x_i)\big(1 - h(x_i;\lambda)\big) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

which preserves features with high mask values while replacing suppressed features with uncertainty-scaled Gaussian noise, allowing the model to adaptively control feature retention and perturbation according to the estimated sample-wise uncertainty.

C.2 Evaluation metrics for real-world multimodal attribution

We evaluate real-world multimodal attribution using three groups of metrics: ROAR, Drop/Incr., and grounding. For ROAR, we follow the evaluation protocol of M2IB with a unified percentile-based threshold across modalities. Specifically, we remove the top 25% most important features, corresponding to the 75th percentile of the attribution scores. For images, the selected pixels are replaced with the channel mean, whereas for text, the selected tokens are replaced with the blank token. The corrupted sample is then evaluated by the relative similarity drop between the original and corrupted image–text similarities, where a larger drop indicates that the removed features are more important. For ROAR retraining, we fine-tune the CLIP backbone rather than the MLP head, using an 80/20 train–validation split for 50 epochs with a learning rate of 5×10^{-4}.

For Drop and Incr., we follow the definitions introduced in Grad-CAM++ and adopt the same percentile-based thresholding strategy as above. Features are perturbed one modality at a time. Drop measures how much the model confidence decreases after removing regions considered unimportant, where lower values indicate better attribution quality. Incr. measures the proportion of samples for which restricting the input to only the explained regions increases the model confidence, where higher values are preferred.

For grounding, we compute the intersection-over-union (IoU) between the binarised attribution map and the ground-truth mask. The attribution map is first normalised to $[0, 1]$ and then binarised with a threshold of 0.5, which provides a balanced decision boundary between salient and non-salient regions. In addition, the insertion and deletion curves are computed following the official implementation of Grad-ECLIP [34].

D Additional Experiments

D.1 Do Real-World Multimodal Data Support Our Assumption?

Similarity-distribution evidence. As shown in Fig. 8, CC3M and Flickr8K display distinct similarity-distribution patterns, providing direct empirical evidence that real-world multimodal data are not well described by a single homogeneous alignment regime. The fitted distributions indicate that datasets differ in their observable mixture patterns: some are better captured by a more concentrated dominant mode, whereas others exhibit clearer multi-component structure. Although the BIC-selected GMM contains only one or two observable components, this should be interpreted cautiously because the analysis is performed on the one-dimensional cosine-similarity projection. Such a projection inevitably compresses the underlying high-dimensional semantic structure, so distinct latent misalignment factors may overlap and appear as merged observable modes. Therefore, the fitted component number should be viewed as a coarse observable proxy rather than the exact number of latent semantic sources. Importantly, the fact that different datasets favour different mixture structures already argues against a universal single-alignment regime and is consistent with our heterogeneous mixture assumption.

Feature-space structural evidence. Table 4 provides complementary evidence from the joint embedding space. *Entropy* measures local semantic diversity, with higher values indicating multiple

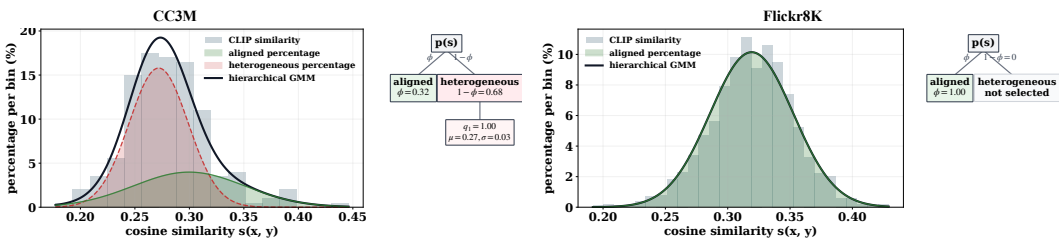

Figure 8: **Directional structure of real-world multimodal data.** Structural analysis on CC3M and Flickr8k shows that, even when paired retrieval remains strong, the feature space still exhibits substantial heterogeneity in local dispersion and cluster organisation, providing empirical support for our mixture-based multimodal misalignment assumption.

Table 4: **Structural analysis of multimodal feature-space heterogeneity across datasets.** Higher entropy and cluster variance indicate greater local dispersion, while silhouette score measures cluster separability. Paired Top-k reflects local image–text alignment through paired-caption retrieval. Despite high retrieval accuracy, substantial feature-space heterogeneity persists across datasets, supporting the proposed mixture-based misalignment assumption.

Dataset	Entropy ↑	Sim Std ↓	Text Consistency ↑	Cluster Var ↑	Silhouette ↑	Paired Top-k ↑
Flickr8k	0.758	0.017	0.637	0.191	0.219	0.967
CC3M	0.730	0.017	0.670	0.172	0.280	0.933
LAION400M	0.445	0.024	0.564	0.236	0.224	0.972

semantic modes; *Text Consistency* captures agreement among neighbouring captions; *Cluster Var* measures feature-cluster dispersion; *Silhouette* quantifies cluster separability; and *Paired Top-k* measures local image–text alignment through paired-caption retrieval. As shown in Table 4, all datasets exhibit strong paired retrieval, indicating largely preserved local alignment. However, non-trivial entropy, imperfect text consistency, and distinct cluster structures reveal persistent feature-space heterogeneity. Dataset-level differences further support this observation: CC3M exhibits stronger cluster separability, suggesting more structured semantic components, whereas LAION400M shows greater cluster variance and lower text consistency, reflecting noisier cross-modal correspondence. Together, these findings suggest that multimodal data are *locally aligned but globally heterogeneous*, supporting our mixture-based assumption.

D.2 Analysis of the roles of β and α

As illustrated in Fig. 9, β and α play distinct roles in TIB. The parameter β controls the overall bottleneck capacity, namely how much raw information is retained after compression, whereas α governs the uncertainty-aware split strength applied to the compressed representation. In this sense, β determines the global information budget, while α modulates how the retained information is adaptively redistributed under sample-wise uncertainty. To further examine the effect of α, we additionally conduct an ablation study in which different values of α are compared under a fixed β, as shown in Fig. 10.

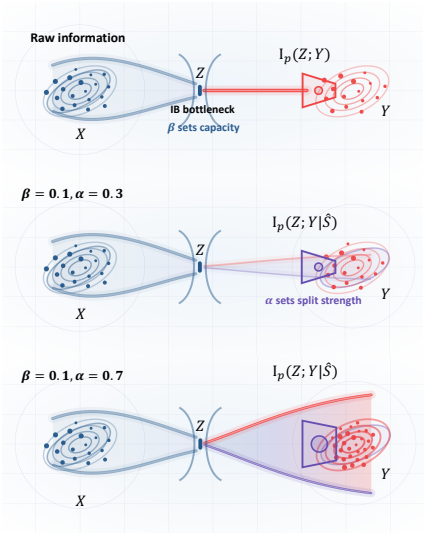

Figure 9: **Roles of α and β in TIB.** β controls the bottleneck capacity, while α controls the uncertainty-based split strength after compression.

We further analyse the sensitivity of TIB to the trade-off parameter β through a best-versus-default study. Specifically, in the *Default* setting, all methods use $\beta = 0.1$, following the standard M2IB configuration; in the *Best* setting, β is selected from a logarithmic grid $10^{-4}, 3 \times 10^{-4}, \ldots, 10^{-1}$. Across this range, TIB remains near-optimal and exhibits a visibly flatter performance curve, indicating substantially lower sensitivity to the precise choice of β. We attribute this behaviour to the adaptive predictive variance $\sigma^2(x)$, which enables sample-wise information scaling and therefore reduces reliance on a single global trade-off parameter.

We also consider whether TIB admits a β-free interpretation. A related idea was introduced by NIB, which defines a narrowing bottleneck as $\tilde{z}(\lambda) = \lambda z + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and proves that for $0 \leq \lambda_1 < \lambda_2 \leq 1$, the mutual information satisfies $I(\tilde{z}(\lambda_1); x) < I(\tilde{z}(\lambda_2); x)$. This establishes a monotonic information path from a fully closed bottleneck ($\lambda = 0$) to a fully open one ($\lambda = 1$), thereby allowing attribution to be defined through path integration without explicitly choosing β.

To extend this idea to TIB, two modifications are required: first, adopting the same path-integral formulation for attribution; and second, replacing the homoscedastic Gaussian channel with the

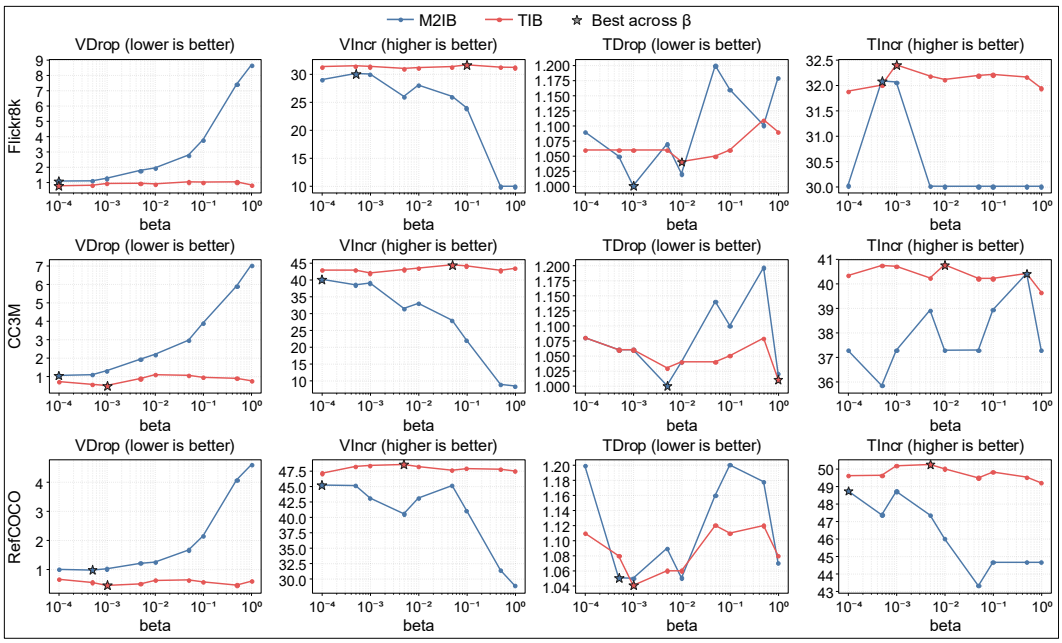

Figure 10: **β-sensitivity analysis.** TIB remains near-optimal over a wide range of β, showing lower sensitivity than competing methods.

heteroscedastic one induced by our predictive variance. This distinction is theoretically important. In NIB, the channel $\tilde{z} = \lambda z + \epsilon$ with input-independent $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is a standard Gaussian channel, under which the monotonicity and β-free narrowing results hold directly. In contrast, under the input-dependent heteroscedastic noise used in TIB, *i.e.* $\sigma^2(x)$, these guarantees no longer hold in a strict sense because the noise scale itself may carry sample-specific information and hence introduce potential information leakage. Nevertheless, if such leakage is sufficiently controlled, the standard assumptions approximately hold, making the NIB-style β-free argument a reasonable approximation.

Empirically, however, the heteroscedastic design still proves beneficial: it captures cross-modal reliability more effectively and leads to improved stability and attribution quality in practice. For example, compared with NIB, the heteroscedastic variant consistently improves both Incr. (↑) and Drop (↓) scores across all three benchmarks: on CC3M from $(39.65, 1.08)$ to $(41.02, 0.99)$, on Flickr8k from $(28.30, 1.49)$ to $(30.11, 1.21)$, and on RefCOCO from $(45.02, 0.97)$ to $(47.24, 0.94)$. These results suggest that, even though the formal β-free guarantee becomes weaker under heteroscedasticity, the resulting formulation is more effective in capturing sample-wise reliability in multimodal attribution.

D.3 Understanding objective contamination from a variational inference perspective

We now construct a variational surrogate for the contaminated IB objective by introducing an encoder $q(z \mid x)$ and a variational decoder $r(y \mid z)$. Recall from Theorem 3.2 that

$$I_p(Z;Y) = \phi I_{p_{\mathrm{A}}}(Z;Y) + (1-\phi)\bar{I}_{p_{\mathrm{M}}}(Z;Y) + I_p(Z;Y;S).$$

Applying the Barber–Agakov bound (see E.3) to the sufficiency terms $I_{p_{\mathrm{A}}}(Z;Y)$ and $I_{p_{\mathrm{M}}}(Z;Y)$, together with a prior $p(z)$ for the compression term, yields the following homoscedastic variational objective:

$$\begin{aligned}\mathcal{J}_{\mathrm{VIB}} =& \mathbb{E}_{p(x,y)}\mathbb{E}_{z\sim q(z|x)}\big[\log r(y \mid z)\big] \\ &- \beta\, \mathbb{E}_{p(x)}\mathrm{KL}(q(z \mid x) \,\|\, p(z))\,.\end{aligned} \tag{9}$$

The detailed derivation is provided in E.2. The interaction-information term $I_p(Z;Y;S)$ remains in the full theoretical formulation as a structural-coupling component. This objective corresponds to a homoscedastic variational realisation of the pairwise contaminated IB formulation, where all samples are implicitly treated with uniform confidence.

Induced interference and amplification of contamination The variational objective in Eq. (9) does not directly optimise the true mutual information. Instead, for each $s \in \{\mathrm{A}, \mathrm{M}\}$, the following identity holds:

$$\begin{aligned} I_{p_s}(Z;Y) =& \mathbb{E}_{p_s(x,y)}\mathbb{E}_{z\sim q(z|x)}\big[\log r(y \mid z)\big] \\ &+ \mathbb{E}_{p_s(z)}\mathrm{KL}\big(p_s(y \mid z) \,\|\, r(y \mid z)\big) \\ &- \mathbb{E}_{p_s(y)} \log p_s(y), \end{aligned} \tag{10}$$

where the KL term quantifies the mismatch between the true conditional distribution $p_s(y \mid z)$ and the variational decoder $r(y \mid z)$. Since this mismatch penalty is nonnegative, maximising the expected log-likelihood in Eq. (9) implicitly trades off against this approximation error. Under misalignment, the conditional distribution $p_{\mathrm{M}}(y \mid z)$ is typically more heterogeneous and less predictable, which makes the mismatch penalty larger and harder to minimise. As a result, the variational relaxation systematically amplifies the influence of misaligned samples, thereby reinforcing the contaminated fitting pressure identified in Proposition 3.3.

E Proofs

E.1 Decomposition of the Contamination IB Fitting Objective

Theorem E.1 (Exact decomposition of the fitting objective)**.** Let Z, Y, and S be random variables defined on a common probability space. Suppose that the marginal distribution of the latent semantic variable S admits the mixture representation

$$p(s) = \phi p_A(s) + (1-\phi)\sum_{j=1}^{K} q_j p_{M_j}(s), \tag{11}$$

where $\phi \in [0,1]$, $q_j \geq 0$, and $\sum_{j=1}^{K} q_j = 1$. Define the aligned semantic fitting term by

$$I_{pA}(Z;Y) \triangleq \mathbb{E}_{s\sim p_A}\big[I_p(Z;Y \mid S=s)\big], \tag{12}$$

and define the direct contamination term by

$$\bar{I}_{p_M}(Z;Y) \triangleq \sum_{j=1}^{K} q_j\, \mathbb{E}_{s\sim p_{M_j}}\big[I_p(Z;Y \mid S=s)\big]. \tag{13}$$

Then the mutual information objective admits the following exact decomposition:

$$I_p(Z;Y) = \underbrace{\phi I_{pA}(Z;Y)}_{\text{Aligned Semantic Fitting}} + \underbrace{(1-\phi)\bar{I}_{pM}(Z;Y)}_{\text{Direct Contamination}} + \underbrace{I_p(Z;Y;S)}_{\text{latent-state dependence}}. \tag{14}$$

Proof. We begin with the standard definition of the interaction information among the three random variables Z, Y, and S:

$$I_p(Z;Y;S) \triangleq I_p(Z;Y) - I_p(Z;Y \mid S). \tag{15}$$

Rearranging (15) gives

$$I_p(Z;Y) = I_p(Z;Y \mid S) + I_p(Z;Y;S). \tag{16}$$

Thus, it remains to decompose the conditional mutual information term $I_p(Z;Y \mid S)$ under the assumed semantic mixture model.

By the definition of conditional mutual information as an expectation over the conditioning variable, we have

$$I_p(Z;Y \mid S) = \mathbb{E}_{s\sim p(s)}\big[I_p(Z;Y \mid S=s)\big]. \tag{17}$$

For a discrete semantic variable S, this can be written explicitly as

$$I_p(Z;Y \mid S) = \sum_{s} p(s) I_p(Z;Y \mid S=s). \tag{18}$$

The continuous case follows identically, with sums replaced by the corresponding integrals.

Substituting the mixture representation in (11) into (18) yields

$$I_p(Z;Y \mid S) = \sum_s \left[\phi p_A(s) + (1-\phi)\sum_{j=1}^{K} q_j p_{M_j}(s)\right] I_p(Z;Y \mid S=s)$$
$$= \phi \sum_s p_A(s) I_p(Z;Y \mid S=s) + (1-\phi)\sum_{j=1}^{K} q_j \sum_s p_{M_j}(s) I_p(Z;Y \mid S=s). \tag{19}$$

The second equality follows solely from the linearity of summation and expectation.

The first term in (19) is precisely the aligned semantic fitting term defined in (12):

$$\sum_s p_A(s) I_p(Z;Y \mid S=s) = I_{pA}(Z;Y). \tag{20}$$

Similarly, the second term is precisely the weighted contribution of the heterogeneous misaligned semantic components defined in (13):

$$\sum_{j=1}^{K} q_j \sum_s p_{M_j}(s) I_p(Z;Y \mid S=s) = \bar{I}_{p_M}(Z;Y). \tag{21}$$

Combining (19), (20), and (21), we obtain the exact conditional decomposition

$$I_p(Z;Y \mid S) = \phi I_{pA}(Z;Y) + (1-\phi)\bar{I}_{p_M}(Z;Y). \tag{22}$$

Finally, substituting (22) into (16) gives

$$I_p(Z;Y) = \phi I_{pA}(Z;Y) + (1-\phi)\bar{I}_{p_M}(Z;Y) + I_p(Z;Y;S), \tag{23}$$

which is exactly the decomposition in (14). The three summands correspond respectively to aligned semantic fitting, direct contamination, and latent-state dependence. This completes the proof. □

Remark E.2. The notation $I_{pA}(Z;Y)$ in (12) should be understood as the expectation of the conditional mutual information $I_p(Z;Y \mid S=s)$ under the aligned semantic marginal $p_A(s)$. Thus, it is a distribution-indexed semantic fitting quantity rather than an unconditional mutual information computed under a full joint law $p_A(z,y)$.

E.2 Variational Relaxation of the Contaminated IB Objective

In this appendix, we derive a variational surrogate for the contaminated IB objective. From Theorem 3.2, the fitting term admits the exact decomposition

$$I_p(Z;Y) = \phi I_{p_{\mathrm{A}}}(Z;Y) + (1-\phi)\bar{I}_{p_{\mathrm{M}}}(Z;Y) + I_p(Z;Y;S), \tag{24}$$

where S denotes the latent alignment state. The term $(1-\phi)\bar{I}_{p_{\mathrm{M}}}(Z;Y)$ corresponds to direct fitting contamination induced by misaligned samples, while $I_p(Z;Y;S)$ captures higher-order latent-state dependence induced by the latent mixture assignment.

Therefore, the full contaminated IB objective is

$$\mathcal{J}_{\mathrm{IB}}^{\mathrm{full}} = \phi I_{p_{\mathrm{A}}}(Z;Y) + (1-\phi)\bar{I}_{p_{\mathrm{M}}}(Z;Y) + I_p(Z;Y;S) - \beta I_p(Z;X). \tag{25}$$

For notational simplicity, we consider a single effective misaligned component p_{M}, *i.e.*

$$\bar{I}_{p_{\mathrm{M}}}(Z;Y) = I_{p_{\mathrm{M}}}(Z;Y).$$

The extension to multiple misaligned modes follows directly by linearity.

Thus, the contaminated IB objective becomes

$$\mathcal{J}_{\mathrm{IB}}^{\mathrm{full}} = \phi I_{p_{\mathrm{A}}}(Z;Y) + (1-\phi) I_{p_{\mathrm{M}}}(Z;Y) + I_p(Z;Y;S) - \beta I_p(Z;X), \tag{26}$$

where

$$p(x,y) = \phi p_{\mathrm{A}}(x,y) + (1-\phi) p_{\mathrm{M}}(x,y)$$

is the mixture distribution and $q(z|x)$ denotes the encoder.

The interaction-information term $I_p(Z;Y;S)$ is retained explicitly in the theoretical formulation as a structural-coupling component, but is not directly variationally relaxed in the practical surrogate objective derived below.

We denote the encoder by $q(z|x)$ and the variational decoder by $r(y|z)$.

Exact decomposition and variational lower bound for $I_{p_s}(Z;Y)$ Let $s \in \{\mathrm{A},\mathrm{M}\}$ denote the aligned/misaligned component. By definition,

$$I_{p_s}(Z;Y) = \mathbb{E}_{p_s(z,y)}\left[\log \frac{p_s(y|z)}{p_s(y)}\right] = \mathbb{E}_{p_s(z,y)}[\log p_s(y|z)] + H_{p_s}(Y). \tag{27}$$

Introduce a variational decoder $r_\theta(y|z)$. Using the identity

$$\mathbb{E}_{p_s(y|z)}[\log p_s(y|z)] = \mathbb{E}_{p_s(y|z)}[\log r_\theta(y|z)] + \mathrm{KL}\big(p_s(y|z) \,\|\, r_\theta(y|z)\big), \tag{28}$$

and taking expectation over $p_s(z)$ yields the *exact* decomposition

$$I_{p_s}(Z;Y) = \underbrace{\mathbb{E}_{p_s(z,y)}[\log r_\theta(y|z)] + H_{p_s}(Y)}_{\text{variational term}} + \underbrace{\mathbb{E}_{p_s(z)}\mathrm{KL}\big(p_s(y|z) \,\|\, r_\theta(y|z)\big)}_{\text{gap / interference } \Delta_s(\theta)\geq 0}. \tag{29}$$

Since the gap term $\Delta_s(\theta) \geq 0$, we obtain the Barber–Agakov lower bound:

$$I_{p_s}(Z;Y) \geq \mathbb{E}_{p_s(z,y)}[\log r_\theta(y|z)] + H_{p_s}(Y). \tag{30}$$

Moreover, with $z \sim q(z|x)$ and $(x,y) \sim p_s(x,y)$,

$$\mathbb{E}_{p_s(z,y)}[\log r_\theta(y|z)] = \mathbb{E}_{(x,y)\sim p_s(x,y)}\,\mathbb{E}_{z\sim q(z|x)}\big[\log r_\theta(y|z)\big]. \tag{31}$$

Exact decomposition and variational upper bound for $I_p(Z;X)$ Under the mixture $p(x)$, the mutual information admits

$$I_p(Z;X) = \mathbb{E}_{p(x)}\mathrm{KL}\big(q(z|x) \,\|\, q(z)\big), \qquad q(z) = \int q(z|x)p(x)\,dx. \tag{32}$$

Introduce a prior $p(z)$. A standard decomposition gives

$$\mathbb{E}_{p(x)}\mathrm{KL}\big(q(z|x) \,\|\, p(z)\big) = I_p(Z;X) + \mathrm{KL}\big(q(z) \,\|\, p(z)\big). \tag{33}$$

Since $\mathrm{KL}(q(z)\|p(z)) \geq 0$, we obtain the variational upper bound

$$I_p(Z;X) \leq \mathbb{E}_{p(x)}\mathrm{KL}\big(q(z|x) \,\|\, p(z)\big), \tag{34}$$

where the nonnegative term $\mathrm{KL}(q(z)\|p(z))$ is another *gap / interference* introduced by the variational relaxation.

Variational inference for $\mathcal{J}_{\mathrm{IB}}^{\mathrm{full}}$ and the induced interference terms First note that the pairwise fitting terms in Eq. (26) can be combined as

$$\phi I_{p_\mathrm{A}}(Z;Y) + (1-\phi) I_{p_\mathrm{M}}(Z;Y). \tag{35}$$

Applying Eq. (29) to both $s = \mathrm{A},\mathrm{M}$ yields

$$\begin{aligned}\phi I_{p_\mathrm{A}}(Z;Y) + (1-\phi) I_{p_\mathrm{M}}(Z;Y) = \mathbb{E}_{p(x,y)}\mathbb{E}_{q(z|x)}[\log r_\theta(y|z)] + \underbrace{\phi H_{p_\mathrm{A}}(Y) + (1-\phi)H_{p_\mathrm{M}}(Y)}_{\text{constant}} \\ + \underbrace{\phi\Delta_\mathrm{A}(\theta) + (1-\phi)\Delta_\mathrm{M}(\theta)}_{\text{variational gaps} \geq 0},\end{aligned} \tag{36}$$

where $\Delta_s(\theta) := \mathbb{E}_{p_s(z)}\mathrm{KL}(p_s(y|z)\|r_\theta(y|z))$.

For the compression term, by Eq. (33),

$$-\beta I_p(Z;X) = -\beta \mathbb{E}_{p(x)}\mathrm{KL}(q(z|x)\|p(z)) + \underbrace{\beta\,\mathrm{KL}(q(z)\|p(z))}_{\text{nonnegative gap}}. \tag{37}$$

Dropping the nonnegative gap yields the standard variational relaxation in Eq. (34).

Combining Eqs. (36) and (37), and retaining the latent-state dependence term explicitly, we obtain:

$$\mathcal{J}_{\mathrm{IB}}^{\mathrm{full}} \geq \underbrace{\mathbb{E}_{p(x,y)}\mathbb{E}_{q(z|x)}[\log r_\theta(y|z)] - \beta\mathbb{E}_{p(x)}\mathrm{KL}(q(z|x)\|p(z))}_{\mathcal{L}_{\mathrm{VIB}}(\theta,q)} + I_p(Z;Y;S) + C \tag{38}$$
$$\text{with} \quad C := \phi H_{p_{\mathrm{A}}}(Y) + (1-\phi)H_{p_{\mathrm{M}}}(Y),$$

where the tightness is controlled by the nonnegative gap terms $\phi\Delta_{\mathrm{A}}(\theta) + (1-\phi)\Delta_{\mathrm{M}}(\theta)$ and $\mathrm{KL}(q(z)\|p(z))$.

Maximising $\mathcal{L}_{\mathrm{VIB}}$ with respect to q and r therefore provides a tractable variational optimisation objective for the pairwise contaminated IB terms. Meanwhile, the additional interaction-information term $I_p(Z;Y;S)$ remains in the full theoretical objective as the residual structural-coupling component induced by the latent mixture process.

This recovers the homoscedastic VIB objective presented in Eq. (9) of the main text.

Remark (where heteroscedasticity enters). The derivation above holds for *any* choice of decoder family $r_\theta(y|z)$. Under mixture-induced misalignment, the conditional $p(y|z)$ is intrinsically heterogeneous, suggesting that a constant-noise (homoscedastic) likelihood can be statistically misspecified. A minimal relaxation is to allow input-dependent uncertainty, *e.g.*

$$r_\theta(y|z,x) = \mathcal{N}\big(y;\, \mu_\theta(z),\, \sigma_\theta^2(x)I\big)\,, \tag{39}$$

which induces a sample-wise temperature $\tau(x) = 1/\sigma_\theta^2(x)$ that scales the effective fitting strength. This observation motivates the heteroscedastic bottleneck formulation in the main text.

E.3 Barber–Agakov bound

In this appendix, we provide the variational derivations used. We first recall the Barber–Agakov bound and then derive the explicit decomposition in Eq. (10), which underlies the variational surrogate in Eq. (9).

Lemma E.3 (Barber–Agakov bound)**.** Let $(Y, Z) \sim p(y, z)$ be any joint distribution and let $r(y \mid z)$ be any conditional density such that $r(y \mid z) > 0$ whenever $p(y \mid z) > 0$. Then

$$I_p(Z;Y) \geq \mathbb{E}_{p(y,z)}\big[\log r(y \mid z)\big] + H_p(Y), \tag{40}$$

where $H_p(Y) := -\mathbb{E}_{p(y)} \log p(y)$ is the entropy of Y under p. Moreover, the gap admits the exact expression

$$I_p(Z;Y) = \mathbb{E}_{p(y,z)}\big[\log r(y \mid z)\big] + H_p(Y) + \mathbb{E}_{p(z)}\mathrm{KL}\big(p(y \mid z) \,\|\, r(y \mid z)\big). \tag{41}$$

Proof. By the definition of mutual information,

$$I_p(Z;Y) = \mathbb{E}_{p(y,z)}\Big[\log \frac{p(y \mid z)}{p(y)}\Big] = \mathbb{E}_{p(y,z)}[\log p(y \mid z)] - \mathbb{E}_{p(y)}[\log p(y)]. \tag{42}$$

We now add and subtract $\log r(y \mid z)$ inside the first expectation:

$$\mathbb{E}_{p(y,z)}[\log p(y \mid z)] = \mathbb{E}_{p(y,z)}[\log r(y \mid z)] + \mathbb{E}_{p(y,z)}\Big[\log \frac{p(y \mid z)}{r(y \mid z)}\Big]. \tag{43}$$

For the second term, using the tower property and the definition of KL divergence,

$$\mathbb{E}_{p(y,z)}\Big[\log \frac{p(y \mid z)}{r(y \mid z)}\Big] = \mathbb{E}_{p(z)}\,\mathbb{E}_{p(y|z)}\Big[\log \frac{p(y \mid z)}{r(y \mid z)}\Big] = \mathbb{E}_{p(z)}\mathrm{KL}\big(p(y \mid z) \,\|\, r(y \mid z)\big) \geq 0. \tag{44}$$

Plugging Eqs. (43) and (44) into (42) yields

$$I_p(Z;Y) = \mathbb{E}_{p(y,z)}[\log r(y \mid z)] - \mathbb{E}_{p(y)}[\log p(y)] + \mathbb{E}_{p(z)}\mathrm{KL}\big(p(y \mid z) \,\|\, r(y \mid z)\big), \tag{45}$$

which is exactly Eq. (41). Dropping the nonnegative KL term gives the bound in Eq. (40). This concludes the proof. □

Corollary E.4 (Explicit BA decomposition under p_s)**.** For each $s \in \{\mathrm{A}, \mathrm{M}\}$, applying Lemma E.3 to $(Y, Z) \sim p_s(y, z)$ yields

$$I_{p_s}(Z;Y) = \mathbb{E}_{p_s(x,y)}\mathbb{E}_{q(z|x)}\big[\log r(y \mid z)\big] + \mathbb{E}_{p_s(z)}\mathrm{KL}\big(p_s(y \mid z) \,\|\, r(y \mid z)\big) - \mathbb{E}_{p_s(y)} \log p_s(y), \tag{46}$$

where $p_s(y, z)$ is the joint induced by $p_s(x, y)q(z \mid x)$ and $p_s(z) = \int p_s(x)q(z \mid x)\, dx$.

Proof. Under $p_s(x, y)$ and encoder $q(z \mid x)$, the joint over (y, z) is $p_s(y, z) = \int p_s(x, y)q(z \mid x)\, dx$ and thus $\mathbb{E}_{p_s(y,z)}[\cdot] = \mathbb{E}_{p_s(x,y)}\mathbb{E}_{q(z|x)}[\cdot]$. Substituting this into Eq. (41) in Lemma E.3 gives Eq. (46). □

From the contaminated objective to a variational inference We now derive the variational surrogate in Eq. (9) starting from the contaminated objective. Recall

$$\mathcal{J}_{\mathrm{IB}} = \phi\Big(I_{p_{\mathrm{A}}}(Z;Y) - I_{p_{\mathrm{M}}}(Z;Y)\Big) + I_{p_{\mathrm{M}}}(Z;Y) - \beta I_p(Z;X). \tag{47}$$

Step 1: Variational lower bounds for the sufficiency terms. Since $0 \le \phi \le 1$, the sufficiency terms can be rewritten as $\phi I_{p_{\mathrm{A}}}(Z;Y) + (1-\phi)I_{p_{\mathrm{M}}}(Z;Y)$, with nonnegative coefficients.

Applying Corollary E.4 and dropping the nonnegative mismatch KL terms $\mathbb{E}_{p_s(z)}\mathrm{KL}(p_s(y \mid z)\|r(y \mid z))$ yields for each $s \in \{\mathrm{A}, \mathrm{M}\}$,

$$I_{p_s}(Z;Y) \ge \mathbb{E}_{p_s(x,y)}\mathbb{E}_{q(z|x)}[\log r(y \mid z)] - \mathbb{E}_{p_s(y)} \log p_s(y). \tag{48}$$

The last term $-\mathbb{E}_{p_s(y)} \log p_s(y) = H_{p_s}(Y)$ depends only on the data distribution, hence it is constant w.r.t. the learnable parameters and can be omitted from optimisation.

Step 2: Variational upper bound for the compression term. For the compression term, using $I_p(Z;X) = \mathbb{E}_{p(x)}\mathrm{KL}(q(z \mid x)\|q(z))$ with the aggregated posterior $q(z) = \int p(x)q(z \mid x)\, dx$, and introducing a prior $p(z)$ gives the standard inequality

$$I_p(Z;X) = \mathbb{E}_{p(x)}\mathrm{KL}(q(z \mid x) \,\|\, q(z)) \le \mathbb{E}_{p(x)}\mathrm{KL}(q(z \mid x) \,\|\, p(z))\,, \tag{49}$$

where the inequality follows from the non-negativity of the KL divergence $\mathrm{KL}(q(z)\|p(z)) \ge 0$.

Step 3: Collecting terms under the mixture distribution. Because $p = \phi p_{\mathrm{A}} + (1-\phi)p_{\mathrm{M}}$, the weighted combination of the log-likelihood expectations in Eq. (48) reduces to an expectation under $p(x, y)$. Up to additive constants independent of q and r, we obtain the homoscedastic variational surrogate:

$$\mathcal{J}_{\mathrm{VIB}} = \mathbb{E}_{p(x,y)}\mathbb{E}_{q(z|x)}[\log r(y \mid z)] - \beta\, \mathbb{E}_{p(x)}\mathrm{KL}(q(z \mid x) \,\|\, p(z))\,, \tag{50}$$

which matches Eq. (9) in the main text.

Accounting for the interaction information $I(Z;S;Y)$. To clarify the relation between the original objective and its variational surrogate, let $S \in \{\mathrm{A}, \mathrm{M}\}$ denote the mixture-component indicator, with $p(S = \mathrm{A}) = \phi$ and $p(S = \mathrm{M}) = 1 - \phi$. Under the joint distribution $p(s, x, y, z) = p(s)p_s(x, y)q(z \mid x)$, the weighted sufficiency terms satisfy

$$\begin{aligned} &\phi I_{p_{\mathrm{A}}}(Z;Y) + (1-\phi)I_{p_{\mathrm{M}}}(Z;Y) \\ &\quad = I(Z;Y \mid S) = I_p(Z;Y) - I(Z;S;Y), \end{aligned} \tag{51}$$

where $I(Z;S;Y) = I(S;Y) - I(S;Y \mid Z)$ denotes the three-way interaction information.

Importantly, this term is not assumed to vanish. Using the exact variational decompositions underlying Eqs. (48) and (49), we obtain

$$\begin{aligned} \mathcal{J}_{\mathrm{IB}} - \mathcal{J}_{\mathrm{VIB}} &= H_p(Y) - I(Z;S;Y) \\ &\quad + \mathbb{E}_{q(z)}\mathrm{KL}(p(y \mid z) \,\|\, r(y \mid z)) \\ &\quad + \beta\, \mathrm{KL}(q(z) \,\|\, p(z)) \\ &= H(Y \mid S) + I(S;Y \mid Z) \\ &\quad + \mathbb{E}_{q(z)}\mathrm{KL}(p(y \mid z) \,\|\, r(y \mid z)) \\ &\quad + \beta\, \mathrm{KL}(q(z) \,\|\, p(z)) \\ &\ge H(Y \mid S), \end{aligned} \tag{52}$$

where $p(y \mid z)$ is induced by the mixture distribution $p(x, y)q(z \mid x)$. Thus, the interaction-information contribution is accounted for through the constant $H(Y \mid S)$ and the nonnegative variational gap $I(S; Y \mid Z)$, rather than being assumed zero. Consequently, Eq. (50) provides a valid variational lower bound, up to an additive constant, of the original contaminated objective, although the two objectives are not generally equivalent. □

E.4 Derivation of the Empirical Heteroscedastic VIB Objective

In this appendix, we derive the empirical objective in Eq. (5) from the heteroscedastic Gaussian predictive model in Eq. (4).

Setup and notation Let $\{(x_i, y_i)\}_{i=1}^N$ be i.i.d. samples drawn from the mixture distribution $p(x, y)$. We consider an IB encoder (posterior) $q(z \mid x)$ and a prior $p(z)$. The variational IB surrogate (to maximise) takes the form

$$\mathcal{J}_{\mathrm{VIB}} = \mathbb{E}_{p(x,y)}\mathbb{E}_{q(z|x)}\big[\log r(y \mid z, x)\big] - \beta\, \mathbb{E}_{p(x)}\mathrm{KL}(q(z \mid x) \,\|\, p(z))\,. \tag{53}$$

We instantiate the decoder with a heteroscedastic isotropic Gaussian:

$$r(y \mid z, x) = \mathcal{N}\big(y; \mu(z),\, \sigma^2(x) I_d\big), \tag{54}$$

where $\mu(z) \in \mathbb{R}^d$ and $\sigma^2(x) > 0$.

Log-likelihood of the heteroscedastic Gaussian By the definition of an isotropic Gaussian density,

$$\log r(y \mid z, x) = -\frac{d}{2}\log(2\pi) - \frac{d}{2}\log\sigma^2(x) - \frac{1}{2\sigma^2(x)}\|y - \mu(z)\|_2^2. \tag{55}$$

Dropping the constant term $-\frac{d}{2}\log(2\pi)$ (independent of all learnable variables), maximizing $\log r(y \mid z, x)$ is equivalent to maximizing

$$\widetilde{\log r}(y \mid z, x) = -\frac{d}{2}\log\sigma^2(x) - \frac{1}{2\sigma^2(x)}\|y - \mu(z)\|_2^2. \tag{56}$$

From population objective to sample average We approximate the expectations in Eq. (53) by the empirical average over $\{(x_i, y_i)\}_{i=1}^N$:

$$\mathcal{J}_{\mathrm{VIB}} \approx \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{q(z|x_i)}\big[\widetilde{\log r}(y_i \mid z, x_i)\big] - \beta\,\mathrm{KL}(q(z \mid x_i) \,\|\, p(z))\right). \tag{57}$$

Monte Carlo estimator over the bottleneck samples For each sample i, draw K Monte Carlo samples $z_{ik} \sim q(z \mid x_i)$. Then

$$\mathbb{E}_{q(z|x_i)}\big[\widetilde{\log r}(y_i \mid z, x_i)\big] \approx \frac{1}{K}\sum_{k=1}^{K}\widetilde{\log r}(y_i \mid z_{ik}, x_i). \tag{58}$$

Substituting Eq. (56) into Eq. (58) yields

$$\frac{1}{K}\sum_{k=1}^{K}\widetilde{\log r}(y_i \mid z_{ik}, x_i) = -\frac{d}{2}\log\sigma_i^2 - \frac{1}{2\sigma_i^2}\cdot\frac{1}{K}\sum_{k=1}^{K}\|y_i - \mu(z_{ik})\|_2^2, \tag{59}$$

where we denote $\sigma_i^2 := \sigma^2(x_i)$ for brevity.

Final empirical minimisation objective Since Eq. (57) is a maximization objective, we define the empirical loss to *minimize* as $\mathcal{J} := -\mathcal{J}_{\mathrm{VIB}}$. Combining Eqs. (57)–(59) gives

$$\mathcal{J} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{d}{2}\log\sigma_i^2 + \frac{1}{2\sigma_i^2}\cdot\frac{1}{K}\sum_{k=1}^{K}\|y_i - \mu(z_{ik})\|_2^2 + \beta\,\mathrm{KL}(q(z \mid x_i) \,\|\, p(z))\right] + \mathrm{const}, \tag{60}$$

where const denotes terms independent of learnable variables. Up to a constant scaling of the data-fit term (absorbed into ℓ), Eq. (60) matches Eq. (5) by defining

$$\ell\big(y_i, \mu(z_{ik})\big) \propto -\|y_i - \mu(z_{ik})\|_2^2, \tag{61}$$

or, more generally, letting $\ell(\cdot, \cdot)$ denote any task-specific similarity whose variance-weighted form is implemented by the heteroscedastic likelihood.

Remark (general task-specific loss). If the application uses a similarity-based log-likelihood (*e.g.* cosine-based likelihood on normalised embeddings), the derivation remains identical: the heteroscedastic Gaussian contributes the universal term $\frac{d}{2}\log\sigma_i^2$ and rescales the data-fit term by $1/\sigma_i^2$, while the specific choice of $\ell(\cdot,\cdot)$ changes only the inner estimator. This yields Eq. (5).

F Synthetic misalignment data generation

Given an image and its original caption, we first extract a set of semantic components using rule-based matching over predefined vocabularies, including objects, entities, actions, attributes, and compositional phrases. Based on these extracted components, we generate two types of textual misalignment. **SELECTIVE BIAS** is constructed by randomly removing one to three semantic components with a weighted preference toward contextual information, resulting in captions that are fluent but semantically incomplete. **PERTURBATION BIAS** is constructed by replacing a single semantic component with another value from the same semantic category or a synonym, producing captions that remain grammatical but contain incorrect semantics. This procedure yields controlled and interpretable image–text misalignment without altering the visual content. The pseudocode is provided in Algorithm 1.

Algorithm 1: Constructing Misaligned Caption Triplets on Flickr8k

Input: Image–caption pairs $(\mathcal{I},\mathcal{C})$, sample size N, alignment probability ϕ, Gaussian parameters $\{(\mu_j,\sigma_j^2)\}_{j=0}^{K}$, slot vocabulary $\mathcal{V}$.
Output: Triplets $\{(x_i, y_i^{\text{aligned}}, y_i^{\text{sel}}, y_i^{\text{pert}})\}_{i=1}^{N}$.

Sample N image–caption pairs;
$q \leftarrow \textsc{RandomSimplex}(K)$;
foreach *sampled pair* (x,y) **do**
  $A \leftarrow \textsc{ExtractSemantics}(y,\mathcal{V})$;
  $y^{\text{aligned}}, y^{\text{sel}}, y^{\text{pert}} \leftarrow y, y, y$;
  `// Sample alignment state and GMM strength`
  $z \sim \textsc{Bernoulli}(\phi)$;
  if $z = 1$ **then**
    $s \sim \mathcal{N}(\mu_0, \sigma_0^2)$;
  else
    $j \sim \textsc{Categorical}(q)$;
    $s \sim \mathcal{N}(\mu_j, \sigma_j^2)$;
  $s \leftarrow \textsc{Clip}(s, 0, 1)$;
  if $z = 0$ ***and*** $A \neq \emptyset$ **then**
    `// Type I: selective omission`
    $D \leftarrow \textsc{SampleSlots}(A, s)$;
    $y^{\text{sel}} \leftarrow \textsc{DeleteSlots}(y, D)$;
    `// Type II: semantic perturbation`
    $B \leftarrow \textsc{SampleSlots}(A, s)$;
    $y^{\text{pert}} \leftarrow \textsc{ReplaceSlots}(y, B, \mathcal{V})$;
  $(y^{\text{sel}}, y^{\text{pert}}) \leftarrow \textsc{Validate}(y^{\text{sel}}, y^{\text{pert}}, y)$;
  output $(x, y^{\text{aligned}}, y^{\text{sel}}, y^{\text{pert}})$;

G Visualisation

We further provide additional qualitative comparisons of multimodal attribution maps in Figs. 11–13, using image–text pairs from the CC3M [31] dataset. For each pair, we visualise the attribution maps produced by Grad-CAM [2], M2IB [5], NIB [6], and TIB under the same image–text correspondence. These visualisations illustrate differences in attribution patterns across methods, particularly in how visual regions are associated with the semantic content of the captions. Compared with existing methods, TIB produces more spatially concentrated and less diffuse attribution maps, highlighting semantically relevant regions more precisely.

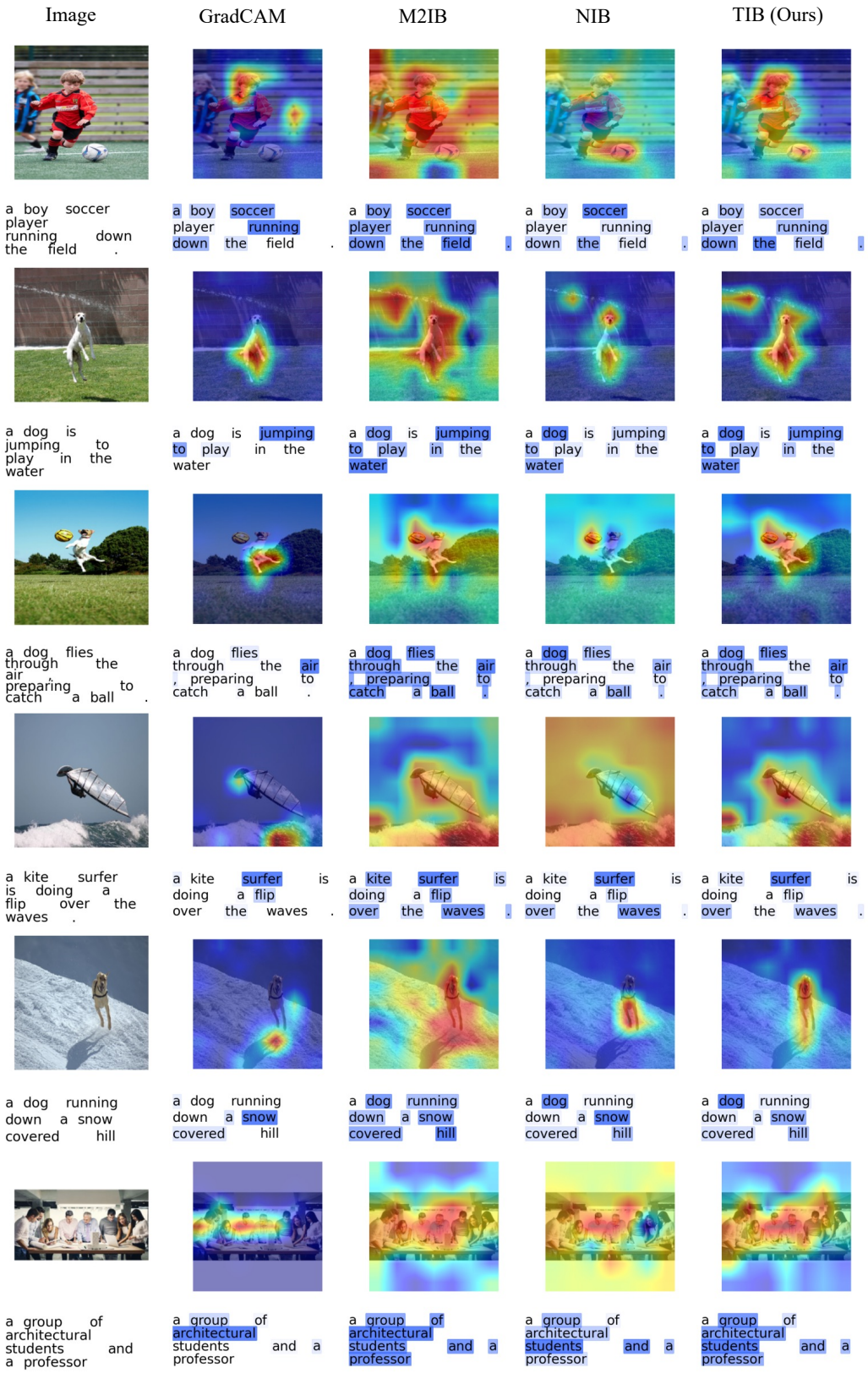

Figure 11: Qualitative comparison of multimodal attribution maps. The maps highlight image regions and text tokens contributing to predictions, illustrating how different methods attribute importance across modalities. More examples on the following pages.

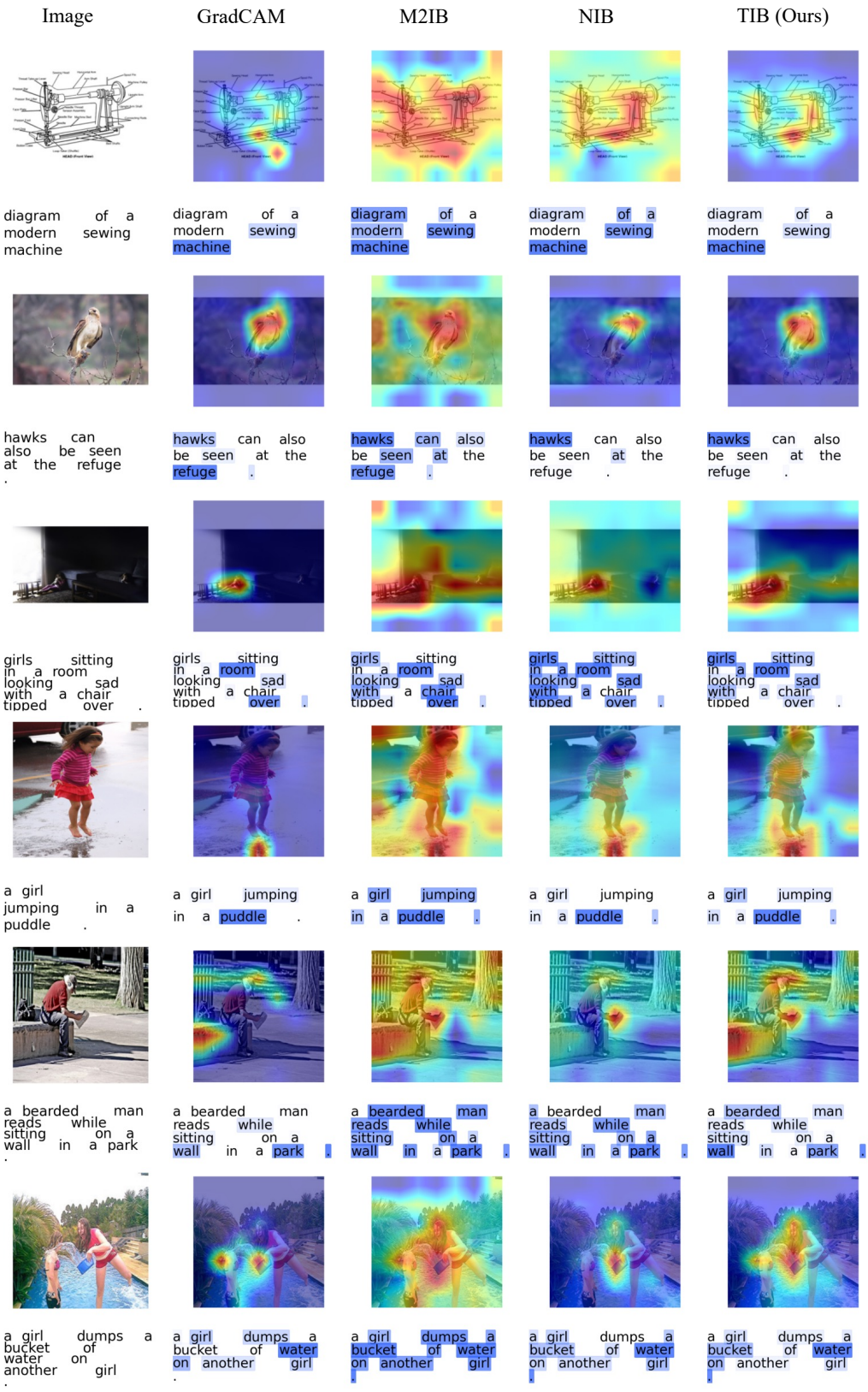

Figure 12: Qualitative comparison of multimodal attribution maps (continued).

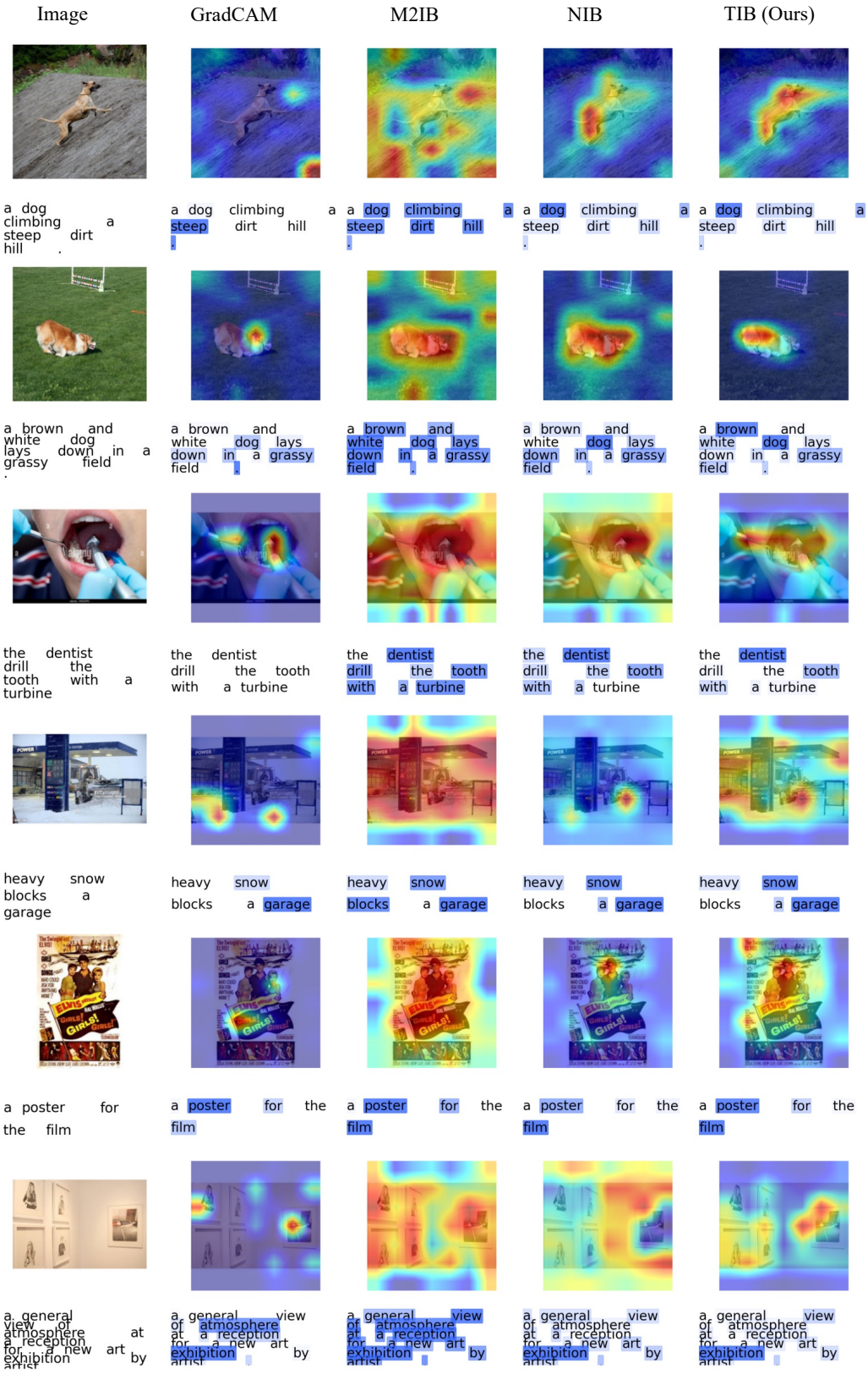

Figure 13: Qualitative comparison of multimodal attribution maps (continued).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarise the paper's main contributions, including the theoretical analysis, the proposed TIB framework, and the experimental evaluation. The scope, assumptions, and empirical claims are consistent with the results presented in the main text and appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discussed the limitation in the appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The proofs are provided in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The codebase is release here for fully reproducibility: https://github.com/Qiming-Huang/TIB.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: Yes, we used the publicly avaiable datasets and have released the codebase.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Yes, in the appendix Sec. C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Justification: Yes, Table 1 reports the mean and variance over five runs with different random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Yes, in the appendix Sec. C.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Yes, in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We used the public avaiable dataset and follow their licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: N/A.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.