Disentangling Modality Contributions for Measuring Multimodal fusion

Anonymous EACL submission

Abstract

Quantifying modality contributions in multimodal systems remains an open challenge, as existing approaches conflate what modality contribution actually means. Most prior work relies on accuracy-based ablations, where performance drops after removing a modality are interpreted as importance. Such outcome-driven metrics, however, cannot distinguish whether a modality is inherently necessary or whether its value arises only through interaction with other modalities. This distinction is particularly critical in cross-attention architectures, where modalities influence each other's representations. In this work, we propose a framework based on Partial Information Decomposition (PID) to evaluate contributions directly from internal embeddings. By decomposing predictive information into unique, redundant, and synergistic components, our method disentangles inherent necessity from cross-modal interaction. To enable scalable analysis, we introduce an inference-only, IPFP-based algorithm that computes layer and dataset-level contributions without retraining. This establishes a principled, representation-level perspective on modality contributions, offering clearer insights than outcome-based metrics alone.

1 Introduction

Recent advances in multimodal learning have enabled models to jointly process information from multiple sensory channels such as vision and language. However, despite their empirical success, many vision-language models (VLMs) fail to achieve genuine multimodal integration. Instead, they often exhibit modality imbalance, a phenomenon where the model disproportionately relies on one modality while neglecting others Peng et al. (2022), Huang et al. (2022), Fan et al. (2023), Wei et al. (2024). As illustrated in Figure 1, this imbalance manifests in different ways: in text-biased examples, the model primarily attends to question phrases while ignoring the visual scene, whereas in image-biased cases, visual context dominates even when textual cues are essential.

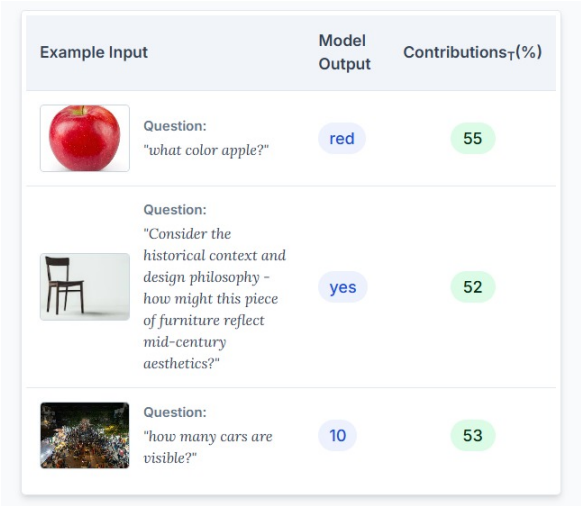

Figure 1: Overview of the proposed modality contribution framework, where text contribution (C_T) and image contribution ($C_I = 100 - C_T$) together form a normalized 100% distribution.

A growing body of work has sought to measure modality contribution, typically by perturbing or masking inputs and observing the effect on model predictions (Gat et al., 2021a), (Wenderoth et al., 2025), (Wang and Wang, 2025). A similar line of work employs gradient and attention-based explanations, such as Integrated Gradients (IG) (Sundararajan et al., 2017) and Grad-CAM (Selvaraju et al., 2017), to attribute importance directly from model sensitivities or attention weights.

Approaches such as MM-SHAP (Parcalabescu and Frank, 2022) and its extensions Wang and Wang (2025), Goldshmidt and Horovicz (2024), and Goldshmidt (2025) estimate each modality's relative influence on the output. However, these methods largely treat modalities as independent sources of information, overlooking the cross-modal interactions that emerge within the model's

internal feature space. Prior work has shown that multimodal fusion is inherently interactional, methods such as EMAP Hessel and Lee (2020), DIME Lyu et al. (2022), and MultiViz Liang et al. (2022) demonstrate that meaningful cross-modal dependencies play a critical role in shaping the final prediction.

Motivated by this connection, we introduce a unified, inference-only framework that jointly quantifies modality attribution and cross-modal interaction through Partial Information Decomposition (PID) (Liang et al., 2023). PID decomposes the total information into redundant, unique, and synergistic components, capturing shared information (r), modality-specific contributions (u_1, u_2), and cross-modal interactions (s). This decomposition naturally disentangles individual modality effects from their joint influence, providing a principled view of how fusion emerges within the model. Previous work has used Sinkhorn-based normalization (Cuturi, 2013) to compute PID, but this requires training additional networks to estimate these relationships. We instead convert this process into an inference-only formulation. Specifically, we propose an Iterative Proportional Fitting Procedure (IPFP) (Bacharach, 1965) based approximation that retains the theoretical grounding of the Sinkhorn approach while removing the need for training. This makes our method parameter-free, efficient, and directly applicable for post-hoc multimodal analysis.

Our key contributions are as follows:

- We propose the first modality contribution metric that jointly captures modality-specific and interactional effects, grounded in interpretable embedding-space analysis.
- We develop a quantification approach based on PID and propose a novel, computationally efficient metric derived using the Iterative Proportional Fitting Procedure (IPFP) algorithm.
- We evaluate our method across diverse VLMs and datasets, supported by synthetic experiments and ablation studies that analyze prompt length, modality quality, and layer-wise cross-attention behavior.

2 Related work

2.1 Gradient and Attention-Based Methods

Gradient-based approaches, such as Integrated Gradients (IG) Sundararajan et al. (2017) attribute importance through path-integrated gradients but suffer from instability, baseline sensitivity, and high attribution noise (Zhuo and Ge, 2024). By relying solely on explicand gradients and neglecting counterfactual ones, IG fails to capture higher-order effects, like redundant or synergistic contributions, where features jointly influence predictions in non-additive ways and violates key Shapley axioms, yielding spurious attributions. More broadly, gradient saliency often reflects the model's implicit density induced by the softmax layer rather than genuine discriminative reasoning (Srinivas and Fleuret, 2020).

Attention-based explanations share similar pitfalls. Attention weights may misrepresent causal polarity (Liu et al., 2022), fail to correlate with true feature importance (Wiegreffe and Pinter, 2019), and can be adversarially altered without affecting predictions (Serrano and Smith, 2019). White-box visualization methods such as attention maps and Grad-CAM Selvaraju et al. (2017) offer coarse interpretability but remain architecture-dependent and cannot disentangle synergistic from suppressive cross-modal effects. Overall, these methods conflate correlation with causation and lack robustness, while provide only coarse unimodal or architecture-specific insights, making them inadequate for diagnosing fine-grained modality interactions or multimodal collapse.

2.2 Perturbation based methods

Early studies on unimodal collapse, the tendency of multimodal models to over-rely on a single modality, focused on uncovering dataset biases that enable strong unimodal performance (Goyal et al., 2017a). Subsequent work introduced perturbation-based tests to quantify modality importance by observing performance degradation when one modality is removed or altered. These approaches fall into two categories: deletion methods, which suppress or mask modality features (Shekhar et al., 2017a; Madhyastha et al., 2018; Frank et al., 2021), and contradiction or foiling methods, which inject misleading inputs such as swapped captions or textual foils (Gat et al., 2021b; Shekhar et al., 2019; Parcalabescu et al., 2022). However, their reliance on accuracy-based evaluation limits interpretability, as accuracy of the models outweighs genuine cross-modal understanding. This dependence has led to inconsistent findings across studies and highlights the need for performance-agnostic approaches that

more reliably capture true modality reliance and detect unimodal collapse.

2.3 SHAPley-based methods

Shapley-based methods are based from Shapley theory and offer a principled framework for attributing model predictions by averaging each feature's marginal contribution over all possible combinations of other features. TokenSHAP Goldshmidt and Horovicz (2024) and PixelSHAP Goldshmidt (2025) extend this framework to tokens and image regions, respectively, yielding fine-grained unimodal attributions but missing cross-modal dependencies. MultiSHAP (Wang and Wang, 2025) models patch-token interactions to separate synergistic and suppressive effects, but it incurs high Monte Carlo cost. MM-SHAP (Parcalabescu and Frank, 2022) introduces a performance-agnostic metric that aggregates absolute Shapley contributions per modality, however it captures only aggregate modality proportions, discarding directional and interaction information and remaining sensitive to masking artifacts.

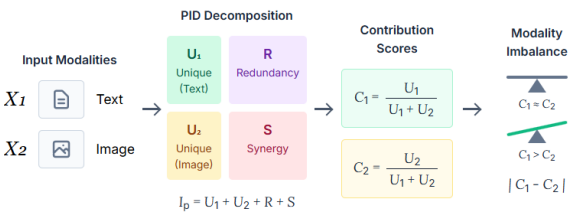

Figure 2: Overview of the proposed modality contribution framework

3 Methodology

Partial Information Decomposition (PID) (Williams and Beer, 2010) extends classical information theory to multiple sources by decomposing the total mutual information $I_p(X_1, X_2; Y)$ into four distinct components:

- **Redundant information** (R): information about the target that is shared by both sources.

$$R = \max_{q \in \Delta_p} I_q(X_1; X_2; Y) \tag{1}$$

- **Unique information from source 1** (U_1): information about the target provided exclusively by the first source.

$$U_1 = \min_{q \in \Delta_p} I_q(X_1; Y \mid X_2) \tag{2}$$

- **Unique information from source 2** (U_2): information about the target provided exclusively by the second source.

$$U_2 = \min_{q \in \Delta_p} I_q(X_2; Y \mid X_1), \tag{3}$$

- **Synergistic information** (S): information that arises only when the two sources are considered together, capturing their interaction.

$$S = I_p(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y) \tag{4}$$

This decomposition explicitly disentangles individual modality contributions (U_1, U_2) from cross-modal interactions (R, S). In multimodal models, each source corresponds to a distinct modality (e.g., text and image), while the target Y represents the fused output embedding. PID thus provides a principled framework to quantify redundant, unique, and synergistic information among modalities.

3.1 Computing PID

Let $X_1 \in \mathcal{X}_1$ and $X_2 \in \mathcal{X}_2$ denote the text and image embeddings, respectively, and $Y \in \mathcal{Y}$ represent the output embeddings. Define the simplex of all joint distributions,

$$\Delta = \mathcal{P}(X_1 \times X_2 \times Y)$$

and the affine feasible set.

$$\Delta_p = \big\{ q \in \Delta : q(x_i, y) = p(x_i, y), \quad \forall x_i \in \mathcal{X}_\rangle,\ y \in \mathcal{Y},\ i \in \{1, 2\} \big\} \tag{5}$$

which preserves the observed modality–target marginals $p(x_1, y)$ and $p(x_2, y)$.

Following Liang et al. (2023), the PID components are obtained via a KL-projection formulation:

$$H_q(Y \mid X_1, X_2) = \log |\mathcal{Y}| - \mathrm{KL}\big(q \,\|\, \tilde{q}\big), \quad \text{where} \quad \tilde{q}(x_1, x_2, y) = \frac{q(x_1, x_2)}{|\mathcal{Y}|} \tag{6}$$

Here, the joint distribution $q(x_1, x_2, y)$ can be represented as a tensor Q of shape $|X_1| \times |X_2| \times |\mathcal{Y}|$. However, optimizing over the full joint tensor $Q \in \mathbb{R}^{|\mathcal{X}_1| \times |\mathcal{X}_2| \times |\mathcal{Y}|}$, is computationally expensive ($\mathcal{O}\left((|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdot |\mathcal{Y}|)^3 \right)$).

To solve this, we first consider the corresponding convex problem that solves Equation (6).

$$
\begin{aligned}
&\arg\min_{Q,\tilde{Q}} F(Q,\tilde{Q}) = \mathrm{KL}(Q \,\Big\|\, \tilde{Q}) \\
\text{s.t.}\quad &\tilde{Q}(x_1,x_2,y) = \frac{Q(x_1,x_2)}{|\mathcal{Y}|} \quad \forall x_1,x_2,y, \\
&\sum_{x_2} Q(x_1,x_2,y) = p(x_1,y), \quad \forall x_1,y, \\
&\sum_{x_1} Q(x_1,x_2,y) = p(x_2,y), \quad \forall x_2,y, \\
&Q \geq 0, \quad \forall x_1,x_2,y, \\
&\sum_{x_1,x_2,y} Q(x_1,x_2,y) = 1.
\end{aligned} \tag{7}
$$

The objective F is jointly convex over the positive orthant with affine constraints, making Equation (7) a convex program.

Theorem 1 (Existence and uniqueness of the KL-projection)**.** *Under the standing assumptions, the convex subproblem in Equation 7 admits a unique minimizer $Q^\star \in \Delta_p$.*

Proof. The result follows from (Csiszár et al., 2004) □

Remark 1. *Uniqueness implies the Q-step is well-posed: for any fixed $\tilde{Q}$ there is a single KL-minimizer in Δ_p, justifying an exact block update*

We solve it via alternating exact minimization over two blocks.

- **Q–step (I-projection)**: Given the current iterate $\tilde{Q}^{(t)}$, the next iterate is obtained via KL projection onto the feasible set Δ_p:

$$
Q^{(t+1)} = \operatorname*{argmin}_{Q\in\Delta_p} \mathrm{KL}(Q \,\|\, \tilde{Q}^{(t)}). \tag{7}
$$

  This projection enforces consistency with the observed marginals $p(x_1,y)$ and $p(x_2,y)$, while remaining information-theoretically close to $\tilde{Q}^{(t)}$. Since the constraints in Δ_p factorize over labels y, the problem decomposes into independent subproblems over each $Q_y(x_1,x_2)$. Each subproblem corresponds to a KL-regularized optimal transport problem, which can be efficiently solved using Sinkhorn iterations (Cuturi, 2013).

- **$\tilde{Q}$-step (closed-form)**: Given the updated coupling $Q^{(t+1)}$, the auxiliary variable $\tilde{Q}$ is obtained by minimizing its KL divergence to $Q^{(t+1)}$:

$$
\tilde{Q}^{(t+1)}(x_1,x_2,y) = \frac{Q^{(t+1)}(x_1,x_2)}{|Y|}
$$

  under the coupling constraint. This update admits a closed-form analytic solution:

$$
\tilde{Q}^{(t+1)} = \arg\min_{\tilde{Q}} \mathrm{KL}(Q^{(t+1)} \,\|\, \tilde{Q})
$$

  subject to the coupling constraint.

This alternation monotonically decreases $F(Q,\tilde{Q})$ and converges globally.

Theorem 2 (Global convergence of alternating KL-projections)**.** *The objective function $(Q^{(t)},\tilde{Q}^{(t)})$ decreases with each iteration:*

$$
F(Q^{(t+1)},\tilde{Q}^{(t+1)}) \leq F(Q^{(t)},\tilde{Q}^{(t)}) \quad \forall t.
$$

The sequence converges to a unique global minimum $F^\star \geq 0$, with all cluster points being global minimizers of Equation (7). Moreover, the sequence $Q(t)$ itself converges to a unique minimizer $Q^\star$.

Proof. The result follows from Dimitri et al. (1999) □

Remark 2. *Theorem 2 guarantees that the alternating KL-projection scheme is both monotonic and globally convergent. Each update strictly decreases the objective until reaching the unique fixed point $Q^\star$, ensuring stable convergence of the IPFP iterations without dependence on initialization.*

Direct computations in floating point can suffer from numerical underflow due to repeated exponentiation. Following Cuturi (2013), we perform each label-specific IPFP update in the log domain for numerical stability. For a fixed label y, let $A_y = \tilde{Q}(\cdot,\cdot,y) > 0$ denote the kernel matrix and define the target marginals as

$$
r_y(x_1) = p(x_1,y), \quad c_y(x_2) = p(x_2,y).
$$

We introduce dual scaling vectors u_y and v_y to enforce these marginal constraints. The log-domain Sinkhorn iterations are thus, given by

$$
\begin{aligned}
\log u_y &\leftarrow \log r_y \;-\; \mathrm{LSE}_{x_2}\big(\log A_y + \log v_y\big), \\
\log v_y &\leftarrow \log c_y \;-\; \mathrm{LSE}_{x_1}\big(\log A_y + \log u_y\big),
\end{aligned} \tag{9}
$$

where $\mathrm{LSE}_{x_2}(M)_i = \log \sum_j \exp(M_{ij})$ denotes the log-sum-exp operator. After convergence, the coupling R_y is recovered by exponentiation:

$$\begin{aligned} R_y &= \exp\big(\log A_y + \log u_y + \log v_y\big) \\ &= \mathrm{diag}(u_y)\, A_y\, \mathrm{diag}(v_y), \end{aligned} \tag{10}$$

which by construction satisfies

$$\begin{aligned} \sum_{x_2} R_y(x_1, x_2) &= r_y(x_1), \\ \sum_{x_1} R_y(x_1, x_2) &= c_y(x_2) \end{aligned} \tag{8}$$

when the log-iterations converge, giving an efficient solution of complexity $\mathcal{O}\left((|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdot |\mathcal{Y}|)\right)$

3.2 Bias Metric

Having obtained scalable estimates of the PID components via our IPFP-based solver, we now construct interpretable modality contribution metrics that directly quantify how much each modality contributes to model predictions.

We define the normalized contribution of modality i as

$$C_i = \frac{U_i}{\sum_j U_j}, \tag{9}$$

where U_i is the unique information provided by modality i. The sum in the denominator ensures that the contributions are expressed as fractions of the total unique information across all modalities. This normalization allows C_i to be interpreted as the relative importance of each modality, independent of the absolute scale of the unique information and also ensures that the contributions sum upto 1. We explain the complete framework in Figure 2

3.3 Why PID enables a contribution score

PID provides a principled decomposition of the total information that multiple sources convey about a target, into redundant, unique, and synergistic components. By isolating the unique information (U_1, U_2) from shared and synergistic effects, PID allows us to clearly distinguish the information contributed by each modality individually from that arising due to cross-modal interactions, preventing confusion between independent and interactional contributions. Importantly, this approach is accuracy-agnostic: it quantifies information content regardless of whether the model makes correct predictions (Parcalabescu and Frank, 2022). Moreover, recent work has highlighted the limitations of perturbation-based methods, advocating for analyses that consider the internal representations of models rather than only the final outputs (Sim et al., 2025). PID naturally aligns with these recommendations, as it operates directly on the model's internal feature distributions, enabling computation of modality contribution scores while capturing cross-modal interactions without relying on model perturbations.

4 Implementation Details

Following Liang et al. (2023), we discretize continuous modality embeddings into finite-support distributions for PID estimation using feature binning and clustering. For high-dimensional embeddings (text, vision, output), we apply K-means clustering to obtain discrete supports $|X_1|, |X_2|, |Y|$.

The IPFP solver runs up to 50 iterations, each with at most 100 log-domain Sinkhorn updates Eq. 8 per label. The loops halts when marginal deviation or relative change in the KL objective $F(Q, \tilde{Q})$ change falls below 1×10^{-8}. A small floor ($\varepsilon \approx 10^{-20}$) prevents numerical underflow, while Laplace smoothing ($\alpha \in [10^{-6}, 10^{-3}]$) stabilizes sparse empirical counts.

5 Experiments

5.1 Synthetic dataset

Our first goal is to evaluate the accuracy of our proposed estimators with respect to the ground truth or human judgment as done in Liang et al. (2023). We sample from a binary bitwise distribution where each bitwise operator's PID can be solved exactly when y and x_i's are discrete and low-dimensional.

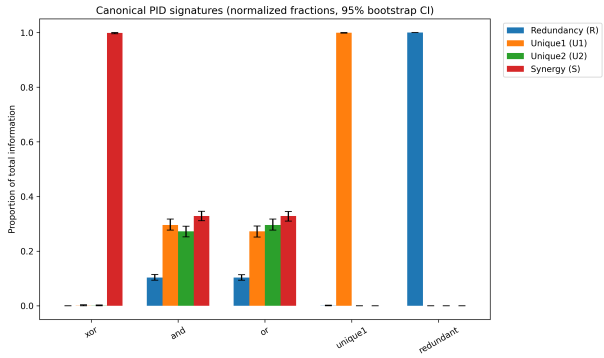

Figure 3: Comparison of our proposed estimators against ground-truth PID values on synthetic datasets constructed using binary bitwise operations

The AND dataset ($y = x_1 \wedge x_2$) exhibits medium synergy because the output is 1 only when both in-

puts are 1, meaning joint information from both inputs is needed to predict y. It has little redundancy since neither input alone reliably predicts $y = 1$, but some unique information is present because if either input is 0, the output is 0. The OR dataset ($y = x_1 \vee x_2$) shows low synergy since $y = 1$ if either input is 1, but high redundancy because both inputs individually provide overlapping information about the output, reducing unique information. In the XOR case ($y = x_1 \oplus x_2$), synergy is high because neither input alone predicts y, and there is zero redundancy or unique information in the individual inputs; only their combination reveals the output. The unique1 dataset ($y = x_1$) contains no synergy or redundancy because x_2 adds no information, and all information about y is unique to x_1. Finally, the redundant dataset, where both inputs are identical (x_1, x_1) and $y = x_1$, has no synergy but maximum redundancy since both predictors share the exact same information about y. As seen in Figure 3, our estimators closely match the ground-truth PID values, confirming the correctness of the method.

5.2 Synthetic Validation of Fusion Mechanisms

To verify that the proposed metric behaves as expected under controlled conditions, we conduct a synthetic validation experiment using two independent Gaussian variables: $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{N}(0, 1)$ To ensure statistical independence, the inputs are decorrelated using Independent Component Analysis (ICA) (Hyvärinen et al., 2001). The synthetic output Y is generated under several predefined fusion rules: additive ($Y = X_1 + X_2$), weighted additive ($Y = X_1 + 10X_2$ and $Y = X_1 + 100X_2$) and and multiplicative ($Y = X_1 \times X_2$). These formulations emulate different multimodal interaction regimes, from balanced contributions to strongly biased or multiplicative relationships. Because the ground-truth modality contributions are analytically known, this setup provides a reliable means of validating the metric's correctness. As expected, additive and multiplicative fusion produce approximately equal contributions from both modalities (around 50%), while weighted fusion shifts the contributions toward the modality with the larger coefficient as shown in Figure 4. These results confirm that the proposed metric faithfully captures true modality influence and reflects the underlying fusion dynamics rather than random variation.

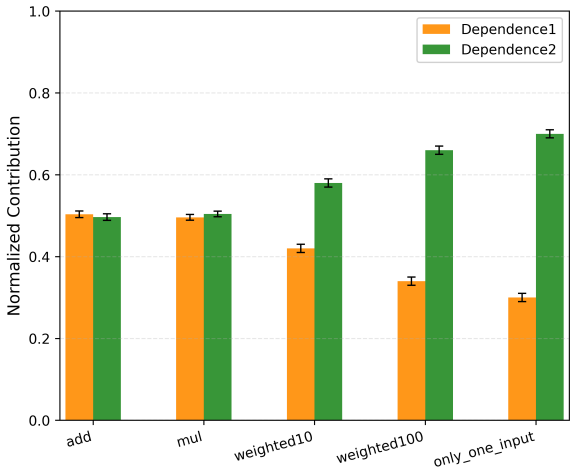

Figure 4: Synthetic validation of the proposed PID-based metric based on rules: additive, weighted additive, and multiplicative

5.3 Real world datasets

We evaluate our PID-based modality contribution metric across multiple datasets, models, and tasks to assess its ability to capture modality influence and cross-modal interactions. Our primary goal is to examine whether the metric reinforces known patterns of modality imbalance in these datasets, especially in the absence of ground-truth contributions, where prior knowledge provides a useful reference for validating the results. cite

5.3.1 Experimental Setup

We apply our method on pretrained models without additional training. For each model, we extract the relevant intermediate representations and compute the PID decomposition, yielding per-sample contributions from each modality. This setup allows us to compare contributions across modalities and across tasks, ensuring that the results reflect model behavior rather than dataset-specific artifacts or random variation (Parcalabescu and Frank, 2022)

5.3.2 Models

We evaluate our metric across models that differ in both fusion strategy and model scale. Specifically, we consider two primary fusion paradigms: (1) cross-attention based fusion, which enables dynamic interaction between modalities, and (2) concatenation-based fusion, which combines modality embeddings in a shared feature space without explicit cross-modal exchange. BLIP

Model	VQA		GQA		FOIL				ScienceQA		NLVR	
	Text	Image	Text	Image	Orig		Foil		Text	Image	Text	Image
					Text	Image	Text	Image				
Paligemma	53.60	46.39	56.44	43.55	55.10	44.90	54.80	45.20	53.29	46.70	48.63	51.36
SmolVLM	60.90	39.10	58.30	41.60	59.80	40.10	60.10	39.80	53.30	46.60	65.18	34.81
LLaVA	59.16	40.84	57.90	42.10	58.50	41.50	58.70	41.30	54.20	45.80	52.40	47.60
BLIP	55.45	44.55	54.58	45.42	58.94	41.05	57.32	42.68	52.90	47.10	53.80	46.10

Table 1: Modality-specific performance (%) of various models across multiple benchmarks. The FOIL dataset is subdivided into Orig and Foil settings, each reporting results with text-only and image-only input.

(Li et al., 2022), a relatively compact 385M-parameter model, exemplifies the cross-attention approach through its Query Transformer (Q-Former), which bridges frozen image and text encoders via lightweight query-based attention for fine-grained alignment. In contrast, LLaVA (Liu et al., 2023) and PaLI-Gemma (Steiner et al., 2024) employ concatenation-based fusion within large-scale architectures (7B parameters), while SMOLVLM (Marafioti et al., 2025), with 500M parameters, serves as an efficiency-oriented variant of the same paradigm. This range from different fusion paradigms to different model sizes, allows us to examine how fusion mechanisms and model capacity jointly influence modality contributions and cross-modal interactions.

5.3.3 Tasks and Datasets

We evaluate our metric on five multimodal benchmarks covering diverse modality dependencies. VQA (Goyal et al., 2017b) and GQA (Hudson and Manning, 2019) represent balanced multimodal tasks that require joint visual–textual reasoning, making them ideal for testing whether our metric captures equal modality influence. NLVR2 (Suhr et al., 2018) focuses on visual reasoning, where models must determine whether textual descriptions align with paired images, providing a setting to evaluate cross-modal grounding and potential visual dominance. ScienceQA (Saikh et al., 2022) features variable modality reliance, as different question types depend more heavily on either text or visuals, allowing us to assess sensitivity to task-specific modality shifts. Finally, FOIL-IT! (Shekhar et al., 2017b) introduces controlled textual discrepancies, testing whether the metric can detect how subtle linguistic perturbations alter modality contributions.

5.3.4 Results

Across VQA and GQA, all major models—BLIP, LLaVA, and PaLI-Gemma—show consistent modality attributions, with text contributing around 55% and image 45%, indicating a balanced integration of linguistic and visual information. This alignment across architectures confirms that our metric captures a stable and interpretable modality balance, consistent with prior findings that both datasets encourage joint reasoning over text and image content. SmolVLM, however, exhibits a slightly higher text bias (60%), likely reflecting its smaller capacity and lighter visual backbone. The near-equal contributions observed for larger models can also be attributed to their VQA-style training objectives, which emphasize textual regularities while maintaining visual grounding (Goyal et al., 2017a). Overall, these results validate that balanced benchmarks like VQA and GQA yield moderate, near-equal modality contributions when analyzed through our information-theoretic framework.

On ScienceQA, modality contributions show greater variability, with text slightly dominating ($\approx$ 53-55%) due to the dataset's heterogeneous question types. BLIP and SmolVLM exhibit more balanced behavior, indicating better adaptation to visual reasoning, whereas LLaVA and PaLI-Gemma remain more text-oriented.

For NLVR2, a visually grounded reasoning task, most models exhibit an increase in image contribution, consistent with its emphasis on spatial and relational understanding. However, SmolVLM remains notably text-oriented ($\approx$ 65% text vs. 35% image), suggesting limited visual grounding likely due to its smaller capacity and lightweight visual encoder. In contrast, BLIP, LLaVA, and PaLI-Gemma demonstrate a more balanced distribution between modalities, aligning with their larger architectures and more expressive fusion mechanisms.

On the FOIL dataset, we note that the modality contribution scores do not deviate significantly from the baseline, indicating that minor textual perturbations do not drastically alter the model's overall modality balance. This behavior is consistent with observations from (Parcalabescu and Frank, 2022), which similarly reported stability under subtle input modifications. Among models, PaliGemma continues to show relatively lower text contribution, while SmolVLM exhibits a stronger bias toward text, reinforcing trends observed in other benchmarks.

6 Ablation

6.1 Fusion Method Ablation

We evaluated how different multimodal fusion strategies affect the computed modality dependence. Table 2 summarizes the mean of the image and text contributions across four fusion methods: cross-attention from image to text, cross-attention from text to image, simple concatenation, and bidirectional fusion.

Fusion	Image	Text
Img→Txt	55.3	44.7
Txt→Img	38.5	61.5
Concat	45.6	54.4
Bi-fusion	54.7	45.3

Table 2: Ablation on the different fusion methods and the modality contributions.

From the results, we observe that cross-attention direction significantly affects the modality balance: image→text slightly favors the image, whereas text→image overweights the text, potentially underestimating image contribution. Concatenation as well as the bidirectional fusion yields moderately balanced dependence indicating a more stable and balanced utilization of image and text information. This suggests that symmetric fusion methods may be preferable in scenarios where both modalities are expected to contribute equally, and that text→image cross-attention may introduce systematic bias toward text.

6.2 Modality perturbations

To further investigate the sensitivity of the model to modality-specific perturbations, we conducted an ablation by varying the textual as well as image input. Specifically, we examined two forms of textual corruption: (1) randomly shuffling the order of words and (2) replacing textual inputs with dummy text (`Lorem Ipsum`). For perturbing the image input, we chose to apply Gaussian noise to it. Table 3 summarizes the mean of the image and text contribution for each condition.

Perturbation	Image	Text
Original	42.6	57.4
Noisy image	44.9	55.1
Dummy text	47.7	52.3
Jumbled text	38.8	61.2

Table 3: Ablation study showing the mean of image and text contributions under different input perturbations

The results show that perturbations shift modality contributions in a consistent and interpretable manner. When noise is added to the images, the metric indicates a slight increase in image dependence, suggesting that the model adjusts its reliance based on the altered input. Similarly, replacing the text with dummy content or jumbling word order leads to increased text contribution, reflecting the model's sensitivity to disruptions in the linguistic signal. The increase in text contribution for jumbled inputs highlights the importance of positional encoding in capturing meaningful semantic structure.

7 Conclusion

We presented a principled framework for quantifying modality contributions in multimodal models using Partial Information Decomposition (PID). By operating on internal representations, our method distinguishes unique, redundant, and synergistic information, separating inherent modality importance from cross-modal interactions. Validation on synthetic data and benchmarks including VQA, GQA, NLVR2, ScienceQA, and FOIL across diverse models (BLIP, LLaVA, PaLI-Gemma, SMOLVLM) shows that the metric produces consistent, interpretable patterns aligned with known dataset biases. These results establish our approach as a scalable, model-agnostic, and performance-independent tool for analyzing multimodal interactions.

8 Limitations

While our PID-based framework provides a principled lens for analyzing modality contributions, it

is not without limitations. First, the PID estimates are sensitive to noise in the embedding space, as small variations in feature distributions can influence the inferred information components. Second, as with other recent analyses, the method may not be fully immune to spurious correlations in the data or model representations, which could affect the attribution scores. Finally, the approach assumes access to intermediate model embeddings, which may not always be available for closed-source or API-restricted systems. Addressing these challenges, particularly improving robustness to noise and mitigating the influence of spurious correlations, remains an important direction for future work.

References

Michael Bacharach. 1965. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310.

Imre Csiszár, Paul C Shields, et al. 2004. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

P Dimitri et al. 1999. Nonlinear programming.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Itai Gat, Idan Schwartz, and Alex Schwing. 2021a. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643.

Itai Gat, Idan Schwartz, and Alexander Schwing. 2021b. Perceptual score: What data modalities does your model perceive?

Roni Goldshmidt. 2025. Attention, please! pixelshap reveals what vision-language models actually focus on. *ArXiv*, abs/2503.06670.

Roni Goldshmidt and Miriam Horovicz. 2024. Tokenshap: Interpreting large language models with monte carlo shapley value estimation.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

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

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! *arXiv preprint arXiv:2010.06572*.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, pages 9226–9259. PMLR.

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

Aapo Hyvärinen, Jarmo Hurri, and Patrik O Hoyer. 2001. Independent component analysis. In *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, pages 151–175. Springer.

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

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. 2023. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393.

Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2022. Multiviz: Towards visualizing and understanding multimodal models. *arXiv preprint arXiv:2207.00056*.

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

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. 2022. Rethinking attention-model explainability through faithfulness violation test. In *International Conference on Machine Learning*, pages 13807–13824. PMLR.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 455–467.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2018. Defoiling foiled image captions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 433–438, New Orleans, Louisiana. Association for Computational Linguistics.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2022. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. *arXiv preprint arXiv:2212.08158*.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

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

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Ravi Shekhar, Ece Takmaz, Raquel Fernández, and Raffaella Bernardi. 2019. Evaluating the representational hub of language and vision models. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 211–222, Gothenburg, Sweden. Association for Computational Linguistics.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470.

Suraj Srinivas and François Fleuret. 2020. Rethinking the role of gradient-based attribution methods for model interpretability. ArXiv preprint.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Zhanliang Wang and Kai Wang. 2025. Multishap: A shapley-based framework for explaining cross-modal interactions in multimodal ai models. *arXiv preprint arXiv:2508.00576*.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. 2024. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27338–27347.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. 2025. Measuring cross-modal interactions in multimodal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21501–21509.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Paul L Williams and Randall D Beer. 2010. Non-negative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Yue Zhuo and Zhiqiang Ge. 2024. Ig 2: Integrated gradient on iterative gradient path for feature attribution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7173–7190.