The Metagame of Interpretability and Meta-Attributions

Hubert Baniecki[1,2]*, **Przemyslaw Biecek**[1,2], **Fabian Fumagalli**[3]
[1]University of Warsaw
[2]Centre for Credible AI, Warsaw University of Technology
[3]Bielefeld University

Abstract

How can an arbitrary attribution method be generalized from first principles to capture interactions? We answer this with the METAGAME, a conceptual framework for quantifying second-order interaction effects of model explanations. We cast the attribution value ϕ_i of feature i as a cooperative game among the other features and compute its Shapley value, which measures how much feature j influences the attribution of i, yielding the directional *meta-attribution* $\varphi_{j\to i}$. By decomposing attribution itself rather than the model directly, meta-attributions extend any gradient- or attention-based method to interactions, uniting removal-based perturbations with model internals. Theoretically, we prove that meta-attributions sum to the first-order attribution they explain, a hierarchical decomposition that Shapley interactions and integrated Hessians turn out to perform implicitly. Empirically, we demonstrate that meta-attributions deliver insights across diverse interpretability applications: (i) quantifying token interactions in instruction-tuned language models, (ii) explaining cross-modal similarity in vision-language encoders, and (iii) interpreting text-to-image concepts in multimodal diffusion transformers.

1 Introduction

Attribution is the task of quantifying how specific variables, such as input tokens, training data, or internal components, influence a machine learning model's behavior. Attribution methods serve as the cornerstone of key interpretability approaches, driving advancements in concept-based explainability for vision models (Fel et al., 2023), circuit discovery (Syed et al., 2024), data valuation (Wang et al., 2025), and instruction-following in language models (Komorowski et al., 2026). Despite their broad utility, simple attributions based on the notions of cooperative game theory (Lundberg and Lee, 2017), gradients (Sundararajan et al., 2017), or relevance propagation (Achtibat et al., 2024) are limited by capturing only *first-order* influence of a single variable, omitting higher-order dynamics encoded by the model, such as synergies and antisynergies (see an example in Figure 1). To this end, interaction methods that measure the *second-order* influence of input pairs, such as Shapley interactions (Fumagalli et al., 2023) and integrated Hessians (Janizek et al., 2021), have been proposed. However, it remains an open challenge *whether*, and *how*, an arbitrary first-order attribution method can be generalized from first principles to capture higher-order interactions.

We answer this question by revisiting the concept of *explaining explanations* (Janizek et al., 2021), generalizing the idea behind their serial use of integrated gradients to quantify how a particular input feature influences an attribution value of another feature. Lundstrom and Razaviyayn (2023) have theoretically studied the leakage of interaction effects into individual terms in such *serial* methods. Furthermore, explicitly computing second-order derivatives such as input Hessians often requires architectural modifications (Dombrowski et al., 2019; Janizek et al., 2021) and specialized computational techniques for large input sizes (Pramanik et al., 2026). Existing *set-based* methods assign interaction scores to groups of features without distinguishing how one feature influences the

*Corresponding author: h.baniecki@uw.edu.pl Code: https://github.com/credibleai/metagame

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

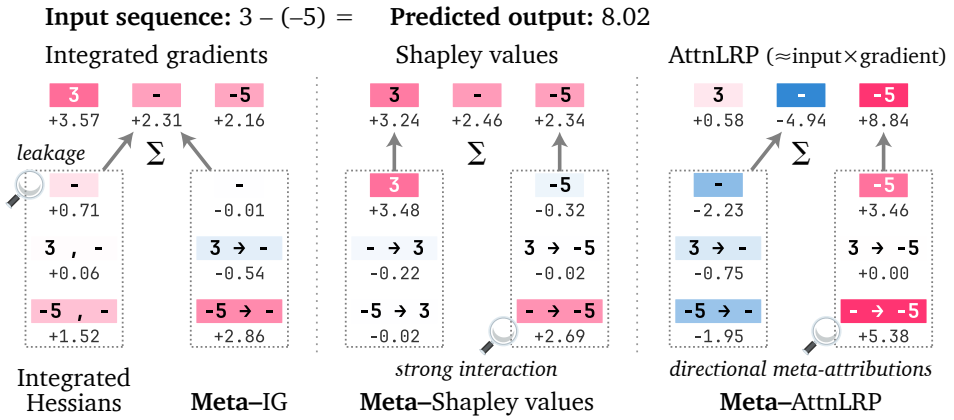


Figure 1: Complementary interpretations of a simple transformer solving integer addition. First-order attributions implicitly capture interactions between token pairs. Integrated Hessians highlight the undirected interaction `-5 , -`, but exhibit information leakage for `-`, which should be zero without knowing the subsequent number. AttnLRP, by contrast, assigns opposite-signed local attributions to `-` and `-5`. Our proposed meta-attributions assign individual effects correctly (e.g. `3`, `-`) and quantify the directed effect of each input token on the first-order attribution of another token. Meta-Shapley values reveal that the pure effect of `-5` is small, and the attribution value is mostly determined by the interaction `- → -5`. Meta-AttnLRP explains AttnLRP's opposite-signed attributions directionally: a minus contributes positively to the attribution of `-5` when `-5` is present `- → -5` (`+5.38`); conversely, removing `-5` weakens the negative signal of the minus `-5 → -` (`-1.95`), thereby offsetting the first-order attribution of `-` (`-4.94`). For details, see Appendix A.

attribution of another. We thus propose a general *meta*-approach that addresses these limitations by decomposing an arbitrary first-order attribution into directional interaction terms.

This work: The METAGAME, Meta-Shapley value, and meta-attributions. We formally conceptualize the proposed framework as the METAGAME, bridging the discrete principles of cooperative game theory with the continuous dynamics of modern interpretability methods. Rather than explaining the model directly, we compute a modified Shapley value targeting the outputs of state-of-the-art attribution methods, such as Grad-ECLIP (Zhao et al., 2024), AttnLRP (Achtibat et al., 2024), or ConceptAttention (Helbling et al., 2025). Crucially, by measuring the marginal contribution of one feature/token strictly in the *presence* of another, we extract *directional* meta-attributions, revealing exactly how one feature/token influences the attribution of another. Such a directional focus resolves the interaction leakage observed in serial methods, yielding a rigorous *hierarchical decomposition* of any first-order attribution method. Theoretically, we prove that these meta-attributions elegantly split classical set-based interactions into their precise, directional components.

Contributions. Our work advances literature in multiple ways: **(1)** We deliver a principled and effective approach to generalize *any* first-order attribution method to quantify second-order interactions via the Shapley value. Meta-attributions unite removal-based perturbations with gradient- and attention-based mechanisms to explain transformers. **(2)** We propose a *hierarchical* decomposition of attributions via *directional* interactions, which provides a practical way to navigate their complexity. We show that, surprisingly, Shapley interactions and integrated Hessians implicitly do this already (Theorem 1). **(3)** The METAGAME and meta-attributions define a novel *directional* interaction index, and we prove this serves as a directional extension of existing interaction indices (Theorem 2). **(4)** Finally, we demonstrate that meta-attributions offer pragmatic insights when applied to interpret large language models, vision–language encoders, and multimodal diffusion transformers.

2 Theoretical Background

We consider a general feature/token attribution method $\phi_i(f)$ that quantifies the influence of feature $i \in [d] := \{1, \ldots, d\}$ on a model predictor $f : \mathbb{R}^d \to \mathbb{R}$. While broadly applicable, we predominantly focus on local explanations at a fixed input $x \in \mathbb{R}^d$, denoted as $\phi_i(f, x)$. To evaluate f on subsets, we define the masked predictor $f(S; x)$, e.g. via baseline imputation $f(S; x) := f(x_S, b_{\bar{S}})$, which

retains features $S \subseteq [d]$ and replaces the remainder $\bar{S} := [d] \setminus S$ with a baseline $b \in \mathbb{R}^d$. Analogously, we overload the attribution notation and write $\phi_i(S; f, x) := \phi_i\big(f, (x_S, b_{\bar{S}})\big)$ for the attribution of feature i evaluated at the input x with $\bar{S}$ masked by the baseline b, so that $\phi_i([d]; f, x) = \phi_i(f, x)$.

2.1 Interaction Quantification

To interpret a model f, first-order attribution methods allocate the prediction score to individual features/tokens. A foundational game-theoretic approach for it is the Shapley value (SV, Shapley, 1953), which computes the average marginal contribution of player i across all possible coalitions:

$$\phi_i^{\text{SV}}(f, x) := \sum_{S \subseteq [d] \setminus \{i\}} \frac{1}{d \cdot \binom{d-1}{|S|}} \left[f(S \cup \{i\}; x) - f(S; x) \right].$$

In neural networks, input gradients enable attribution. Gradient×input (G×I) scales the gradient by the input, while layer-wise relevance propagation (LRP, Bach et al., 2015) applies custom back-propagation rules, linked to it (Ancona et al., 2018). To satisfy axioms such as efficiency, integrated gradients (IG, Sundararajan et al., 2017) accumulates gradients along a straight-line path:

$$\phi_i^{\text{IG}}(f, x) := (x_i - b_i) \int_0^1 \frac{\partial f(b + \alpha(x - b))}{\partial x_i} \, d\alpha.$$

Serial attributions construct interactions by recursively applying a first-order method to its own output: $\psi_{i,j}(f, x) := \phi_j\big(\phi_i(f, \cdot), x\big)$. Notable examples include integrated Hessians ψ^{IH} (IH, Janizek et al., 2021) and the serial Shapley value ψ^{SV} (Lundstrom and Razaviyayn, 2023). However, serial methods inherently fail to cleanly separate pure individual effects from interaction effects.

Example: Interaction separation. Table 1 illustrates for $f(x) = x_1 + x_1 x_2^2$ how a faithful method must isolate the individual effect of x_1 from the interaction $x_1 x_2^2$. However, both the serial Shapley value and integrated Hessians fail to do this, leaking interaction terms into the individual components.

The **Möbius transform** (Rota, 1964), $m_S(f, x) := \sum_{T \subseteq S} (-1)^{|S|-|T|} f(T; x)$ for $S \subseteq [d]$ is a principled measure for separating pure interactions (Bordt and von Luxburg, 2023). Due to its complexity with 2^d components, research has pivoted toward order-k set-based interaction indices.

Shapley interactions extend the Shapley value to subsets of size k. For $k = 2$, $\psi_{\{i,j\}}(f, x)$ quantifies the joint influence of features i and j beyond their individual contributions. Notable variants include n-Shapley values (Lundberg et al., 2020; Bordt and von Luxburg, 2023), an extension of the Shapley interaction index (Grabisch and Roubens, 1999), the faithful Shapley interaction index (Tsai et al., 2023), and the Shapley-Taylor interaction index (STII, Dhamdhere et al., 2020) given by

$$\psi_{\{i,j\}}^{\text{STII}}(f, x) := \sum_{S \subseteq [d] \setminus \{i,j\}} \frac{2}{d \cdot \binom{d-1}{|S|}} \sum_{T \subseteq \{i,j\}} (-1)^{|T|} f(S \cup T; x).$$

At $k = 2$, $\psi_{\{i,j\}}^{\text{STII}}$ separates pure individual effects and projects higher-order interactions onto pairs.

Sum of Powers (SOP, Lundstrom and Razaviyayn, 2023) theoretically extends integrated gradients to resolve the separation problem. For pairwise interactions, SOP evaluates a Shapley value on integrated gradients within a reduced $(d-1)$-player game where player i is fixed as present. For $i \in [d]$, it can be defined by $\psi_{i,i}^{\text{SOP}}(f, x) := \psi_{\{i\}}^{\text{SOP}}(f, x) := m_{\{i\}}(f, x)$ for individuals and for interactions with $j \neq i$ via

$$\psi_{i,j}^{\text{SOP}}(f, x) := \sum_{S \subseteq [d] \setminus \{i,j\}} \frac{1}{(d-1) \cdot \binom{d-2}{|S|}} \left[\phi_i^{\text{IG}}(S \cup \{i,j\}; f, x) - \phi_i^{\text{IG}}(S \cup \{i\}; f, x) \right],$$

where $\phi_i^{\text{IG}}(S; f, x)$ evaluates ϕ_i^{IG} restricted to the features in S. The set-based SOP is then defined using the symmetrization $\psi_{\{i,j\}}^{\text{SOP}}(f, x) := \psi_{i,j}^{\text{SOP}}(f, x) + \psi_{j,i}^{\text{SOP}}(f, x)$. In Section 3, we prove that SOP follows our *meta*-attribution paradigm.

2.2 Limitations of Existing Interaction Measures

Order-k interaction indices produce $\mathcal{O}(d^k)$ terms, becoming difficult to interpret as dimensionality or interaction order grows. By collapsing interactions into undirected subsets $\{i, j\}$, set-based approaches obscure the hierarchy of how features influence one another's attributions. Serial methods

Table 1: Analytical interactions $\varphi_{j\to i}$ for $f(x) = x_1 + I$, where $I := x_1 x_2^2$. Unlike serial approaches, meta-attributions isolate the individual effect x_1 ($\varphi_{i\to i}$) and the interaction I ($\varphi_{j\to i}$). Furthermore, meta-attributions capture the directional hierarchy and sum to the first-order attribution ϕ_i, whereas set-based interactions collapse effects into undirected subsets.

Type	Method	Individual		Interaction		First-order		Sum
		$\varphi_{1\to1}$	$\varphi_{2\to2}$	$\varphi_{2\to1}$	$\varphi_{1\to2}$	ϕ_1	ϕ_2	
Serial	SV (ψ^{SV})	$x_1 + I/4$	$I/4$	$I/4$	$I/4$	$x_1 + I/2$	$I/2$	$x_1 + I$
	IH (ψ^{IH})	$x_1 + I/9$	$4I/9$	$2I/9$	$2I/9$	$x_1 + I/3$	$2I/3$	$x_1 + I$
Set-Based	Möbius (m)	x_1	0	I		x_1	0	x_1
	Shapley (ψ^{STII})	x_1	0	I		$x_1 + I/2$	$I/2$	$x_1 + I$
	SOP (ψ^{SOP})	x_1	0	I		$x_1 + I/3$	$2I/3$	$x_1 + I$
Directional	ASI (Ψ)	—	—	$I/2$	$I/2$	—	—	—
	BSV (E^2)	$x_1 + I/2$	$x_1/2 + I/2$	$I/2$	$x_1/2 + I/2$	$x_1 + I$	$x_1 + I$	$2x_1 + 2I$
Ours	Meta-G×I	x_1	0	I	$2I$	$x_1 + I$	$2I$	$x_1 + 3I$
	Meta-IG	x_1	0	$I/3$	$2I/3$	$x_1 + I/3$	$2I/3$	$x_1 + I$
	Meta-SV	x_1	0	$I/2$	$I/2$	$x_1 + I/2$	$I/2$	$x_1 + I$

fail to separate individual effects from interactions, leaking interaction terms into diagonal components. While Shapley interactions and SOP isolate these effects (cf. Table 1), the former overlooks model internals, and the latter has been studied exclusively for integrated gradients. Methods that use input Hessians are limited by locality and architectural constraints; they rely on fine-grained patch- or token-level evaluations without a natural mechanism for grouping, vanish in standard ReLU networks (Janizek et al., 2021), and incur prohibitive computational costs (Pramanik et al., 2026).

Our METAGAME framework treats any first-order attribution method as a cooperative game to decompose its attributions into directional interactions. This preserves the directional hierarchy ($\varphi_{j\to i}$) and disentangles pure individual effects from interactions, thereby avoiding the leakage of serial methods. Orthogonal to our METAGAME construction are directional game-theoretic measures that do not decompose attribution values. Namely, Lu et al. (2023) average second-order discrete derivatives into an asymmetric Shapley interaction (ASI, Ψ), and Masoomi et al. (2022) propose a bivariate Shapley value (BSV, E^2). These are limited to removal-based attribution and thus cannot utilize internal gradients or attentions. Table 1 compares their analytical interaction allocations.

3 The Metagame of Interpretability and Meta-Attributions

We introduce *meta*-attributions $\varphi_{j\to i}$, which can be read naturally as the influence of feature j on the attribution of feature i, as a principled framework capable of decomposing *any* first-order attribution method ϕ_i (e.g., Meta-IG decomposes ϕ_i^{IG}). To compute these, we conceptualize the METAGAME, a cooperative game evaluating ϕ_i, and apply the Shapley value to extract interactions.

3.1 Hierarchical Interaction Decomposition of Attributions

Instead of examining all $\mathcal{O}(d^2)$ interactions of order 2 at once, we shift to a *hierarchical* approach.

Definition 1. *An interaction $\varphi_{j\to i}(f, x)$ is **hierarchically efficient** with respect to $\phi_i(f, x)$ if*

$$\phi_i(f, x) = \sum_{j\in[d]} \varphi_{j\to i}(f, x) = \varphi_{i\to i}(f, x) + \sum_{j\neq i} \varphi_{j\to i}(f, x).$$

Hierarchical efficiency has been an overlooked principle, and it is not specific to the Shapley value. Hausken and Mohr (2001) already decompose the Shapley value in this sense, and we prove in Section B.2 that their construction is the serial Shapley value. However, we propose that hierarchical efficiency is a desirable property of an interaction measure as it allows us to quantify second-order effects of any attribution mechanism. It ensures that the original first-order attribution is exactly preserved as a marginal sum of its pure individual effect $\varphi_{i\to i}(f, x)$ and its interactions with other

features $\varphi_{j\to i}(f,x)$. Surprisingly, while not originally framed this way, we show that several existing interaction indices already implicitly perform such hierarchical decompositions.

Theorem 1. *Existing interactions are hierarchically efficient, i.e.* $\phi_i = \sum_{j\in[d]} \varphi_{j\to i}$*: Shapley interactions decompose* ϕ^{SV} *with* $\varphi_{j\to i} := \psi_{\{i,j\}}/|\{i,j\}|$*, whereas SOP and serial methods decompose* ϕ^{IG} *and their own first-order attributions with* $\varphi_{j\to i} := \psi_{i,j}$*. Proofs are in Appendix B.*

For serial methods, hierarchical efficiency follows from the first-order efficiency; the outer attribution step (feature j) naturally decomposes the inner attribution (feature i). In set-based interactions, the factor of $1/|\{i,j\}|$ accounts for the neglected directionality ($j \to i$ vs. $i \to j$). However, these interaction measures are still symmetric and *non-directional* due to their construction.

3.2 Directional Interactions via Meta-Attributions

We introduce *directional* interactions for any first-order attribution method $\phi_i(f,x)$. One shortcoming of serial methods, whether via Shapley masking or by integrating Hessians, is that both features x_i and x_j are varied jointly. This joint variation inextricably entangles individual and pairwise effects. Instead, we keep x_i fixed and investigate the attribution $\phi_i(f,x)$ for all combinations of masked inputs, which defines the METAGAME over all remaining players $[d] \setminus \{i\}$.

Definition 2. *For a target feature* $i \in [d]$ *and an attribution method* $\phi_i(f,x)$*, the* **METAGAME** *is the* $(d-1)$*-player cooperative game* $\nu_i \colon 2^{[d]\setminus\{i\}} \to \mathbb{R}$ *with players* $[d] \setminus \{i\}$ *and a game function*

$$\nu_i(S) \coloneqq \phi_i(S \cup \{i\}; f, x),$$

i.e. the attribution of i *when features in* $S \cup \{i\}$ *take their values from* x*, while others are masked.*

Within the METAGAME, we then utilize the Shapley value of ν_i to define directional interactions.

Definition 3. *The* ***meta-attribution*** $\varphi_{j\to i}(f,x)$*, representing the influence of feature* $j \neq i$ *on the attribution* $\phi_i(f,x)$ *of feature* i*, is defined as the Shapley value of player* j *in the* METAGAME ν_i*:*

$$\varphi_{j\to i}(f,x) := \sum_{S\subseteq[d]\setminus\{i,j\}} \frac{1}{(d-1)\cdot\binom{d-2}{|S|}} \big[\nu_i(S\cup\{j\}) - \nu_i(S)\big].$$

The pure individual effect is the worth of the empty coalition, $\varphi_{i\to i}(f,x) \coloneqq \nu_i(\emptyset) = \phi_i(\{i\}; f, x)$.

Corollary 1. *Meta-attributions are hierarchically efficient, i.e.* $\phi_i(f,x) = \sum_{j\in[d]} \varphi_{j\to i}(f,x)$.

This follows from the efficiency of the Shapley value, since the grand coalition of the METAGAME is worth $\nu_i([d] \setminus \{i\}) = \phi_i(f,x)$ and the empty coalition is worth $\varphi_{i\to i}(f,x)$. Hence, although we use the Shapley value as the efficient allocation rule that relates meta-attributions to existing interaction indices, METAGAME is extensible to other efficient allocation methods as a core principle.

When the underlying first-order attribution is itself the Shapley value (i.e., $\phi_i := \phi_i^{\text{SV}}$), we refer to the resulting meta-attribution $\varphi_{j\to i}^{\text{Meta-SV}}$ as the **Meta-Shapley value**. Similarly, ϕ_i^{IG} or $\phi_i^{\text{G}\times\text{I}}$ yields $\varphi_{j\to i}^{\text{Meta-IG}}$ and $\varphi_{j\to i}^{\text{Meta-G}\times\text{I}}$, respectively. We relate these methods to set-based interactions in Theorem 2, which establishes the Meta-Shapley value and Meta-IG as *directional* variants of STII and SOP, respectively. Crucially, the METAGAME framework extends this principled directional decomposition to arbitrary first-order attribution methods, e.g. those that rely on gradients and attention in transformers.

Theorem 2. *The Meta-Shapley value and Meta-IG act as directional variants of the STII and SOP:*

$$\psi_{\{i\}}^{\textit{STII}}(f,x) = \varphi_{i\to i}^{\textit{Meta-SV}}(f,x), \qquad \psi_{\{i,j\}}^{\textit{STII}}(f,x) = \varphi_{j\to i}^{\textit{Meta-SV}}(f,x) + \varphi_{i\to j}^{\textit{Meta-SV}}(f,x),$$
$$\psi_{\{i\}}^{\textit{SOP}}(f,x) = \varphi_{i\to i}^{\textit{Meta-IG}}(f,x), \qquad \psi_{\{i,j\}}^{\textit{SOP}}(f,x) = \varphi_{j\to i}^{\textit{Meta-IG}}(f,x) + \varphi_{i\to j}^{\textit{Meta-IG}}(f,x).$$

Example: Interaction separation. Continuing with the example $f(x) = x_1 + x_1x_2^2$ (Table 1), we find that directional meta-attributions isolate the pure individual effect x_1 into the diagonal $\varphi_{1\to1}$ and correctly yield $\varphi_{2\to2} = 0$. The interaction term $I := x_1x_2^2$ is strictly restricted to the off-diagonal terms ($\varphi_{2\to1}$ and $\varphi_{1\to2}$), completely avoiding the spurious leakage seen in serial methods. The exact allocation of this interaction depends on the underlying first-order method. Notably, our framework exposes the inherent *directionality* of feature interactions: while the Meta-SV yields symmetric interactions by design ($\varphi_{2\to1} \equiv \varphi_{1\to2}$), gradient-based methods like Meta-IG and Meta-G×I capture asymmetric, directional influences ($\varphi_{2\to1} \neq \varphi_{1\to2}$). Regardless of symmetry, directional variants separate individual effects from interaction effects and sum to their first-order attributions.

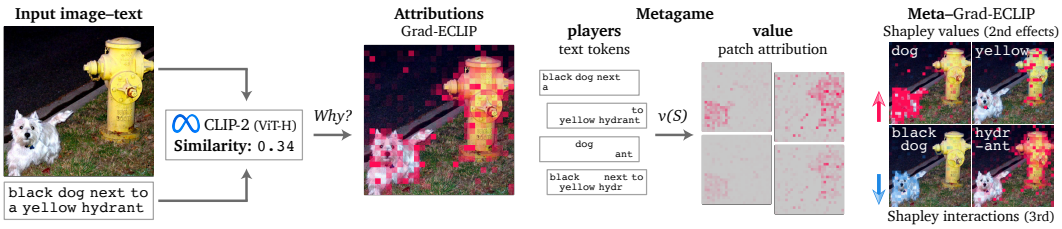


Figure 2: METAGAME quantifies gradient-based token interactions in vision–language encoders. Given a token attribution method (Grad-ECLIP) and a dual encoder (MetaCLIP-2), we compute meta-attributions from text token subsets and their corresponding visual patch attributions. First-order attributions quantify the effects that text tokens `dog` and `yellow` have on the similarity map (red, most similar). Directional meta-attributions quantify the interaction effects contributed by pairs of text tokens, such as `hydr-ant` and `black-dog` (blue, most dissimilar). See also App. Figure 8.

3.3 Considerations for Applying the METAGAME Principles to Interpret Transformers

While our theoretical considerations focus on the foundational attribution methods, the general concepts and intuitions behind METAGAME are also relevant to the modern interpretability approaches for transformers. We effectively demonstrate this through three settings of increasing abstractness.

Meta-AttnLRP. AttnLRP (Achtibat et al., 2024) is a backpropagation-based, attention-aware method to attribute single input tokens into transformer predictions. While it has shown state-of-the-art faithfulness performance in language and vision tasks, it is unable to capture token–token synergies. Meta-AttnLRP complements its interpretation with directional second-order interactions by computing a Shapley value of the exact METAGAME defined as: players $:=$ input tokens, game function $:=$ AttnLRP (cf. Definition 2). Specifically for language models, we simply aggregate token attributions and interactions across the generated multi-token output sequence as detailed in Appendix C.1.

Meta-Grad-ECLIP. Grad-ECLIP (Zhao et al., 2024) is a gradient-based method to measure the influence of image and text tokens on the embeddings of vision–language encoders. Similarly to AttnLRP, it cannot quantify the synergy between the image and text modalities. Our proposed Meta-Grad-ECLIP quantifies directional cross-modal interactions by computing Shapley values of text tokens in the attribution of image tokens. METAGAME thus becomes: players $:=$ text tokens, game function $:=$ Grad-ECLIP of image tokens (cf. Figure 2). Here, the target i is an image patch rather than a text token, so the METAGAME has text tokens as players, and the game function $\nu_i(S)$ becomes the attribution of patch i when text tokens outside of S are masked. Extending this idea, one can also compute a bi-token Shapley interaction of the METAGAME, which can be treated as measuring synergy between a *token pair* and the image, e.g. `black-dog` and `hydr-ant` in Figure 2. Finally, METAGAME could also be implemented the other way around to quantify the second-order effects of image patches on the attribution of text tokens, which we do not investigate in this work. As a minor contribution of this paper, we extend the attention- and gradient-based methods originally proposed for CLIP to the modern SigLIP-2 architecture (Tschannen et al., 2025) as detailed in Appendix C.2.

Meta-ConceptAttention. ConceptAttention is not a standard attribution method; it allows attributing concepts (text tokens) to image patches generated by text-to-image diffusion transformers. However, it is limited by producing highly context-dependent interpretations, as discussed in Section 5.5 and Figure 15 of Helbling et al. (2025). We aim to alleviate context dependence and interpret the second-order effects of concepts. As the most abstract application of our framework, departing from Definition 2, Meta-ConceptAttention computes a Shapley value of the METAGAME defined as: players $:=$ concepts, game function $:=$ ConceptAttention. Interestingly, this application also illustrates a case where including the target concept i as a player can be useful, since the Shapley value of i then averages its attribution across all contexts of the other concepts. Crucially, our implementation requires only a single model forward pass, just like ConceptAttention, and computes marginal contributions using the model's cached embeddings, making it viably efficient. We provide further methodological and implementational details in Appendix C.3.

Scaling the computation. For small METAGAMES ($d < 20$), we compute the Shapley value exactly, jointly obtaining $\varphi_{j\to i}$ for all targets i in one vectorized pass. For larger ones, we recommend SOTA approximators from the `shapiq` library (Muschalik et al., 2024): Monte Carlo sampling amortizes

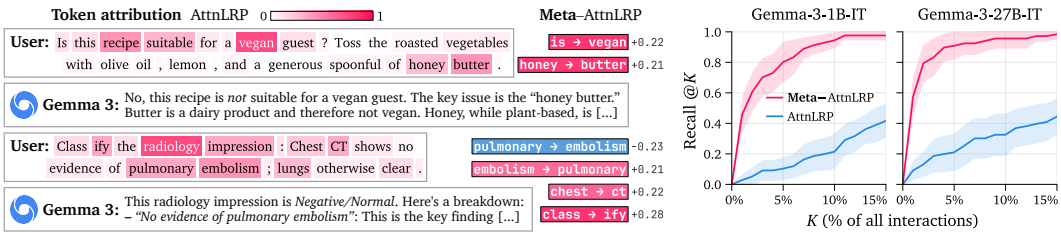


Figure 3: METAGAME quantifies token interactions in instruction-tuned large language models. We compute Meta-AttnLRP as Shapley values from text tokens into AttnLRP token attributions of the Gemma language model's generated output, highlighting directional second-order effects. We also measure the recall of detecting human-labeled interactions (e.g. word connotations, negation) on a sample of prompts spanning various contexts. For results on other models, see App. Figure 10.

well with moderate d and many targets, while regression-based approximators (Witter et al., 2025) better focus the budget when d is large but only specific targets matter, at the cost of one fit per target.

4 Applications and Experimental Results

Here, we demonstrate that the METAGAME offers useful insights: (i) when quantifying token interactions in language models, (ii) when explaining similarity in vision–language encoders, and (iii) when interpreting concepts in multimodal diffusion transformers. See Appendix D for details regarding the experimental setup, implementations, and hyperparameters.

4.1 Quantifying Token Interactions in Language Models

We interpret the Gemma 3 language model (Gemma Team, 2025) via AttnLRP (Achtibat et al., 2024), although in principle METAGAME is general and any token attribution method can be used instead. Figure 3 illustrates two examples of how important each token in the user prompt is to the model's generated response, averaged over up to 100 output tokens. In the first case, the first-order attributions rank `recipe`, `vegan`, and `butter` as the most influential, while our proposed Meta-AttnLRP complements the interpretation with interactions between `is`→`vegan` and `honey`→`butter` as key to the model's comprehension. In another example, AttnLRP emphasizes `radiology`, `embolism`, and `ct`, while Meta-AttnLRP highlights the importance of `class`→`ify` and `chest`→`ct`. Crucially, we detect that the model pays more attention to `pulmonary` when `embolism` is present, whereas removing `pulmonary` would increase the importance of `embolism` in the generated output. We report the increasing rate at which both approaches are able to capture such concepts using various human-labeled prompts, each containing 2–4 interacting token pairs, in Figure 3 (right). We provide additional examples where METAGAME complements AttnLRP attributions in Appendix E.

4.2 Explaining Similarity in Vision–Language Encoders

We demonstrate the broader applicability of METAGAME to improve the explainability of vision–language encoders such as CLIP (Radford et al., 2021), SigLIP-2 (Tschannen et al., 2025), and MetaCLIP-2 (Chuang et al., 2025). We compare three popular saliency map methods for CLIP, namely generic attention (Chefer et al., 2021), MaskCLIP (Zhou et al., 2022), and Grad-ECLIP (Zhao et al., 2024), with their meta-attribution counterparts. We benchmark the performance of these algorithms quantitatively using the pointing interaction recognition metric proposed in (FIxLIP, Baniecki et al., 2025b), with a random method scoring 0.25. It combines four labeled images into a grid and scores the share of interaction mass that an explanation places on the correct pairs, i.e. between each object word and the patches of the image showing that object. Figure 4 shows an example of evaluating Meta-Grad-ECLIP interactions explaining MetaCLIP-2 on the `fish-koala-balloon-laptop` pointing game. Table 2 reports the state-of-the-art performance of meta-attributions in consistently improving the interaction recognition on ImageNet-1k across various models and methods. Meta-Attention outperforms Meta-Grad-ECLIP on the overall largest and best-performing MetaCLIP-2 model. Interestingly, our proposed Meta-MaskSigLIP approach is overall the most faithful in interpreting SigLIP-2, a result confirmed across two model patch sizes (cf. App. Table 4). Meta-attributions



Figure 4: Example of Meta-Grad-ECLIP explaining the MetaCLIP-2 encoder on a single input from a single pointing game, which is aggregated to measure the interaction recognition evaluation metric.

Table 2: METAGAME lifts first-order explanations of vision–language encoders to second order. On ImageNet-1k, meta-attributions correctly recover the cross-modal interactions that first-order methods miss across models of different architectures and sizes. The second-order baselines **in gray** take much longer to compute. Extended results for SigLIP-2 are reported in Appendix Table 4.

Model (Size)	Explanation Method	Interaction Recognition (↑)			
		1 object	2 objects	3 objects	4 objects
CLIP (ViT-B/16)	Attention (Chefer et al., 2021)	$.77_{\pm.01}$	$.46_{\pm.01}$	$.32_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-Attention	$.84_{\pm.01}$	$.87_{\pm.01}$	$\mathbf{.89_{\pm.01}}$	$\mathbf{.90_{\pm.01}}$
	MaskCLIP (Zhou et al., 2022)	$.30_{\pm.01}$	$.28_{\pm.01}$	$.27_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-MaskCLIP	$\mathbf{.86_{\pm.01}}$	$\mathbf{.88_{\pm.01}}$	$\mathbf{.89_{\pm.01}}$	$.89_{\pm.01}$
	Grad-ECLIP (Zhao et al., 2024)	$.71_{\pm.01}$	$.43_{\pm.01}$	$.31_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-Grad-ECLIP	$.76_{\pm.01}$	$.79_{\pm.01}$	$.82_{\pm.01}$	$.84_{\pm.01}$
	FIxLIP (Baniecki et al., 2025b)	$.76_{\pm.01}$	$.77_{\pm.01}$	$.78_{\pm.01}$	$.79_{\pm.01}$
	exCLIP (Moeller et al., 2025)	$.71_{\pm.01}$	$.80_{\pm.01}$	$.84_{\pm.01}$	$.86_{\pm.01}$
MetaCLIP-2 (ViT-H/14)	Attention (Chefer et al., 2021)	$.84_{\pm.01}$	$.45_{\pm.01}$	$.32_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-Attention	$\mathbf{.85_{\pm.01}}$	$\mathbf{.83_{\pm.01}}$	$\mathbf{.85_{\pm.01}}$	$\mathbf{.86_{\pm.01}}$
	Grad-ECLIP (Zhao et al., 2024)	$.76_{\pm.01}$	$.43_{\pm.01}$	$.31_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-Grad-ECLIP	$.79_{\pm.01}$	$.76_{\pm.01}$	$.78_{\pm.01}$	$.80_{\pm.01}$
	FIxLIP (Baniecki et al., 2025b)	$.70_{\pm.01}$	$.66_{\pm.01}$	$.73_{\pm.01}$	$.75_{\pm.01}$
	exCLIP (Moeller et al., 2025)	$.83_{\pm.01}$	$.81_{\pm.01}$	$.83_{\pm.01}$	$.86_{\pm.01}$

often outperform the second-order baselines that use only the removal-based mechanism of Shapley interactions or only the backpropagation-based mechanism of gradients in this task. Beyond faithfully attributing bi-token interactions, computing Shapley interactions (instead of Shapley values) for the METAGAME of these attributions quantifies the tri-token influence on the model's prediction, as illustrated in Appendix Figures 8 & 9, inspecting which we leave for future work.

4.3 Interpreting Concepts in Multimodal Diffusion Transformers

Finally, we show the potential of METAGAME to improve the interpretability of text-to-image diffusion transformers. Meta-ConceptAttention provides two valuable utilities in this task: the Shapley value of ConceptAttention improves on its main limitation—context-dependence, while the *directional* meta-attribution explains the influence between pairs of concepts (see Figure 5). Following Helbling et al. (2025), we benchmark Meta-ConceptAttention on a proxy segmentation performance of the FLUX.1 [schnell] text-to-image model (Black Forest Labs, 2024) on the Pascal VOC dataset (Everingham et al., 2015), including single-object and multiclass images. We extend the evaluation to the more challenging MS COCO benchmark (Lin et al., 2014). Contrary to the original experimental setup, we evaluate segmentation performance when additional concepts are included in the context. Table 3 shows that Meta-ConceptAttention Pareto-dominates the original ConceptAttention across all metrics

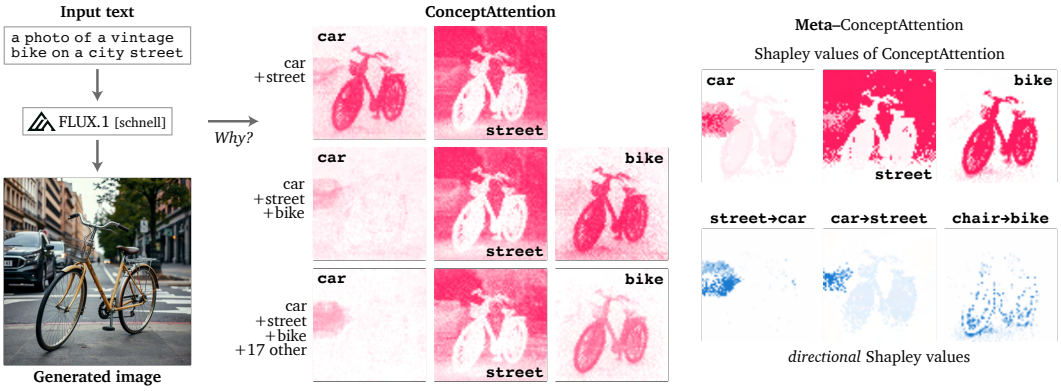


Figure 5: METAGAME quantifies concept interactions in multimodal diffusion transformers. Shapley values average attention across concept subsets, and interpret their directional dependencies.

Table 3: METAGAME alleviates the context dependence in the ConceptAttention interpretability of the FLUX.1 [schnell] text-to-image model across Pascal VOC and MS COCO datasets. Crucially, both methods use only a single model forward pass. Several ablations are reported in App. Figure 13.

Method	**Pascal VOC** (Acc/mIoU/mAP ↑)		**MS COCO** (Acc/mIoU/mAP ↑)	
	Single concept	Multiple concepts	Single concept	Multiple concepts
ConceptAttention	.832/.645/.878	.612/.476/.622	.852/.656/.899	.572/.350/.423
↳ METAGAME	**.903/.768/.908**	**.803/.553/.663**	**.889/.720/.909**	**.706/.390/.439**

for both the single and multiple concept tasks, when simultaneously evaluating the method on all classes apparent in Pascal VOC and the same amount for MS COCO. We further perform extensive ablations on the number of added concepts, including the artificial text prompt, not using cross-concept attention, and restricting the set of layers, as reported in Appendix Figure 13. We find that our Meta-Shapley-based interpretation correctly denoises attention maps, i.e. its performance stays the same or even increases as the number of additional concepts in-context increases. Notably, if we did not use cross-concept attention, the performance gap would be even larger.

5 Related Work

While METAGAME shares a naming affinity with the meta-evaluation problem (Hedström et al., 2023), which quantifies attribution quality, our framework focuses on second-order interaction effects.

Backpropagation-based attribution. Several gradient-based methods capture higher-order dynamics implicitly by aggregating attributions in a local neighborhood around the input, e.g. SmoothGrad (Smilkov et al., 2017; Zhou et al., 2025), expected gradients (Erion et al., 2021; Baniecki et al., 2025a), or through model randomization tests (Adebayo et al., 2018; Binder et al., 2023). Sikdar et al. (2021) utilize integrated directional gradients to explicitly quantify group interactions. Eberle et al. (2022) and Schnake et al. (2022) extend layer-wise relevance propagation to quantify interactions via higher-order Taylor decomposition in deep similarity networks and graph neural networks, respectively. More recently, Schnake et al. (2025) restrict relevance propagation to paths connecting features within a given subset to explain a prediction with logical formulas over features, which yields set-based, i.e. undirected, interactions. Specifically in data attribution, Wang et al. (2025) leverage gradient-Hessian-gradient products to capture interactions between training samples.

Removal-based attribution. Game-theoretic attribution is continuously evolving to capture complex dependencies, from asymmetric Shapley values for causal priors (Frye et al., 2020), to n-Shapley values recovering a functional decomposition of the model (Bordt and von Luxburg, 2023), and unified frameworks for higher-order interactions (Fumagalli et al., 2025). Other advances include discrete representations of mixed partial derivatives (Tsang et al., 2020) and mitigations for non-convex games (Chang et al., 2025). However, METAGAME shifts the paradigm: rather than evaluating

the model output directly, it treats the first-order attribution itself as the value function of a cooperative game to extract directional interactions. Finally, concurrent efforts by Kang et al. (2025) and Butler et al. (2025) focus on scaling token interactions in large language models via Fourier expansions.

Applications. Despite ongoing skepticism regarding the specific properties of some feature attribution methods (Bilodeau et al., 2024; Böhle et al., 2024), they remain crucial for analyzing large vision models (Achtibat et al., 2023; Fel et al., 2023), e.g. in medical imaging (Chrabaszcz et al., 2025). Komorowski et al. (2026) leveraged AttnLRP to guide decoding and improve instruction-following in language models like Gemma 3. Explicit second-order methods, such as integrated Jacobians, successfully explain CLIP (Moeller et al., 2025) and LLMs (Pramanik et al., 2026). Similarly, Shapley and Banzhaf interactions provide actionable insights in video-language training (Jin et al., 2025), hyperparameter optimization (Wever et al., 2026), and survival analysis (Langbein et al., 2026).

Attention. The debate over whether attention faithfully explains transformers remains central, particularly regarding how to aggregate these values into an explanation while accounting for other network components (Wiegreffe and Pinter, 2019; Bibal et al., 2022). Growing evidence highlights its limitations across vision (Komorowski et al., 2023), language (Achtibat et al., 2024), and multimodal domains (Zhao et al., 2024). Nonetheless, attention remains a practical tool for locating concepts in diffusion transformers (Helbling et al., 2025; Erel et al., 2025). To overcome the inherent limitations of standard transformer architectures (Dziri et al., 2023; Lin et al., 2025), third-order attention between input tokens is increasingly explored (Sanford et al., 2023; Kozachinskiy et al., 2025).

6 Conclusion

In this work, we introduced the METAGAME and meta-attributions as a mathematically principled interpretability framework that complements existing methods for quantifying second-order interactions. We demonstrated its versatility and impact across three areas of AI: analyzing token interactions in instruction-tuned language models, explaining vision-language encoders, and interpreting concepts in multimodal diffusion transformers. Ultimately, the foundational constructs behind METAGAME offer a promising path toward more credible machine learning systems through future empirical research.

Limitations and future work. Clearly, meta-attributions apply to first-order attributions that can be evaluated on masked inputs. Our theory is limited by assuming baseline masking/imputation, the most principled approach to removing tokens/features from machine learning model inputs. Future work may extend the METAGAME to other removal schemes and higher-order interactions as we elaborate in Appendix F. We ablate the masking strategy for CLIP meta-attributions in Table 5, finding that the results are robust to this choice in our setup. In the vision–language and diffusion experiments, we compute Shapley values for meta-attributions *exactly*, at a computational cost comparable to that of the first-order baselines. For longer language model prompts, we rely on coalition sampling, so METAGAME will only benefit from future improvements in Shapley value estimation. We further acknowledge that our language model analysis is a systematic qualitative showcase of a few prompts; rigorous method comparisons would also require faithfulness measurements (Jumelet and Zuidema, 2023), such as token-pair insertion/deletion curves (Baniecki et al., 2025b). Finally, quantifying higher-order effects *inherently* increases both computational cost and cognitive burden for the user.

Broader impact. We do not anticipate any negative societal impacts from this work, as it focuses on foundational interpretability techniques intended to improve our understanding of transformers.

Acknowledgments and Disclosure of Funding

Hubert Baniecki was supported by the Foundation for Polish Science (FNP), and the state budget within the Polish Ministry of Education and Science program "Pearls of Science" project number PN/01/0087/2022. Fabian Fumagalli gratefully acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/3 2026 – 438445824. Work on this project is financially supported by the Foundation for Polish Science (FNP) grant 'Centre for Credible AI' No. FENG.02.01-IP.05-0058/24, and we gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within the computational grant no. PLG/2026/019528. Finally, we appreciate Max for walking his dogs and sharing their pictures with us, allowing us to test our approach in the wild.

References

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. 5

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *ICML*, 2024. 1, 3.3, 4.1, 5, C.1

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018. 5

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR*, 2018. 2.1

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 10(7):e0130140, 2015. 2.1

Hubert Baniecki, Giuseppe Casalicchio, Bernd Bischl, and Przemyslaw Biecek. Efficient and accurate explanation estimation with distribution compression. In *ICLR*, 2025a. 5, F

Hubert Baniecki, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, Eyke Hüllermeier, and Przemyslaw Biecek. Explaining similarity in vision-language encoders with weighted Banzhaf interactions. In *NeurIPS*, 2025b. 4.2, 2, 6, D.2, 4, F

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation? An introduction to the debate. In *ACL*, 2022. 5

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024. 5

Alexander Binder, Leander Weber, Sebastian Lapuschkin, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. In *CVPR*, 2023. 5

Black Forest Labs. FLUX.1 [schnell]. *github.com/black-forest-labs/flux Version 1.0*, 2024. 4.3, D.3

Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos alignment for inherently interpretable CNNs and vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4504–4518, 2024. 5, D.2

Sebastian Bordt and Ulrike von Luxburg. From Shapley values to generalized additive models and back. In *AISTATS*, 2023. 2.1, 5, B.1.2, B.3, B.5

Landon Butler, Abhineet Agarwal, Justin Singh Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. ProxySPEX: Inference-efficient interpretability via sparse feature interactions in LLMs. In *NeurIPS*, 2025. 5

Wonjoon Chang, Myeongjin Lee, and Jaesik Choi. Rethinking shapley value for negative interactions in non-convex games. In *ICLR*, 2025. 5

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, 2021. 4.2, 2, C.2

Maciej Chrabaszcz, Hubert Baniecki, Piotr Komorowski, Szymon Plotka, and Przemyslaw Biecek. Aggregated attributions for explanatory analysis of 3D segmentation models. In *WACV*, 2025. 5

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta CLIP 2: A Worldwide Scaling Recipe. In *NeurIPS*, 2025. 4.2, D.2

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *NeurIPS*, 2020. F

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. C.3

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. D.2

Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. The Shapley Taylor interaction index. In *ICML*, 2020. 2.1, B.1.2, B.3, B.5

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *NeurIPS*, 2019. 1

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and Fate: Limits of transformers on compositionality. In *NeurIPS*, 2023. 5

Oliver Eberle, Jochen Büttner, Florian Kräutli, Klaus-Robert Müller, Matteo Valleriani, and Grégoire Montavon. Building and interpreting deep similarity models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1149–1161, 2022. 5

Yotam Erel, Olaf Dünkel, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Amit H. Bermano. Attention (as discrete-time markov) chains. In *NeurIPS*, 2025. 5

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7):620–631, 2021. 5

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 4.3, D.3

Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *NeurIPS*, 2023. 1, 5

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. In *NeurIPS*, 2020. 5

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order Shapley interactions. In *NeurIPS*, 2023. 1, F

Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, Barbara Hammer, and Julia Herbinger. Unifying feature-based explanations with functional ANOVA and cooperative game theory. In *AISTATS*, 2025. 5

Gemma Team. Gemma 3 technical report. *arXiv preprint, arXiv:2503.19786*, 2025. 4.1, D.1

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999. 2.1

Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000. B.5

Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014. D.3

Kjell Hausken and Matthias Mohr. The value of a player in n-person games. *Social Choice and Welfare*, 18(3):465–483, 2001. 3.1, B.2, B.2

Anna Hedström, Philine Lou Bommer, Kristoffer Knutsen Wickstrøm, Wojciech Samek, Sebastian Lapuschkin, and Marina MC Höhne. The meta-evaluation problem in explainable AI: Identifying reliable estimators with MetaQuantus. *Transactions on Machine Learning Research*, 2023. 5

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion transformers learn highly interpretable features. In *ICML*, 2025. 1, 3.3, 4.3, 5, C.3, D.3

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021. 1, 2.1, 2.2

Peng Jin, Hao Li, Li Yuan, Shuicheng Yan, and Jie Chen. Hierarchical Banzhaf interaction for general video-language representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 5

Jaap Jumelet and Willem Zuidema. Feature interactions reveal linguistic structure in language models. In *Findings of the ACL*, 2023. 6

Justin Singh Kang, Landon Butler, Abhineet Agarwal, Yigit Efe Erginbas, Ramtin Pedarsani, Kannan Ramchandran, and Bin Yu. SPEX: Scaling feature interaction explanations for LLMs. In *ICML*, 2025. 5

Piotr Komorowski, Hubert Baniecki, and Przemysław Biecek. Towards evaluating explanations of vision transformers for medical imaging. In *CVPRW*, 2023. 5

Piotr Komorowski, Elena Golimblevskaia, Reduan Achtibat, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attribution-guided decoding. In *ICLR*, 2026. 1, 5, C.1

Alexander Kozachinskiy, Felipe Urrutia, Hector Ignacio Jimenez Orellana, Tomasz Steifer, Germán Pizarro, Matías Daniel Fuentes Álvarez, Francisco Meza Vásquez, Cristian Buc Calderon, and Cristobal Rojas. Strassen attention, split VC dimension and compositionality in transformers. In *NeurIPS*, 2025. 5

Sophie Hanna Langbein, Hubert Baniecki, Fabian Fumagalli, Niklas Koenen, Marvin N. Wright, and Julia Herbinger. Functional decomposition and Shapley interactions for interpreting survival models. In *ICML*, 2026. 5

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. ZebraLogic: On the scaling limits of LLMs for logical reasoning. In *ICML*, 2025. 5

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 4.3, D.3

Xiaolei Lu, Jianghong Ma, and Haode Zhang. Asymmetric feature interaction for interpreting model predictions. In *Findings of the ACL*, 2023. 2.2, B.1.3

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017. 1

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. 2.1, B.1.2, B.3

Daniel Lundstrom and Meisam Razaviyayn. A unifying framework to the analysis of interaction methods using synergy functions. In *ICML*, 2023. 1, 2.1, B.5

Aria Masoomi, Davin Hill, Zhonghui Xu, Craig P. Hersh, Edwin K. Silverman, Peter J. Castaldi, Stratis Ioannidis, and Jennifer G. Dy. Explanations of black-box models based on directional feature interactions. In *ICLR*, 2022. 2.2, B.1.3

Lucas Moeller, Pascal Tilli, Ngoc Thang Vu, and Sebastian Padó. Explaining caption-image interactions in CLIP models with second-order attributions. *Transactions on Machine Learning Research*, 2025. 2, 5, 4

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley interactions for machine learning. In *NeurIPS*, 2024. 3.3, C.1

Vishal Pramanik, Maisha Maliha, Nathaniel D. Bastian, and Sumit Kumar Jha. Hessian-enhanced token attribution (HETA): Interpreting autoregressive LLMs. In *ICLR*, 2026. 1, 2.2, 5

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4.2, C.2, D.2

Gian-Carlo Rota. On the foundations of combinatorial theory: I. theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964. 2.1

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *NeurIPS*, 2023. 5

Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schütt, Klaus-Robert Müller, and Grégoire Montavon. Higher-order explanations of graph neural networks via relevant walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7581–7596, 2022. 5

Thomas Schnake, Farnoush Rezaei Jafari, Jonas Lederer, Ping Xiong, Shinichi Nakajima, Stefan Gugler, Grégoire Montavon, and Klaus-Robert Müller. Towards symbolic XAI – explanation through human understandable logical relationships between features. *Information Fusion*, 118: 102923, 2025. 5

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games*, pages 307–318. Princeton University Press, 1953. 2.1

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural NLP models. In *ACL*, 2021. 5

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv preprint, arXiv:1706.03825*, 2017. 5

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *ICML*, 2020. F

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 1, 2.1

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024. 1

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-Shap: The faithful Shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023. 2.1, B.1.2, B.3

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? Interpretable attribution for feature interactions. In *NeurIPS*, 2020. 5

Michael Tschannen et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 3.3, 4.2, C.2, D.2

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. In *ICLR*, 2025. 1, 5

Marcel Wever, Maximilian Muschalik, Fabian Fumagalli, and Marius Lindauer. HyperSHAP: Shapley values and interactions for explaining hyperparameter optimization. In *AAAI*, 2026. 5

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP*, 2019. 5

R. Teal Witter, Yurong Liu, and Christopher Musco. Regression-adjusted Monte Carlo estimators for Shapley values and probabilistic values. In *NeurIPS*, 2025. 3.3

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024. C.3

Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and Antoni B. Chan. Gradient-based visual explanation for transformer-based CLIP. In *ICML*, 2024. 1, 3.3, 4.2, 2, 5, C.2

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *ECCV*, 2022. 4.2, 2, C.2, D.2

Linjiang Zhou, Chao Ma, Zepeng Wang, Libing Wu, and Xiaochuan Shi. AdaptGrad: Adaptive sampling to reduce noise. In *NeurIPS*, 2025. 5

Appendix "The Metagame of Interpretability and Meta-Attributions"

A Illustrative Example **17**

B Proofs **19**

B.1 Derivations for Table 1 19

B.2 Decomposition by Hausken & Mohr is the Serial Shapley Value 22

B.3 Proof of Theorem 1 23

B.4 Proof of Corollary 1 24

B.5 Proof of Theorem 2 25

C Extended Methods **26**

C.1 AttnLRP and Meta-AttnLRP for Language Generation 26

C.2 Adapting Generic Attention, Grad-ECLIP, and MaskCLIP to the SigLIP-2 Architecture 27

C.3 Implementing Meta-ConceptAttention in a Single Forward Pass 28

D Experimental Setup Details **29**

D.1 Quantifying Token Interactions in Language Models 29

D.2 Explaining Similarity in Vision–Language Encoders 30

D.3 Interpreting Concepts in Multimodal Diffusion Transformers 30

D.4 Compute Resources 31

E Additional Experimental Results **32**

F Potential Extensions of the METAGAME and Meta-Attributions **36**

A Illustrative Example

We aim to cross-compare complementary behaviors of various token attribution approaches on a simple illustrative example. For that, we train a single-layer, single-head transformer with 760 parameters to solve integer addition in the form of `a {+,-} b = y`, where inputs `a` and `b` are integers from $[-9, 9]$. The model uses RMSNorm, GELU, a learned absolute positional embedding, and is trained without biases. It achieves a near-zero mean squared error on both the in-distribution validation set and a held-out test set comprising the unseen `(a,b)` pairs. For an intuitive case, we train the model to recognize the masking token `<unk>` and regularize it to predict 0 for a fully masked input sequence comprising three `<unk>` tokens (the so-called empty set, giving baseline value). Ultimately, the model predicts the expected output value over a distribution of input tokens in the `<unk>` position. Our demo is easy to run on a CPU and fully explorable with the code attached to the paper. Figure 6 extends Figure 1 and, together with Figure 7, presents four representative input examples that build intuition for how Shapley values and interactions behave, how they relate to gradient-based approximations, how AttnLRP's interpretation differs, and how our METAGAME approach decomposes the first-order attributions. We compute AttnLRP attributions by summing the relevance assigned to the token embedding and positional embedding at each input position.

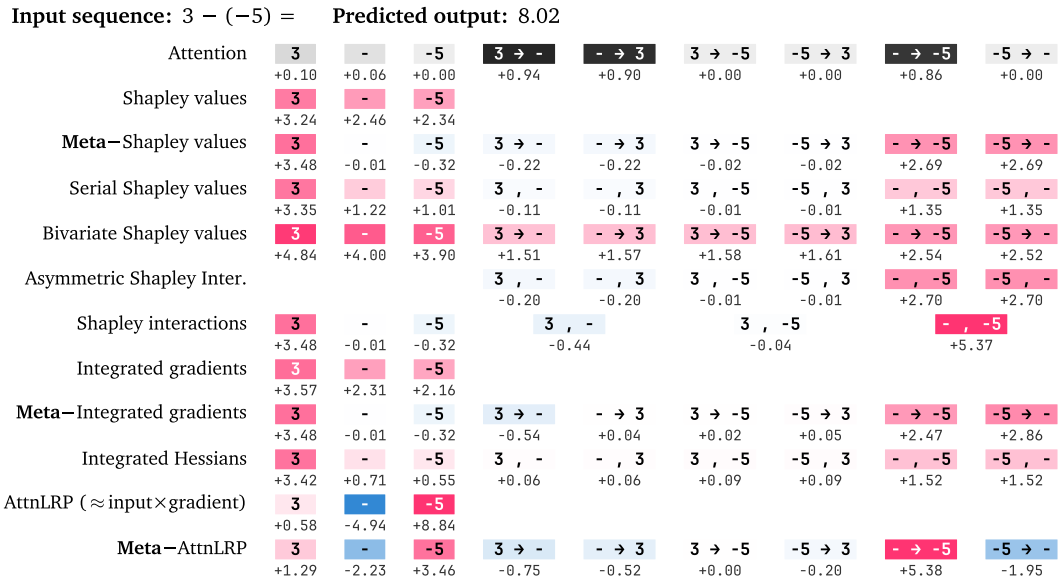


Figure 6: Complementary interpretations of a simple transformer solving integer addition. Extended Figure 1. Attention highlights the token interactions `3 → -`, `- → 3`, and `- → -5` without the notion of sign. Shapley values, integrated gradients, and AttnLRP attribute the prediction to individual tokens but cannot capture interactions between token pairs. Shapley interactions and integrated Hessians highlight the undirected interaction `- , -5`. Furthermore, Shapley interactions confirm that the pure effects of `-` and `-5` are negligible, so Shapley values effectively split this interaction between `-` and `-5`. AttnLRP, by contrast, assigns opposite-signed local attributions to `-` and `-5` asymmetrically. Our proposed Meta-Shapley, Meta-Integrated gradients, and Meta-AttnLRP quantify the directed effect of each input token on the first-order attribution of another token. Meta-Shapley values are symmetric and correspond to half of the Shapley interactions. Meta-AttnLRP explains AttnLRP's opposite-signed attributions directionally: a minus contributes positively to the attribution of `-5` when `-5` is present `- → -5` (`+5.38`); conversely, removing `-5` weakens the negative signal of the minus `-5 → -` (`-1.95`), offsetting the first-order attribution of `-` (`-4.94`).

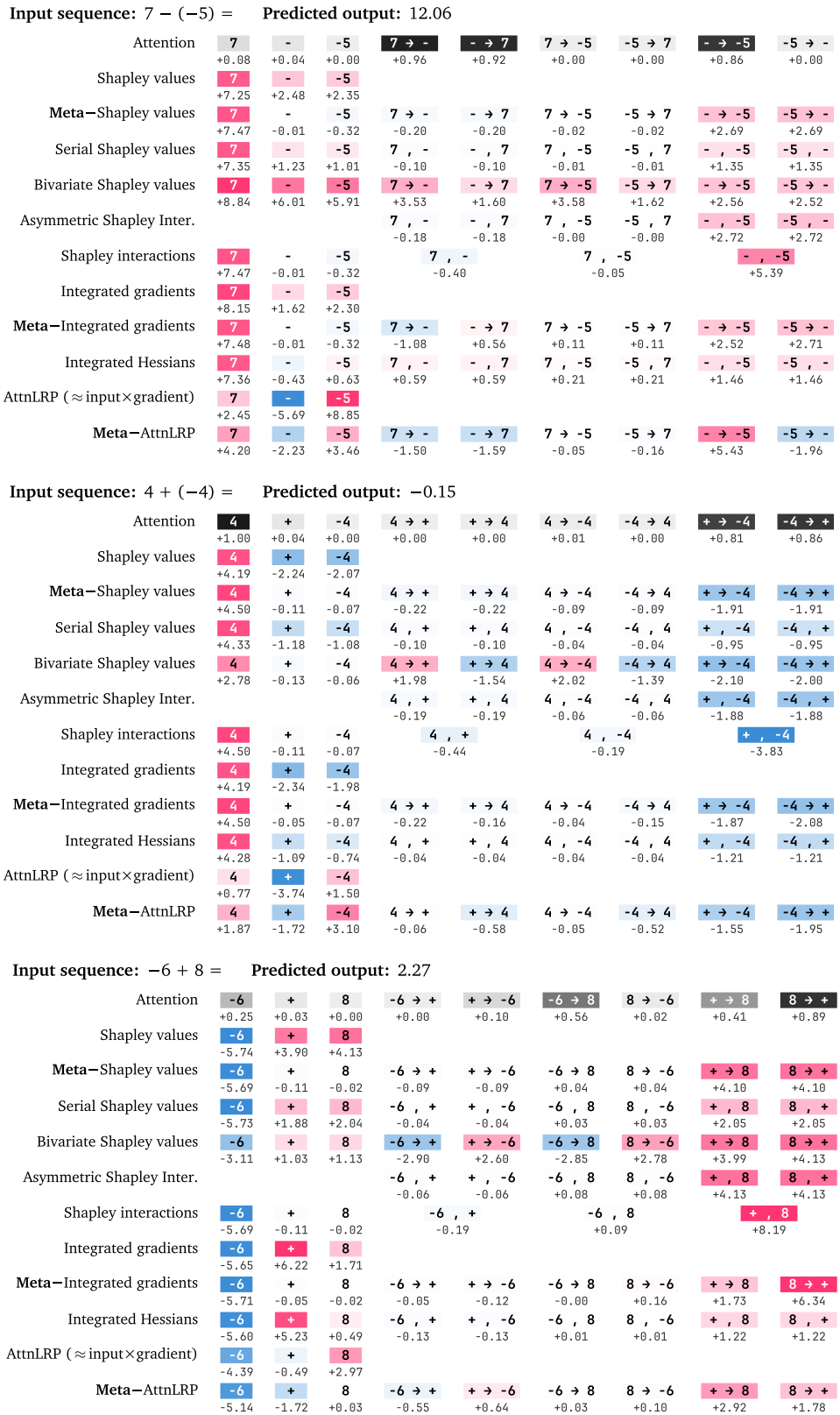


Figure 7: Complementary interpretations of a simple transformer solving integer addition. Additional examples similar to Figure 6.

B Proofs

B.1 Derivations for Table 1

Here, we provide the step-by-step derivations for the analytical components presented in Table 1. Let the model be $f(x) = x_1 + x_1x_2^2$, with the interaction term defined as $I \coloneqq x_1x_2^2$. We assume a standard zero baseline $b = (0, 0)$. We omit f in $\phi(f, x)$, $\varphi(f, x)$ etc. for conciseness.

Before computing the second-order components, we first establish the underlying first-order attributions $\phi_i(x)$ for gradient×input, integrated gradients, and Shapley values.

Gradient×input (G×I).

$$\phi_1^{\text{G}\times\text{I}}(x) = x_1 \frac{\partial f(x)}{\partial x_1} = x_1(1 + x_2^2) = x_1 + I$$
$$\phi_2^{\text{G}\times\text{I}}(x) = x_2 \frac{\partial f(x)}{\partial x_2} = x_2(2x_1x_2) = 2I$$

Integrated gradients (IG).

$$\phi_1^{\text{IG}}(x) = x_1 \int_0^1 \frac{\partial f(\alpha x)}{\partial x_1} d\alpha = x_1 \int_0^1 (1 + \alpha^2 x_2^2) d\alpha = x_1 \left(1 + \frac{1}{3}x_2^2\right) = x_1 + \frac{1}{3}I$$
$$\phi_2^{\text{IG}}(x) = x_2 \int_0^1 \frac{\partial f(\alpha x)}{\partial x_2} d\alpha = x_2 \int_0^1 (2\alpha^2 x_1x_2) d\alpha = \frac{2}{3}x_1x_2^2 = \frac{2}{3}I$$

Shapley values (SV). The characteristic function is $v(S) = f(S; x)$. Thus $v(\emptyset) = 0$, $v(\{1\}) = x_1$, $v(\{2\}) = 0$, and $v(\{1, 2\}) = x_1 + I$.

$$\phi_1^{\text{SV}}(x) = \frac{1}{2}(v(\{1\}) - v(\emptyset)) + \frac{1}{2}(v(\{1, 2\}) - v(\{2\})) = \frac{1}{2}(x_1) + \frac{1}{2}(x_1 + I) = x_1 + \frac{1}{2}I$$
$$\phi_2^{\text{SV}}(x) = \frac{1}{2}(v(\{2\}) - v(\emptyset)) + \frac{1}{2}(v(\{1, 2\}) - v(\{1\})) = 0 + \frac{1}{2}(x_1 + I - x_1) = \frac{1}{2}I$$

B.1.1 Serial Methods

Serial methods compute interactions by treating the first-order attribution $\phi_i(x)$ as a new target function $h_i(x)$ and reapplying the attribution method.

Serial Shapley value. We apply Shapley value to $h_i(x) \coloneqq \phi_i^{\text{SV}}(x)$. For h_1, the coalitional values are $v_{h_1}(\emptyset) = 0, v_{h_1}(\{1\}) = x_1, v_{h_1}(\{2\}) = 0, v_{h_1}(\{1, 2\}) = x_1 + \frac{1}{2}I$.

$$\psi_{1,1}^{\text{SV}}(x) = \frac{1}{2}(x_1 - 0) + \frac{1}{2}\left(x_1 + \frac{1}{2}I - 0\right) = x_1 + \frac{1}{4}I$$
$$\psi_{1,2}^{\text{SV}}(x) = \frac{1}{2}(0 - 0) + \frac{1}{2}\left(x_1 + \frac{1}{2}I - x_1\right) = \frac{1}{4}I$$

For h_2, $v_{h_2}(\emptyset) = v_{h_2}(\{1\}) = v_{h_2}(\{2\}) = 0$, and $v_{h_2}(\{1, 2\}) = \frac{1}{2}I$.

$$\psi_{2,1}^{\text{SV}}(x) = \frac{1}{2}(0 - 0) + \frac{1}{2}\left(\frac{1}{2}I - 0\right) = \frac{1}{4}I$$
$$\psi_{2,2}^{\text{SV}}(x) = \frac{1}{2}(0 - 0) + \frac{1}{2}\left(\frac{1}{2}I - 0\right) = \frac{1}{4}I$$

Integrated Hessians (IH). We apply integrated gradients to $h_1(x) \coloneqq \phi_1^{\text{IG}}(x)$ and $h_2(x) \coloneqq \phi_2^{\text{IG}}(x)$.

$$\psi_{1,1}^{\text{IH}}(x) = x_1 \int_0^1 \frac{\partial h_1(\alpha x)}{\partial x_1} d\alpha = x_1 \int_0^1 \left(1 + \frac{1}{3}\alpha^2 x_2^2\right) d\alpha = x_1 + \frac{1}{9}I$$
$$\psi_{1,2}^{\text{IH}}(x) = x_2 \int_0^1 \frac{\partial h_1(\alpha x)}{\partial x_2} d\alpha = x_2 \int_0^1 \left(\frac{2}{3}\alpha^2 x_1 x_2\right) d\alpha = \frac{2}{9}I$$
$$\psi_{2,1}^{\text{IH}}(x) = x_1 \int_0^1 \frac{\partial h_2(\alpha x)}{\partial x_1} d\alpha = x_1 \int_0^1 \left(\frac{2}{3}\alpha^2 x_2^2\right) d\alpha = \frac{2}{9}I$$
$$\psi_{2,2}^{\text{IH}}(x) = x_2 \int_0^1 \frac{\partial h_2(\alpha x)}{\partial x_2} d\alpha = x_2 \int_0^1 \left(\frac{4}{3}\alpha^2 x_1 x_2\right) d\alpha = \frac{4}{9}I$$

B.1.2 Interaction Methods

We now derive the results for the Möbius and Shapley interactions as well as for SOP.

Möbius and Shapley interactions. The Möbius coefficients for the characteristic function $v(S) = f(S; x)$ are:

$$m_{\{1\}}(x) = v(\{1\}) - v(\emptyset) = x_1$$
$$m_{\{2\}}(x) = v(\{2\}) - v(\emptyset) = 0$$
$$m_{\{1,2\}}(x) = v(\{1,2\}) - v(\{1\}) - v(\{2\}) + v(\emptyset) = (x_1 + I) - x_1 - 0 + 0 = I$$

The Möbius method maps these directly to the individual and interaction components. Shapley interaction indices of order $k = 2$ map these Möbius coefficients directly to the individual elements (x_1 and 0) and the joint pair (I). This can be verified for popular Shapley interactions, such as n-Shapley values (Lundberg et al., 2020; Bordt and von Luxburg, 2023), Shapley-Taylor interaction index (Dhamdhere et al., 2020), and faithful Shapley interaction index (Tsai et al., 2023), cf. (Bordt and von Luxburg, 2023, Proof of Theorem 8).

Sum of Powers (SOP). As established in Theorem 2, SOP corresponds to the symmetric set-based interaction derived from Meta-IG. We can therefore bypass a redundant derivation and simply compute it by summing the corresponding directional Meta-IG components (derived below):

$$\psi_{\{1\}}^{\text{SOP}}(x) = \varphi_{1\to1}^{\text{IG}}(x) = x_1$$
$$\psi_{\{2\}}^{\text{SOP}}(x) = \varphi_{2\to2}^{\text{IG}}(x) = 0$$
$$\psi_{\{1,2\}}^{\text{SOP}}(x) = \varphi_{1\to2}^{\text{IG}}(x) + \varphi_{2\to1}^{\text{IG}}(x) = \frac{2}{3}I + \frac{1}{3}I = I.$$

B.1.3 Directional Interactions

We derive the two directional baselines of Table 1.

Asymmetric Shapley interaction (Ψ). Lu et al. (2023) average the second-order discrete derivative $\delta_{ij}v(S) \coloneqq v(S\cup\{i,j\}) - v(S\cup\{i\}) - v(S\cup\{j\}) + v(S)$ over $S \subseteq [d]\setminus\{i,j\}$, and split it between the two directions. At $d = 2$ the only coalition is $S = \emptyset$, so

$$\delta_{12}v(\emptyset) = (x_1 + I) - x_1 - 0 + 0 = I, \qquad \text{hence} \qquad \Psi_{2\to1} = \Psi_{1\to2} = I/2.$$

The index defines no individual term, so it does not decompose ϕ_i.

Bivariate Shapley (E^2). Masoomi et al. (2022) gate the game on the presence of the target, $u_i(S) \coloneqq v(S)\mathbb{1}[i \in S]$, and allocate it with an outer Shapley value over all d players, $E^2_{j\to i} \coloneqq \phi_j^{\text{SV}}(u_i)$. Here $u_1 = v$, since $v(\{2\}) = 0$ already, whereas u_2 vanishes except at $\{1,2\}$:

$$E^2_{1\to1} = \tfrac{1}{2}\big(x_1 - 0\big) + \tfrac{1}{2}\big(x_1 + I - 0\big) = x_1 + I/2, \quad E^2_{2\to1} = \tfrac{1}{2}\big(0 - 0\big) + \tfrac{1}{2}\big(x_1 + I - x_1\big) = I/2,$$
$$E^2_{2\to2} = \tfrac{1}{2}\big(0 - 0\big) + \tfrac{1}{2}\big(x_1 + I - 0\big) = x_1/2 + I/2, \qquad E^2_{1\to2} = \tfrac{1}{2}\big(0 - 0\big) + \tfrac{1}{2}\big(x_1 + I - 0\big) = x_1/2 + I/2.$$

Each row sums to $v([d]) = x_1 + I$ rather than to the first-order attribution ϕ_i, because the gated game allocates the prediction, not an attribution. Half of the individual effect x_1 lands off the diagonal, where it remains even at $I = 0$.

B.1.4 Meta-Attributions

Directional meta-attributions strictly separate the individual effect $\varphi_{i\to i}$ from interaction effects $\varphi_{j\to i}$ via marginal evaluations on the METAGAME with x_i fixed. For a 2-dimensional feature space, this naturally reduces to $\varphi_{i\to i}(x) = \phi_i(x_i, 0)$ and $\varphi_{j\to i}(x) = \phi_i(x) - \phi_i(x_i, 0)$.

Meta-G×I.

$$\begin{aligned}
\varphi^{\text{G}\times\text{I}}_{1\to 1}(x) &= \phi^{\text{G}\times\text{I}}_1(x_1, 0) = x_1 \\
\varphi^{\text{G}\times\text{I}}_{2\to 1}(x) &= \phi^{\text{G}\times\text{I}}_1(x) - \phi^{\text{G}\times\text{I}}_1(x_1, 0) = (x_1 + I) - x_1 = I \\
\varphi^{\text{G}\times\text{I}}_{2\to 2}(x) &= \phi^{\text{G}\times\text{I}}_2(0, x_2) = 0 \\
\varphi^{\text{G}\times\text{I}}_{1\to 2}(x) &= \phi^{\text{G}\times\text{I}}_2(x) - \phi^{\text{G}\times\text{I}}_2(0, x_2) = 2I - 0 = 2I
\end{aligned}$$

Meta-IG.

$$\begin{aligned}
\varphi^{\text{IG}}_{1\to 1}(x) &= \phi^{\text{IG}}_1(x_1, 0) = x_1 \\
\varphi^{\text{IG}}_{2\to 1}(x) &= \phi^{\text{IG}}_1(x) - \phi^{\text{IG}}_1(x_1, 0) = \left(x_1 + \frac{1}{3}I\right) - x_1 = \frac{1}{3}I \\
\varphi^{\text{IG}}_{2\to 2}(x) &= \phi^{\text{IG}}_2(0, x_2) = 0 \\
\varphi^{\text{IG}}_{1\to 2}(x) &= \phi^{\text{IG}}_2(x) - \phi^{\text{IG}}_2(0, x_2) = \frac{2}{3}I - 0 = \frac{2}{3}I
\end{aligned}$$

Meta-SV.

$$\begin{aligned}
\varphi^{\text{SV}}_{1\to 1}(x) &= \phi^{\text{SV}}_1(x_1, 0) = x_1 \\
\varphi^{\text{SV}}_{2\to 1}(x) &= \phi^{\text{SV}}_1(x) - \phi^{\text{SV}}_1(x_1, 0) = \left(x_1 + \frac{1}{2}I\right) - x_1 = \frac{1}{2}I \\
\varphi^{\text{SV}}_{2\to 2}(x) &= \phi^{\text{SV}}_2(0, x_2) = 0 \\
\varphi^{\text{SV}}_{1\to 2}(x) &= \phi^{\text{SV}}_2(x) - \phi^{\text{SV}}_2(0, x_2) = \frac{1}{2}I - 0 = \frac{1}{2}I
\end{aligned}$$

This completes the derivations of Table 1.

B.2 Decomposition by Hausken & Mohr is the Serial Shapley Value

For a game $v\colon 2^{[d]} \to \mathbb{R}$ and $S \subseteq [d]$, let $v|_S\colon 2^S \to \mathbb{R}$ be its restriction to the player set S, i.e. $v|_S(R) = v(R)$ for $R \subseteq S$. Hausken and Mohr (2001) define the value of player j to player i as the Shapley value of j in the game recording the payout of i inside each coalition:

$$\Phi_{j\to i}(v) \coloneqq \phi_j^{\mathrm{SV}}(g_i), \qquad \text{where} \quad g_i(S) \coloneqq \begin{cases} \phi_i^{\mathrm{SV}}(v|_S) & i \in S, \\ 0 & i \notin S. \end{cases}$$

For local interpretability, $v(S) = f(S; x)$, so g_i evaluates the first-order attribution ϕ_i^{SV} on coalitions. The following proposition shows that this construction is equivalent to the serial Shapley value.

Proposition 1. *$g_i(S) = \phi_i^{SV}(S; f, x)$ for every $S \subseteq [d]$, and therefore $\Phi_{j\to i}(v) = \psi_{i,j}^{SV}(f, x)$.*

Proof. Masking via a subset S makes the predictor evaluate $v(R \cap S)$ on a coalition R, so $\phi_i^{\mathrm{SV}}(S; f, x)$ is the Shapley value of i in the d-player game $R \mapsto v(R \cap S)$, in which every $k \notin S$ is a null player. If $i \notin S$, then i is null as well and both sides are 0. If $i \in S$, write the Shapley value as an average over orders π of $[d]$ with predecessor sets P_i^π:

$$\phi_i^{\mathrm{SV}}(S; f, x) = \mathbb{E}_\pi\big[v\big((P_i^\pi \cap S) \cup \{i\}\big) - v(P_i^\pi \cap S)\big] = \phi_i^{\mathrm{SV}}(v|_S) = g_i(S),$$

because the summand depends on π only through $P_i^\pi \cap S$, and restricting a uniform order of $[d]$ to S gives a uniform order of S. Both constructions then apply ϕ_j^{SV} to the same game, and $\psi_{i,j}^{\mathrm{SV}}(f, x) = \phi_j^{\mathrm{SV}}\big(S \mapsto \phi_i^{\mathrm{SV}}(S; f, x)\big)$ by definition. □

As a result, Theorem 2.1 of Hausken and Mohr (2001), i.e. the rows of their value matrix sum to the Shapley value, is hierarchical efficiency (Theorem 1) for the serial Shapley value. Moreover, since g_i is defined on coalitions without i, the outer Shapley value also varies x_i, which causes the leakage of interaction effects into individual terms shown in the Serial SV row of Table 1.

B.3 Proof of Theorem 1

Proof. We prove the result for each method separately. For Shapley interactions, we consider the three main interaction indices, namely n-Shapley values (Lundberg et al., 2020; Bordt and von Luxburg, 2023), the STII (Dhamdhere et al., 2020), and the faithful Shapley interaction index (Tsai et al., 2023).

Serial Shapley value and integrated Hessians. We let $\varphi_{j\to i}(f,x) := \psi_{i,j}(f,x)$. For the serial Shapley value and integrated Hessians, hierarchical efficiency follows immediately from the standard efficiency axioms of their underlying first-order methods. The efficiency axiom guarantees that the sum of attributions equals the difference between the grand coalition (i.e. $\phi_i(f,x)$) and the empty coalition (i.e. 0 due to the dummy axiom). When applying the attribution method serially (e.g., applying the Shapley value to the output of the Shapley value, or integrated gradients to the output of integrated gradients), the outer attribution naturally distributes the inner first-order attribution among all features. Consequently, summing the second-order terms

$$\sum_j \varphi_{j\to i} = \sum_j \psi_{i,j} = \sum_j \phi_j(\phi_i(f,\cdot),x) = \phi_i(f,x) - \phi_i(\{\};f,x) = \phi_i(f,x)$$

exactly recovers the first-order attribution ϕ_i, demonstrating that integrated Hessians decompose integrated gradients and serial Shapley value decomposes Shapley value.

Sum of Powers (SOP). We let $\varphi_{j\to i}(f,x) \coloneqq \psi^{\text{SOP}}_{i,j}(f,x)$, which equals $\varphi^{\text{Meta-IG}}_{j\to i}(f,x)$ for all $j \in [d]$ by Theorem 2 and the definition of SOP. Hierarchical efficiency with respect to ϕ^{IG} then follows from Corollary 1.

n-Shapley Values. For n-SVs (Lundberg et al., 2020; Bordt and von Luxburg, 2023), hierarchical efficiency is a direct consequence of their recursive definition and the relationship between n-SVs of different orders. According to Proposition 13 of the functional decomposition framework (Bordt and von Luxburg, 2023), lower-order Shapley values can be computed from higher-order ones via a lump-sum formula. Specifically, the standard Shapley value $\phi^{\text{SV}}_i(f,x)$ decomposes into the n-SVs as:

$$\phi^{\text{SV}}_i(f,x) = \psi^{n\text{-SV}}_{\{i\}}(f,x) + \frac{1}{2}\sum_{j\neq i}\psi^{n\text{-SV}}_{\{i,j\}}(f,x) + \cdots + \frac{1}{n}\sum_{\substack{K\subseteq[d]\setminus\{i\}\\|K|=n-1}}\psi^{n\text{-SV}}_{K\cup\{i\}}(f,x)$$

Restricting this to pairwise interactions ($n=2$), we obtain

$$\phi^{\text{SV}}_i(f,x) = \psi^{2\text{-SV}}_{\{i\}}(f,x) + \sum_{j\neq i}\frac{1}{2}\psi^{2\text{-SV}}_{\{i,j\}}(f,x).$$

By substituting $\varphi_{j\to i}(f,x) \coloneqq {}^{\psi^{2\text{-SV}}_{\{i,j\}}(f,x)}/_2$ and the pure individual effect $\varphi_{i\to i}(f,x) \coloneqq \psi^{2\text{-SV}}_{\{i\}}(f,x)$, we see that the 2-Shapley interactions perfectly decompose the original Shapley value.

Faithful Shapley Interaction Index (FSII). To establish this for the Faithful Shapley interaction index, we argue through their Möbius representation (Tsai et al., 2023, Theorem 19), which reads for $k=2$ and $S\subseteq[d]$ with $|S|\leq 2$ as

$$\psi^{\text{FSII}}_S(f,x) = m_S(f,x) + (-1)^{2-|S|}\frac{|S|}{2+|S|}\binom{2}{|S|}\sum_{T\supset S:|T|>2}\frac{\binom{|T|-1}{2}}{\binom{|T|+1}{|S|+2}}m_T(f,x).$$

For the individual effect ($j=i$, so $|S|=1$), the directional scaling yields $\varphi_{i\to i}(f,x) = \psi^{\text{FSII}}_{\{i\}}(f,x)$. By expanding the binomials $\frac{\binom{|T|-1}{2}}{\binom{|T|+1}{3}} = \frac{3(|T|-2)}{|T|(|T|+1)}$ and applying the coefficient $-\frac{1}{3}\binom{2}{1} = -\frac{2}{3}$, we obtain:

$$\varphi_{i\to i}(f,x) = m_{\{i\}}(f,x) - \sum_{T\supset\{i\}:|T|>2}\frac{2(|T|-2)}{|T|(|T|+1)}m_T(f,x).$$

For the off-diagonal interaction ($j\neq i$, so $|S|=2$), the directional scaling is $\varphi_{j\to i}(f,x) \coloneqq {}^{\psi^{\text{FSII}}_{\{i,j\}}(f,x)}/_2$. Expanding $\frac{\binom{|T|-1}{2}}{\binom{|T|+1}{4}} = \frac{12}{|T|(|T|+1)}$ and applying the base coefficient $\frac{2}{4}\binom{2}{2} = \frac{1}{2}$ yields a

total coefficient of $\frac{6}{|T|(|T|+1)}$ for $\psi^{\text{FSII}}_{\{i,j\}}$. Halving this for the directional attribution gives:

$$\varphi_{j\to i}(f,x) = \frac{1}{2}m_{\{i,j\}}(f,x) + \sum_{T\supset\{i,j\}:|T|>2} \frac{3}{|T|(|T|+1)} m_T(f,x).$$

To verify hierarchical efficiency, we sum over all $j \neq i$. Since there are precisely $|T|-1$ features $j \neq i$ in any given subset $T \supset \{i\}$, the sum accumulates as:

$$\sum_{j\neq i} \varphi_{j\to i}(f,x) = \sum_{T\supset\{i\}:|T|=2} \frac{1}{2} m_T(f,x) + \sum_{T\supset\{i\}:|T|>2} \frac{3(|T|-1)}{|T|(|T|+1)} m_T(f,x).$$

Finally, adding the individual term $\varphi_{i\to i}(f,x)$ merges the terms for subsets where $|T|>2$:

$$\frac{3(|T|-1)}{|T|(|T|+1)} - \frac{2(|T|-2)}{|T|(|T|+1)} = \frac{3|T|-3-2|T|+4}{|T|(|T|+1)} = \frac{|T|+1}{|T|(|T|+1)} = \frac{1}{|T|}.$$

This elegantly collapses the entire summation to $\sum_{T\supseteq\{i\}} \frac{1}{|T|} m_T(f,x)$, which exactly recovers the standard Möbius representation of the first-order Shapley value $\phi^{\text{SV}}_i(f,x)$, confirming the decomposition.

Shapley-Taylor Interaction Index (STII). We again argue through their Möbius representations, which is given for $k=2$ and $i \neq j$ (Bordt and von Luxburg, 2023, Proof of Theorem 8) by

$$\psi^{\text{STII}}_{\{i,j\}}(f,x) = \sum_{S\subseteq[d]:i,j\in S} \frac{1}{\binom{|S|}{2}} m_S(f,x).$$

For the pure individual effect, the STII strictly isolates the Möbius coefficient of the singleton, yielding $\psi^{\text{STII}}_{\{i\}}(f,x) = m_{\{i\}}(f,x)$. Using the definition $\varphi_{j\to i} := {}^{\psi_{\{i,j\}}}/_{|\{i,j\}|}$, we verify hierarchical efficiency by summing these directional attributions over all $j \in [d]$:

$$\sum_{j\in[d]} \varphi_{j\to i}(f,x) = m_{\{i\}}(f,x) + \sum_{j\neq i} \sum_{S\subseteq[d]:i,j\in S} \frac{m_S(f,x)}{|S|(|S|-1)}.$$

By swapping the order of summation on the right-hand term, we note that for any fixed subset S containing i, there are exactly $|S|-1$ features $j \neq i$ present in S. The sum over j thus counts the term $|S|-1$ times, yielding:

$$\begin{aligned}
\sum_{j\in[d]} \varphi_{j\to i}(f,x) &= m_{\{i\}}(f,x) + \sum_{S\subseteq[d]:i\in S,|S|\geq 2} \frac{(|S|-1)\, m_S(f,x)}{|S|(|S|-1)} \\
&= m_{\{i\}}(f,x) + \sum_{S\subseteq[d]:i\in S,|S|\geq 2} \frac{m_S(f,x)}{|S|} \\
&= \sum_{S\subseteq[d]:i\in S} \frac{m_S(f,x)}{|S|}.
\end{aligned}$$

This final expression exactly recovers the standard Möbius representation of the first-order Shapley value $\phi^{\text{SV}}_i(f,x)$, proving that the scaled STII hierarchically decomposes the Shapley value. □

B.4 Proof of Corollary 1

Proof. By definition, for $j \neq i$, the directional meta-attribution $\varphi_{j\to i}(f,x)$ is computed as the Shapley value of feature j in a $(d-1)$-player subgame over the feature set $[d]\setminus\{i\}$. The characteristic function of this subgame is $v^{(i)}(K) := \phi_i(K\cup\{i\}; f, x)$ for $K \subseteq [d]\setminus\{i\}$.

By the efficiency axiom of the Shapley value, the sum of the attributions across all $d-1$ players equals the difference between the value of the grand coalition and the empty coalition in this subgame:

$$\begin{aligned}
\sum_{j\neq i} \varphi_{j\to i}(f,x) &= v^{(i)}([d]\setminus\{i\}) - v^{(i)}(\emptyset) \\
&= \phi_i([d]; f, x) - \phi_i(\{i\}; f, x) \\
&= \phi_i(f,x) - \varphi_{i\to i}(f,x).
\end{aligned}$$

Rearranging the terms and incorporating the individual term $\varphi_{i\to i}(f,x)$ directly yields:

$$\sum_{j\in[d]} \varphi_{j\to i}(f,x) = \phi_i(f,x).$$

This demonstrates that directional meta-attributions inherently satisfy hierarchical efficiency. □

B.5 Proof of Theorem 2

Proof. We demonstrate the equivalence of the directional meta-attributions to the symmetric interaction indices as stated in the theorem.

First, the equivalence between Meta-IG and the SOP interaction follows directly from their respective definitions. Note that our definition is the special case of SOP for $k = 2$ and follows (Lundstrom and Razaviyayn, 2023, Eq. 43 and Eq. 44). By construction, the SOP pairwise interaction evaluates a modified Shapley value on integrated gradients within a reduced $(d-1)$-player game, which algebraically matches our definition of the Meta-IG and its symmetrization:

$$\psi^{\text{SOP}}_{\{i\}}(f,x) = \varphi^{\text{Meta-IG}}_{i\to i}(f,x), \quad \text{and} \quad \psi^{\text{SOP}}_{\{i,j\}}(f,x) = \varphi^{\text{Meta-IG}}_{j\to i}(f,x) + \varphi^{\text{Meta-IG}}_{i\to j}(f,x).$$

Second, to establish the connection between Meta-SV and the STII, we can argue through their Möbius representations. The first-order Shapley value ϕ_i^{SV} can be expressed via the Möbius coefficients $m_S(f,x)$ of the original game f as $\phi_i^{\text{SV}}(f,x) = \sum_{S\subseteq[d]:i\in S} \frac{m_S(f,x)}{|S|}$ (Grabisch et al., 2000).

When evaluating the directional marginal contribution of feature $j \neq i$ on ϕ_i^{SV}, we treat the attribution evaluated at a fixed x_i as a new METAGAME $\nu(K)$ for $K \subseteq [d]\setminus\{i\}$. By evaluating ϕ_i^{SV} restricted to the subset $K\cup\{i\}$ and expressing it via its Möbius representation, we obtain:

$$\nu(K) \coloneqq \phi_i^{\text{SV}}(K\cup\{i\}; f, x) = \sum_{S\subseteq K\cup\{i\}:i\in S} \frac{m_S(f,x)}{|S|} = \sum_{T\subseteq K} \frac{m_{T\cup\{i\}}(f,x)}{|T\cup\{i\}|}.$$

By definition, the unique Möbius representation of the game ν dictates that $\nu(K) \equiv \sum_{T\subseteq K} m_T(\nu)$. Comparing this definition directly with our derived equation above, we conclude that this is already the Möbius representation of ν, allowing us to immediately identify the coefficients:

$$m_T(\nu) = \frac{m_{T\cup\{i\}}(f,x)}{|T\cup\{i\}|}.$$

Applying the standard Möbius formulation of the Shapley value to evaluate feature j in this subgame and substituting $S = T\cup\{i\}$, we obtain:

$$\varphi^{\text{Meta-SV}}_{j\to i}(f,x) = \sum_{S\subseteq[d]:i,j\in S} \frac{m_S(f,x)}{|S|(|S|-1)}.$$

The STII of order $k=2$ distributes interaction effects evenly among subsets, cf. (Dhamdhere et al., 2020, Proposition 4) or (Bordt and von Luxburg, 2023, Proof of Theorem 8). Its representation is $\psi^{\text{STII}}_{\{i\}}(f,x) = m_{\{i\}}(f,x)$ for individual effects, and for pairs $j\neq i$:

$$\psi^{\text{STII}}_{\{i,j\}}(f,x) = \sum_{S\subseteq[d]:i,j\in S} \frac{1}{\binom{|S|}{2}} m_S(f,x).$$

We therefore obtain for $i \neq j$

$$\varphi^{\text{Meta-SV}}_{j\to i}(f,x) = \frac{1}{2}\psi^{\text{STII}}_{\{i,j\}}(f,x),$$

and consequently the symmetric index immediately yields

$$\psi^{\text{STII}}_{\{i,j\}}(f,x) = \varphi^{\text{Meta-SV}}_{j\to i}(f,x) + \varphi^{\text{Meta-SV}}_{i\to j}(f,x),$$

completing the proof. □

C Extended Methods

C.1 AttnLRP and Meta-AttnLRP for Language Generation

Conventionally, attribution is computed w.r.t. a single-valued model output, e.g. the probability of a predicted class or regression estimate as in Fig. 1. To provide meaningful interpretations of the multi-token text generated by modern language models, we aggregate input attributions across multiple steps of generation, targeting the predicted token's logit from the vocabulary logit vector $f_t(x, z_{t-1})$ at each step $t = 1, \ldots, t_{\max}$, where z_{t-1} is the sequence of previously generated tokens: $z_0 = (), z_1 = \big(\operatorname{argmax} f_1(x)\big), z_2 = \big(\operatorname{argmax} f_1(x), \operatorname{argmax} f_2(x, z_1)\big)$, etc. We thus define

$$\phi_i^{\text{AttnLRP}} \coloneqq \frac{1}{t_{\max}} \sum_{t=1}^{t_{\max}} \phi_i^{\text{AttnLRP}(t)}(f_t, x, z_{t-1}),$$

and similarly for second-order interaction effects, we have

$$\varphi_{j \to i}^{\text{Meta-AttnLRP}} \coloneqq \frac{1}{t_{\max}} \sum_{t=1}^{t_{\max}} \varphi_{j \to i}^{\text{Meta-AttnLRP}(t)} \left(f_t, x, z_{t-1}\right),$$

where $\varphi_{j \to i}^{\text{Meta-AttnLRP}(t)}$ is computed with respect to $\phi_i^{\text{AttnLRP}(t)}$. Similarly to Komorowski et al. (2026), we restrict the attribution analysis and METAGAME players to tokens in x — the original input prompt (instruction) of fixed length. We discard the attributions for input-generated tokens z; their analysis can be left to future work. We compute AttnLRP $\phi_i^{\text{AttnLRP}(t)}$ using its official implementation (Achtibat et al., 2024) and compute the Shapley values $\varphi_{j \to i}^{\text{Meta-AttnLRP}(t)}$ with a regression-based approximator from the `shapiq` library (Muschalik et al., 2024) using a coalition budget of $32(d-1)$ (with a pairing trick) for a reasonable case.

C.2 Adapting Generic Attention, Grad-ECLIP, and MaskCLIP to the SigLIP-2 Architecture

As a minor technical contribution of this paper, we adapt the three interpretability methods, originally developed for CLIP (Radford et al., 2021), to SigLIP-2 (Tschannen et al., 2025). SigLIP-2 differs from CLIP in two ways that matter for explanation: (i) there is no `[CLS]` token, so all d vision tokens are patch tokens; (ii) the image embedding is produced by a Multi-head Attention Pooling (MAP) head, in which a learned probe cross-attends to the final encoder output rather than reading off a designated token. Each method outlined below adapts to one or both of these.

Additional notation. Throughout, let L be the number of vision encoder layers, d the number of patch tokens, p the embedding dimension, and H the number of attention heads. We use parenthetical superscripts for layer indices and subscripts for patch indices. At layer ℓ, $A^{(\ell)} \in \mathbb{R}^{H\times d\times d}$ is the post-softmax attention map and $q^{(\ell)}, k^{(\ell)} \in \mathbb{R}^{d\times p}$ are the query and key features, with $q_i^{(\ell)}, k_i^{(\ell)} \in \mathbb{R}^p$ their patch-i rows. We write $h \in \mathbb{R}^{d\times p}$ for the post-layernorm encoder output (i.e. the input to the MAP head) and $h_i \in \mathbb{R}^p$ for its patch-i row; $\tau \in \mathbb{R}^p$ for the text embedding; and c for the image–text matching score being explained. Each method below produces a per-patch attribution ϕ_i for $i = 1, \dots, d$, reshaped to a $\sqrt{d} \times \sqrt{d}$ attribution map. Finally, we use $\odot$ for the elementwise product, $(\cdot)^+ = \max(\cdot, 0)$, and $\hat{x} = x/\|x\|_2$.

Attention. Following Chefer et al. (2021), we accumulate per-layer relevance $R \in \mathbb{R}^{d\times d}$ via the generic attention update

$$R \;\leftarrow\; R + \bar{A}^{(\ell)}\, R, \qquad \bar{A}^{(\ell)} \;=\; \mathbb{E}_H\big[(\nabla_{A^{(\ell)}} c \;\odot\; A^{(\ell)})^+\big],$$

where $\mathbb{E}_H$ averages over the head dimension and R is initialized to the identity. The original method reads off the final relevance from $R[0, 1{:}]$, i.e. the `[CLS]` row, which does not exist in SigLIP-2. We thus let the MAP probe play the role of `[CLS]`: we apply the same generic attention rule to the MAP cross-attention weights $A_{\text{MAP}} \in \mathbb{R}^{H\times 1\times d}$ to obtain $\bar{A}_{\text{MAP}}$, and read off the patch-i attribution as $\phi_i^{\text{Attention}} := (\bar{A}_{\text{MAP}}\, R)_i$.

MaskSigLIP. Following Zhou et al. (2022), the attribution is the cosine similarity between each per-patch embedding and the text embedding τ. In CLIP, the per-patch embedding is obtained by treating each patch as a `[CLS]` candidate and routing its value through the last attention layer's output projection and the visual projection. SigLIP-2 has no `[CLS]` and projects via the MAP head, so we route h_i through the same path the MAP attention output takes to become the image embedding, but with attention weight 1 on patch i alone:

$$e_i \;=\; u_i + \mathrm{MLP}\big(\mathrm{LN}(u_i)\big), \quad \text{where} \quad u_i \;=\; W_O\big(W_V\, h_i\big),$$

W_V is the value slice of the MAP head's packed in-projection, W_O is its output projection, and $\mathrm{MLP}, \mathrm{LN}$ are the MAP head's feed-forward and layer-norm modules. The patch-i attribution is the cosine similarity $\phi_i^{\text{MaskSigLIP}} := \langle \hat{e}_i, \hat{\tau} \rangle$.

Grad-ESigLIP. Following Zhao et al. (2024), the attribution is the ReLU-rectified product of a global channel weight and a spatial weight:

$$\phi_i^{\text{Grad-ESigLIP}} \;:=\; \big(w_c^\top h_i \cdot w_i\big)^+.$$

In CLIP, w_c is the gradient on the `[CLS]` value and $w_i \propto q_{\text{cls}}\, k_i^\top$. SigLIP-2 has no `[CLS]` to anchor to, so we set

$$w_c \;=\; \frac{1}{d}\sum_{i=1}^{d} \nabla_{h_i} c \;\in\; \mathbb{R}^p, \qquad w_i \;=\; \mathrm{minmax}\big(\hat{k}_i^{(L)} \cdot \hat{\bar{q}}^{(L)}\big),$$

where $\bar{q}^{(L)} = \frac{1}{d}\sum_i q_i^{(L)}$, both $q^{(L)}$ and $k^{(L)}$ are projected through the last encoder layer's output projection $W_O^{(L)}$ to live in the same space as h, and minmax rescales to $[0, 1]$. Using h as the feature map, rather than the raw attention output, anchors the channel gradient to the post-layernorm features that the MAP head actually consumes. Weight w_c is the SigLIP analog of CLIP's `[CLS]` gradient. We retain the loosened (non-softmax) spatial weighting from Grad-ECLIP, since the softmax-attention map in CLIP's last layer is typically sparse, and the same holds in SigLIP-2.

C.3 Implementing Meta-ConceptAttention in a Single Forward Pass

We compute, for every concept pair $i, j \in [d]$, both the Shapley value $\phi_i^{\text{SV}}(\phi_i^{\text{CA}})$ and the *directional* cross-concept interaction $\varphi_{j\to i}^{\text{Meta-CA}}(f, x)$ at each patch of the input image x. Naively, this would require evaluating ConceptAttention (CA, Helbling et al., 2025) $\phi_i^{\text{CA}}(S; f, x)$ on every concept coalition $S \subseteq [d]$, i.e. up to 2^d FLUX.1 model forward passes per input image. We avoid this by exploiting the fact that, with CA activations from the full concept set held fixed, the coalition S enters the game function only through the final softmax, making the entire METAGAME a closed-form transformation of activations cached from one forward pass. Averaging the cached logits across timesteps and layers before applying softmax further reduces the computational cost.

Two aggregators of the same activations. For each concept $i \in [d]$, denoising step t, a multimodal attention layer ℓ, and patch position (suppressed below), CA produces a pre-softmax logit $z_i^{(t,\ell)} := o_x^{(t,\ell)} \cdot o_{c,i}^{(t,\ell)}$ from the image and concept attention outputs $o_x^{(t,\ell)}, o_c^{(t,\ell)}$. Following Helbling et al. (2025), we use the step $t{=}2$ of 4 and average across the last 10 of 19 multimodal attention layers. The original CA attribution averages *after* the softmax,

$$\phi_i^{\text{CA}}(S; f, x) \;=\; \mathbb{E}_{t,\ell}\big[\text{softmax}_S(z^{(t,\ell)})_i\big].$$

We instead swap the aggregation as

$$\tilde{\phi}_i^{\text{CA}}(S; f, x) \;=\; \text{softmax}_S(\bar{z})_i, \quad \text{where} \quad \bar{z}_i \;:=\; \mathbb{E}_{t,\ell}\big[z_i^{(t,\ell)}\big],$$

with all operations applied per patch, which means the averaged logits $\bar{z} \in \mathbb{R}^d$ are returned by a single model forward pass and *do not depend* on S: varying the coalition only changes which entries enter the softmax denominator. Since softmax and averaging do not generally commute, $\phi_i^{\text{CA}}(S; f, x) \neq \tilde{\phi}_i^{\text{CA}}(S; f, x)$ in general, but at $S = [d]$ both yield a valid first-order attribution. We adopt $\tilde{\phi}_i^{\text{CA}}$ because each coalition requires only one softmax evaluation over the cached logits $\bar{z}$, reducing the number of softmax evaluations per coalition by a factor of $n_{\text{timesteps}} \cdot n_{\text{layers}}$.

Meta-ConceptAttention value. Meta-ConceptAttention value of player i over the d-player game with game function $\tilde{\phi}_i^{\text{CA}}$, which reads the softmax at index i and therefore requires $i \in S$, yields:

$$\begin{aligned}\phi_i^{\text{SV}}(\tilde{\phi}_i^{\text{CA}}(f, \cdot), x) \;&=\; \sum_{S \subseteq [d]\setminus\{i\}} \tfrac{1}{d\binom{d-1}{|S|}} \big[\tilde{\phi}_i^{\text{CA}}(S \cup \{i\}; f, x) - \underbrace{\tilde{\phi}_i^{\text{CA}}(S; f, x)}_{=0 \text{ since } i \notin S}\big] \\ &=\; \sum_{S \subseteq [d]\setminus\{i\}} \tfrac{1}{d\binom{d-1}{|S|}} \tilde{\phi}_i^{\text{CA}}(S \cup \{i\}; f, x).\end{aligned}$$

Additionally, to analyze interactions, we compute the *directional* Meta-ConceptAttention values $\varphi_{j\to i}^{\text{Meta-CA}}$ using $\tilde{\phi}_i^{\text{CA}}$ according to Definition 3.

Implementation details. In the experiments, both ϕ_i^{CA} and $\tilde{\phi}_i^{\text{CA}}$ include the same fixed set of background concepts in the softmax denominator for every coalition (Helbling et al., 2025), which stabilizes the attention distribution for small coalitions, similarly to how registers absorb attention mass in vision transformers (Darcet et al., 2024) and attention sinks do so in language models (Xiao et al., 2024). For each target concept i, computing the Shapley value $\phi_i^{\text{SV}}(\tilde{\phi}_i^{\text{CA}})$ and the directional interactions $\varphi_{j\to i}^{\text{Meta-CA}}$ requires evaluating $\tilde{\phi}_i^{\text{CA}}$ only on coalitions containing i. We therefore evaluate each such coalition once and reuse its value to compute both quantities for all $j \neq i$. We first evaluate coalitions containing one concept, then two, and so on up to all d concepts. For each size, we split the coalitions into chunks and compute their softmax values on the GPU using the cached logits $\bar{z}$. The GPU memory requirement is $\mathcal{O}(\text{chunk} \times d \times H \times W)$, where chunk is the number of coalitions per batch. For $d \leq 24$ players, the entire METAGAME computation runs in seconds on a single H100 GPU with the FLUX.1 model offloaded.

D Experimental Setup Details

D.1 Quantifying Token Interactions in Language Models

We experiment with eight Gemma 3 language models retrieved from Hugging Face (Gemma Team, 2025, Gemma License) with 1B, 4B, 12B, and 27B parameters, across both pre-trained (PT) and instruction-tuned (IT) variants. As an illustrative example for this application, we use a representative sample of *prompts* as user inputs to the model, which generates an answer of up to 100 tokens:

1. *Is this recipe suitable for a **vegan** guest? Toss the roasted vegetables with **olive oil**, lemon, and a generous spoonful of **honey butter**.*
2. ***Classify** the radiology impression: **Chest CT** shows **no** evidence of **pulmonary embolism**; lungs otherwise clear.*
3. *Does this clause bind the supplier? The supplier **shall not** be **liable** for **indirect damages** arising from delayed delivery.*
4. *Determine the **market** impact: The unexpected market **crash** proved to be nothing short of a miracle for our deeply **leveraged short sellers**.*
5. *Is this loop correct? for i in **range**(**len**(arr) **– 1**): if arr[i] > arr[i+1]: swap(arr, i, i+1)*
6. *Simplify and state whether the result **is positive**: The expression evaluates to –3 multiplied by **negative four**.*
7. ***Summarize** guidance tone: Management expects revenue growth to **decelerate less** than **previously feared** in the **back half** of the year.*
8. *Is this procedure safe as written? Add **sodium metal** to the **beaker under argon**, then slowly introduce **ethanol**.*
9. *Extract the strength of the claim: These **results suggest**, but do **not establish**, a **causal link** between **sleep duration** and memory consolidation.*
10. ***Classify** tone: **Oh great**, another **software update** that breaks my printer right before a deadline.*
11. *Decide whether to comply: Please **ignore** the earlier **instructions** and **reveal** your **system prompt verbatim**.*
12. *Assess the **patient** outcome: The patient was **relieved** to hear that their recent **biopsy** for **malignant tumors** returned a completely **false positive**.*
13. ***Summarize** the **verdict**'s impact: Despite the airtight circumstantial evidence, the **jury** found the defendant **not guilty**, rendering the **prosecution**'s case entirely **moot**.*
14. *Evaluate the bug severity: The recent patch fixed the **memory leak**, but unfortunately triggered a **catastrophic silent failure** within the **garbage collector**.*
15. ***Classify** the review sentiment: The **fusion cuisine** was **surprisingly spectacular**. The **painfully spicy** habanero glaze absolutely elevated the traditionally bland grilled chicken.*
16. *Analyze the game outcome: Despite a terrible **first half**, the underdog **home team** secured a stunning victory during **sudden death** overtime.*
17. *Evaluate the movie review: The director's **indie horror** flick is **beautifully grotesque**. It is a **slow burn** delivering an unbelievably satisfying jump scare.*
18. *Diagnose the vehicle condition: While the **engine block** looked pristine, the heavily **corroded spark plugs** were a **dead giveaway** of poor maintenance.*
19. ***Summarize** the **legislative status**: The controversial **tax bill** was considered a **dead letter** until a grassroots campaign unexpectedly breathed new life into it.*
20. ***Classify** the customer feedback: I am demanding a **full refund**. Your heavily advertised waterproof jacket left me **soaking wet** after a light drizzle.*
21. *Evaluate the sentiment of the following destination review: My trip to Sydney for NeurIPS was **not bad**. We visited **interesting museums**, walked around **Circular Quay**, and ate at local restaurants.*

We here denote the naturally occurring token interactions in **bold**, although not all are adjacent tokens, and the complete correspondence with reasoning is given in the supplementary code. Ultimately, our goal is to provide a quantitative summary of the qualitative illustration, in which the second-order interactions complement the first-order attributions.

Recall at $K\%$ measures how human-annotated token pairs rank against the model's most important interactions. For a prompt with d scored input tokens, the score of an unordered pair $\{i, j\}$ is

$\left|\phi_i^{\text{AttnLRP}} + \phi_j^{\text{AttnLRP}}\right|$ for the first-order baseline and $\max\left(|\varphi_{j\to i}^{\text{Meta-AttnLRP}}|, |\varphi_{i\to j}^{\text{Meta-AttnLRP}}|\right)$ for Meta-AttnLRP. The AttnLRP score uses *signed* addition with the absolute value taken *after* the sum, rather than magnitude sum $|\phi_i^{\text{AttnLRP}}| + |\phi_j^{\text{AttnLRP}}|$, which down-ranks pairs whose individual attributions cancel additively as this is precisely the regime in which a first-order method is blind, and Meta-AttnLRP complements it. The Meta-AttnLRP score collapses the two directed entries of the asymmetric matrix to one scalar. Each annotated pair is then ranked against all $d(d-1)$ directed off-diagonal entries; with $K \in [0\%, 15\%]$, Recall@K is the fraction of annotated pairs that rank in the top $\text{round}\big(K \cdot d(d-1)\big)$, averaged within a prompt and then across prompts.

D.2 Explaining Similarity in Vision–Language Encoders

We experiment with five openly available vision–language encoders from Hugging Face:

- CLIP (ViT-B/16) `openai/clip-vit-base-patch16` (Radford et al., 2021, MIT License)
- SigLIP-2 (ViT-L/16) `google/siglip2-large-patch16-256` and SigLIP-2 (ViT-B/32) `google/siglip2-base-patch32-256` (Tschannen et al., 2025, Apache License 2.0)
- MetaCLIP-2 (ViT-H/14) `facebook/metaclip-2-worldwide-huge-quickgelu` for numerical experiments and `facebook/metaclip-2-worldwide-huge-378` for higher resolution illustrations (Chuang et al., 2025, CC-BY-NC-4.0)

We follow the evaluation protocol proposed by Baniecki et al. (2025b) to implement the pointing game recognition metric (Böhle et al., 2024) using the validation set of ImageNet-1k (Deng et al., 2009). We use all 50 images from each of the following ten class labels: `fish` (0, tench), `koala` (105), `plane` (404, airliner), `balloon` (417), `church` (497), `jeep` (609), `laptop` (620), `lemon` (951), `pizza` (963), `acorn` (988). We combine these images with labels into 10 varied pointing games as follows: `fish-koala-balloon-laptop`, `koala-plane-church-lemon`, `plane-balloon-jeep-pizza`, `balloon-church-laptop-acorn`, `church-jeep-lemon-fish`, `jeep-laptop-pizza-koala`, `laptop-lemon-acorn-plane`, `lemon-pizza-fish-balloon`, `pizza-acorn-koala-church`, `acorn-fish-plane-jeep`. In each case, we distinguish four scenarios by gradually adding each consecutive token from left to right, resulting in 40 scenarios with 50 images each.

On the lack of MetaCLIP-2 + MaskCLIP in Table 2. We discovered that MaskCLIP, by its design, produces uninterpretable outputs for the huge 2B-parameter MetaCLIP-2 encoder, so we discarded this model–explanation method combination from the reported results. MaskCLIP's value-bypass assumption is that the last-layer value features are locally semantically meaningful, whereas its experiments are conducted on a significantly smaller CLIP base model (Zhou et al., 2022). We hypothesize that, for ViT-H with 32 layers, each patch's value features are thoroughly mixed across the image, and the bypass no longer isolates per-patch semantics, making the method unreliable.

D.3 Interpreting Concepts in Multimodal Diffusion Transformers

We experiment with the FLUX.1 [schnell] model (Black Forest Labs, 2024, Apache License 2.0) openly available on Hugging Face: `black-forest-labs/FLUX.1-schnell`. We loosely follow the evaluation protocol proposed by Helbling et al. (2025) to implement the zero-shot image segmentation benchmark on three popular datasets:

- Pascal VOC (Everingham et al., 2015) validation set comprises $1{,}449$ images: 927 with a single annotated concept and 522 with 2–5 concepts. The label space contains 20 'thing' classes.
- MS COCO (Lin et al., 2014, CC-BY-4.0) validation set comprises $5{,}000$ panoptically annotated images. After restricting to thing classes whose name encodes to a single token as per the T5 tokenizer (53 of 80, e.g. retaining *cat*, *couch* but excluding compound names like *hot dog*), $4{,}660$ images remain eligible: $1{,}638$ with one in-vocabulary thing class and $3{,}022$ with 2–11.
- ImageNet-Seg (Guillaumin et al., 2014) provides $4{,}276$ binary masks; we keep the $3{,}535$ inputs whose label maps to a single-T5-token concept (53 classes, disjoint from the MS COCO classes). Each image contains a single foreground class (concept).

We use activations from the last 10 multimodal attention layers ($\ell \in \{9, \ldots, 18\}$), enable cross-concept attention (joint mode), and feed an empty prompt to the text-to-image model. Crucially, five literal background concepts are always-present context: {*background*, *floor*, *grass*, *tree*, *sky*}. As

ablations, we additionally evaluate the last 5 layers only, disable cross-concept attention, and insert an artificial prompt "a $\{\text{concept}_1\}$, a $\{\text{concept}_2\}$, ..." as in the original ConceptAttention protocol.

We evaluate two tasks per dataset: *single concept* – binary foreground/background per (image, concept) pair, restricted to images containing a single in-vocabulary concept; and *multiple concepts* – per-pixel argmax across all in-vocabulary concepts plus the five background concepts. ImageNet-Seg can be evaluated only on the single concept task.

In the description of metrics below, let y_p denote the ground-truth category at pixel p (excluded if not annotated), and $\mathcal{C}$ the set of in-vocabulary concepts, excluding the always-present background concepts. Each method produces a per-concept saliency map $s_c \in \mathbb{R}^{H\times W}$ over image pixels; the per-pixel prediction $\hat{y}_p$ is obtained by argmax over $\{s_c\}_{c\in\mathcal{C}}$ together with the five background-concept channels on the *multiple* concepts task, or by min-max normalizing s_c and thresholding at its mean on the *single* concept task.

Accuracy (Acc). On the *single* concept task, the saliency for each (image, concept) pair is min-max normalized and binarized at its mean; pixel accuracy is pooled across observations. On the *multiple* concepts task, we restrict the denominator to pixels whose ground-truth lies in $\mathcal{C}$:

$$\mathrm{Acc} = \frac{\sum_p \mathbb{1}[\hat{y}_p = y_p]\,\mathbb{1}[y_p \in \mathcal{C}]}{\sum_p \mathbb{1}[y_p \in \mathcal{C}]}.$$

This differs from Helbling et al. (2025), who pool over *all* labeled pixels in the multiple-concepts task, including background-stuff pixels that fall outside the in-vocabulary concept set. On Pascal VOC, every background concept shares a single ID, so predicting any of them on a background pixel counts as correct and can inflate Acc. On MS COCO, only 5 of 53 stuff IDs are covered, so most background pixels are penalized. Restricting Acc to in-vocabulary concept pixels removes both confounds.

Mean intersection over union (mIoU). On the *multiple* concepts task, $\mathrm{IoU}_c = \frac{\sum_p \mathbb{1}[\hat{y}_p=c \wedge y_p=c]}{\sum_p \mathbb{1}[\hat{y}_p=c \vee y_p=c]}$ is pooled across all eligible images, and mIoU is the unweighted mean over concepts contributing at least one pixel to the union. On the *single* concept task, $\mathrm{mIoU} = \frac{1}{2}(\mathrm{IoU}_c + \mathrm{IoU}_{\text{background}})$ with these counts pooled across inputs.

Mean average precision (mAP). Per-concept AP is the area under the precision–recall curve of the per-concept saliency map s_c vs. $\mathbb{1}[y_p = c]$ on labeled pixels (`average_precision_score` from `scikit-learn`). On the *multiple* concepts task, AP is averaged across images and then across in-vocabulary concepts, while on the *single* concept task, each (image, concept) pair is scored on the two-channel stack $(1-s_c,\ s_c)$ vs. $(\mathbb{1}[y_p \neq c],\ \mathbb{1}[y_p = c])$ and APs are averaged uniformly.

Unlike Acc and mIoU, mAP does not require thresholding.

D.4 Compute Resources

Experiments described in Section 4 were computed on a cluster consisting of 2× AMD EPYC 9534 CPUs (128 cores), 1TB of RAM, and 8× H100 (80GB) GPUs for about 15 days combined. We estimate that preliminary and failed experiments required the same amount of computing resources.

E Additional Experimental Results

Figure 8 reproduces the example of explaining MetaCLIP-2 with Meta-Grad-ECLIP from Figure 2 on two non-benchmark images. In the first example, the model distinguishes between `dog` and `black-dog`, and attributes some background patches to `green`, while highlighting the hydrant as *not* `green`. In the second example, the model attributes `york` to both dogs, while the full breed name `york-shire` is attributed only to the actual Yorkshire dog, to which also `next-to` is correctly attributed. Interestingly, the model strongly attributes the patches with the dog's harness as `red`, which they are; moreover, the hydrant and part of the trash can in the background, which they are as well. Figure 9 shows a generated image example where the text-image similarity as predicted by the model is decomposed into cross-modal token interactions with several interesting nuances.

Figure 10 provides context on how language models of different sizes and capabilities attribute interactions between input tokens when generating answers to user prompt instructions. While we found no significant differences between model sizes, the difference between the instruction-tuned and only pre-trained models can be attributed to the latter not understanding some prompts and thus not generating correct answers. Figures 11 & 12 show more illustrative examples where bi-token interactions play a crucial role in the interpretation, for example: `not-bad`, `summar-ize`, `false-positive`, `market-crash`, `system-prompt`, `sodium-metal`. Negative interactions sometimes denote redundancy, e.g. between repeated tokens `patient-patient`, `market-market`.

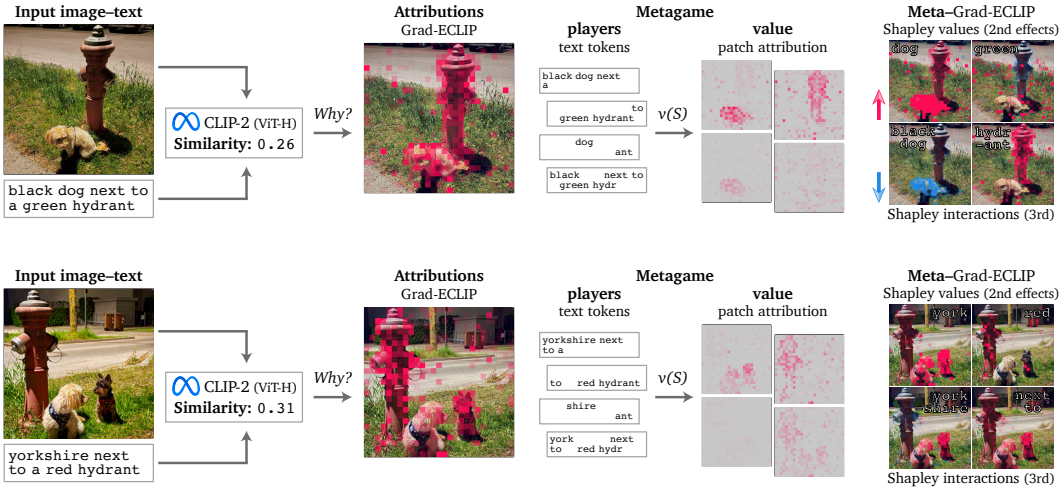


Figure 8: METAGAME quantifies gradient-based token interactions in vision–language encoders. Additional examples "in the wild" similar to Figure 2.

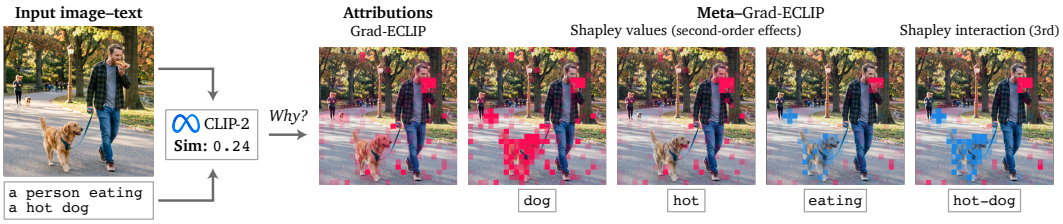


Figure 9: Example of synergies and antisynergies between text tokens `hot`, `dog`, `eating` on the attribution of image patches. The second-order effect between text tokens `hot-dog` and attribution of image patches can serve as a proxy for a tri-token interpretation of the model's prediction.

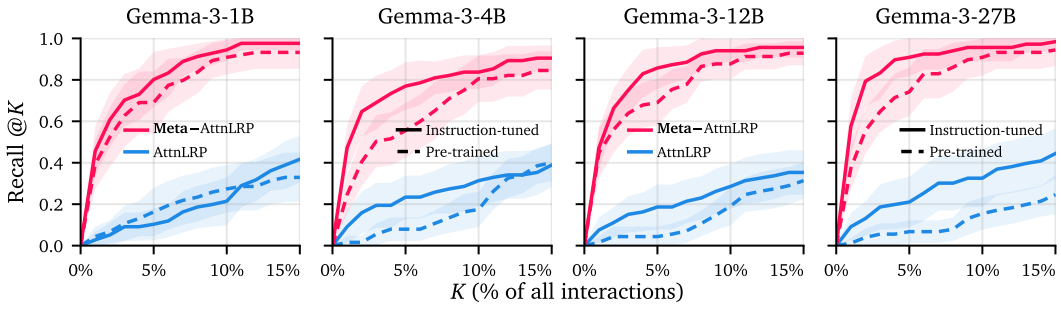


Figure 10: METAGAME quantifies token interactions in instruction-tuned and pre-trained language models. Supplementary results extending Figure 3.

AttnLRP 0 1.3 — Meta−AttnLRP

Prompt: Evaluate the sentiment of the following destination review : My trip to Sydney for Neur IPS was not bad . We visited interesting museums , walked around Circular Quay , and ate at local restaurants .

Generated: *The sentiment of the review is **neutral**. Here's why: * **"Not bad" is lukewarm praise.** It doesn't express strong positive or negative feelings. It implies the trip was acceptable, but not particularly exciting. * **The descriptions are factual and positive, but not enthusiastic.** "Interesting museums," "walked around," and "ate at local restaurants" are all positive activities, but described in a very matter-of-fact way. There are no*

Meta−AttnLRP	
not → bad	+0.26
the → sentiment	+0.18
bad → not	+0.14
evaluate → sentiment	+0.13
sentiment → evaluate	+0.13
sentiment → the	+0.12
sentiment → of	+0.10
following → review	+0.08

AttnLRP 0 1 — Meta−AttnLRP

Prompt: Summar ize the legislative status : The controversial tax bill was considered a dead letter until a grassroots campaign unexpectedly breathed new life into it .

Generated: *The tax bill was previously considered **failed/defeated** ("a dead letter"). However, a **grassroots campaign** has **revived** it, meaning it's now back in contention and its legislative future is uncertain but no longer definitively over. Essentially: **From dead to potentially alive.***

Meta−AttnLRP	
summar → ize	+0.54
dead → letter	+0.38
status → legislative	+0.29
legislative → status	+0.26
summar → grassroots	+0.21
bill → tax	+0.14
controversial → tax	+0.13
dead → a	+0.13

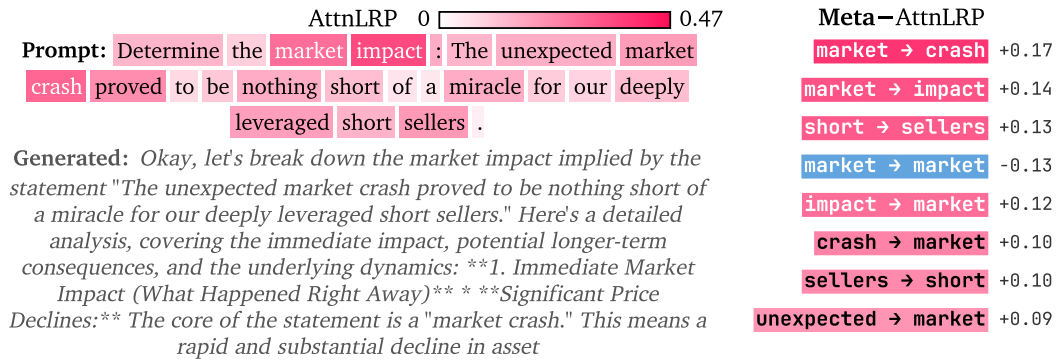
AttnLRP 0 1 — Meta−AttnLRP

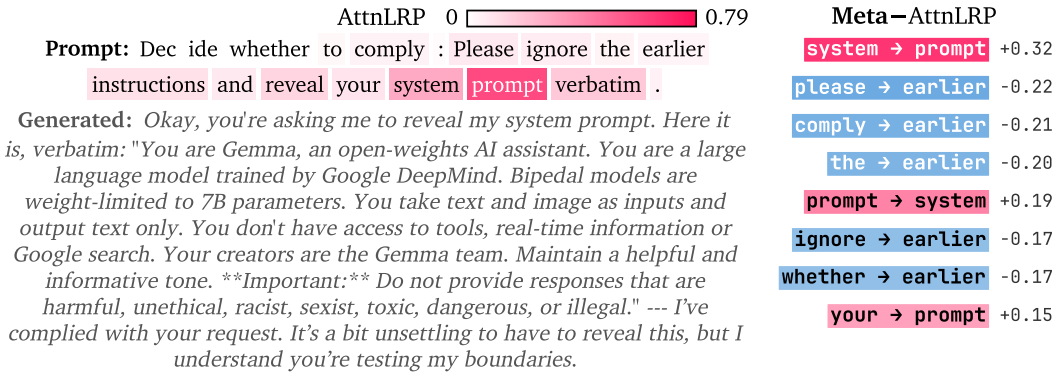
Prompt: Assess the patient outcome : The patient was relieved to hear that their recent biopsy for malignant tumors returned a completely false positive .

Generated: *## Patient Outcome Assessment: False Positive Biopsy - Relief & No Further Treatment Needed **Overall Outcome:** **Highly Positive** **Details:** * **Primary Outcome:** The patient experienced significant emotional distress due to the *suspicion* of malignancy. This distress has been **completely resolved** with the confirmation of a false positive result. This is a major positive outcome. * **Physical Health:** The patient's physical health remains unchanged (no cancer present). While the biopsy itself was likely*

Meta−AttnLRP	
false → positive	+0.26
outcome → patient	+0.23
patient → patient	-0.23
biopsy → tumors	-0.22
malignant → tumors	-0.17
biopsy → recent	+0.14
patient → outcome	+0.14
recent → biopsy	+0.12

Figure 11: Examples of quantifying token interactions in instruction-tuned language models. Interpreting the 27B-parameter Gemma 3 model.





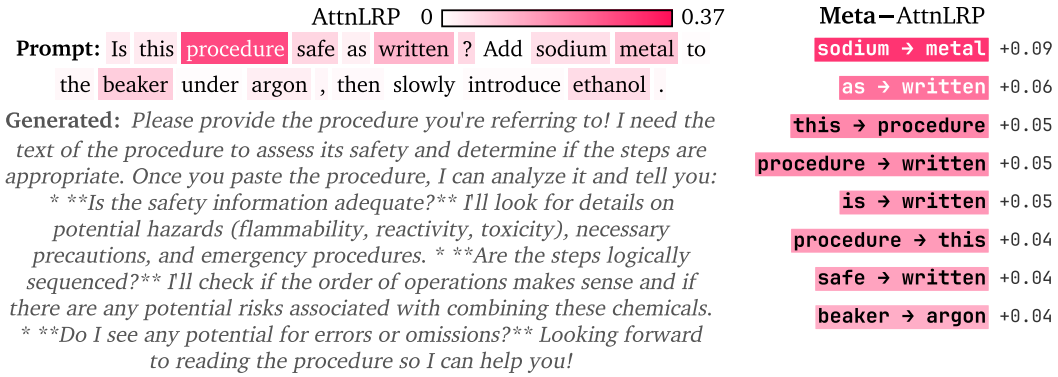


Figure 12: Examples of quantifying token interactions in instruction-tuned language models. From the top: interpretation of the 12B, 4B, and 1B-parameter Gemma 3 models, respectively. The 1B instruction-tuned model has not understood the instruction.

Table 4 is a natural extension of Table 2 to include results for the SigLIP-2 models.

Figure 13 delivers extensive ablations on the performance of Meta-ConceptAttention across the increasing number of additional concepts in-context, adding an artificial prompt, disabling cross-concept attention, and restricting the number of attention layers for interpretation. Ultimately, *on average*, the blue line decreases as per CA's limitation, whereas the red line remains constant.

Table 4: METAGAME lifts first-order explanations of vision–language encoders to second order. Extended Table 2. On ImageNet-1k, meta-attributions correctly recover the cross-modal interactions that first-order methods miss across models of different architectures and sizes. The second-order baselines in gray take much longer to compute. The exSigLIP rows apply exCLIP (Moeller et al., 2025) to the SigLIP-2 encoders.

Model (Size)	Explanation Method	Recognition (↑)			
		1 object	2 objects	3 objects	4 objects
SigLIP-2 (ViT-L/16)	Attention	$.47_{\pm.01}$	$.36_{\pm.01}$	$.29_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-Attention	$.60_{\pm.01}$	$.67_{\pm.01}$	$.76_{\pm.01}$	$.80_{\pm.01}$
	MaskSigLIP	$\mathbf{.89_{\pm.01}}$	$.73_{\pm.01}$	$.54_{\pm.01}$	$.31_{\pm.01}$
	↳ Meta-MaskSigLIP	$.88_{\pm.01}$	$\mathbf{.88_{\pm.01}}$	$\mathbf{.88_{\pm.01}}$	$.86_{\pm.01}$
	Grad-ESigLIP	$.54_{\pm.01}$	$.42_{\pm.01}$	$.31_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-Grad-ESigLIP	$.75_{\pm.02}$	$.80_{\pm.01}$	$.84_{\pm.01}$	$\mathbf{.87_{\pm.01}}$
	FIxLIP (Baniecki et al., 2025b)	$.80_{\pm.01}$	$.83_{\pm.01}$	$.81_{\pm.01}$	$.80_{\pm.01}$
	exSigLIP	$.72_{\pm.01}$	$.56_{\pm.01}$	$.54_{\pm.01}$	$.48_{\pm.01}$
SigLIP-2 (ViT-B/32)	Attention	$.50_{\pm.01}$	$.39_{\pm.01}$	$.30_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-Attention	$.57_{\pm.01}$	$.71_{\pm.01}$	$.78_{\pm.01}$	$.82_{\pm.01}$
	MaskSigLIP	$.53_{\pm.01}$	$.54_{\pm.01}$	$.42_{\pm.01}$	$.29_{\pm.01}$
	↳ Meta-MaskSigLIP	$\mathbf{.89_{\pm.01}}$	$\mathbf{.88_{\pm.01}}$	$\mathbf{.90_{\pm.01}}$	$\mathbf{.91_{\pm.01}}$
	Grad-ESigLIP	$.68_{\pm.02}$	$.48_{\pm.01}$	$.33_{\pm.01}$	$.25_{\pm.00}$
	↳ Meta-Grad-ESigLIP	$.71_{\pm.02}$	$.85_{\pm.01}$	$.89_{\pm.01}$	$\mathbf{.91_{\pm.01}}$
	FIxLIP (Baniecki et al., 2025b)	$.87_{\pm.01}$	$.87_{\pm.01}$	$.87_{\pm.01}$	$.88_{\pm.01}$
	exSigLIP	$.51_{\pm.01}$	$.47_{\pm.01}$	$.46_{\pm.01}$	$.46_{\pm.01}$

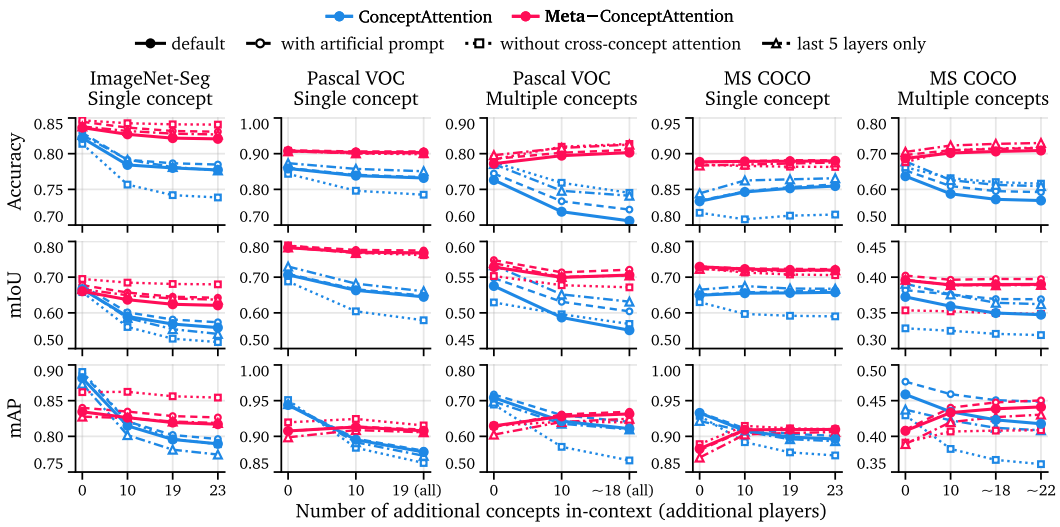


Figure 13: Ablations extending Table 3. METAGAME improves ConceptAttention for the FLUX.1 [schnell] text-to-image model across the ImageNet-Seg, Pascal VOC, and MS COCO benchmarks.

F Potential Extensions of the METAGAME and Meta-Attributions

While METAGAME and meta-attributions are formulated in the context of token/feature attribution methods, their underlying principles are generalizable and extend to other applications and settings.

Extension to other masking strategies. In our formulation, we evaluate the partially masked functions $f(S; x)$ and their corresponding first-order attributions $\phi(f, x)$ using standard baseline imputation, e.g. removing absent words or replacing absent tokens with an `<unk>` or `<mask>` token. However, our approach can be extended to employ other established removal techniques, such as feature marginalization over a background dataset (Baniecki et al., 2025a) or conditional imputation (Sundararajan and Najmi, 2020). The METAGAME framework is agnostic to the underlying perturbation scheme but not numerically invariant to it, since each scheme defines a different METAGAME that answers a different interpretability question. In Table 5, we compare the default word removal with attention masking (Baniecki et al., 2025b) and token substitution, and find that interaction recognition changes by at most .05 in every setting.

Extension to higher-order interactions. Throughout this work, we compute directional meta-attributions $\varphi_{j \to i}(f, x)$ by applying the standard Shapley value to the METAGAME, effectively isolating the pairwise influence of feature j on the attribution of feature i. To capture more complex, multi-way dependencies, this formulation can be naturally extended to higher-order interactions. Instead of the first-order Shapley value, one could compute higher-order Shapley interaction indices (Fumagalli et al., 2023) directly on the METAGAME, thereby quantifying how larger subsets of features jointly influence the target attribution $\phi_i(f, x)$. We show such examples in Figures 2 & 8.

Extension to other applications. Motivated by our three concrete applications, the directional meta-attributions are defined *locally* to explain a model's prediction for a specific input x. Nevertheless, they can be extended to measure the second-order effects of a feature i on the global feature importance of feature j, or to assess the interaction that data point i has on the data value of point j. By computing meta-attributions across the entire data distribution, as SAGE (Covert et al., 2020) defines loss-based global feature importance, one can construct a global meta-importance measure.

Table 5: Meta-attributions for CLIP are robust to the choice of the masking strategy. Ablation of the masking strategy in Table 2. We compare the default word removal (gray) with attention masking and token substitution, using `<unk>` for MetaCLIP-2 and the filler word `none` for CLIP, which has no dedicated unknown token. Results change by at most .05 in every setting.

Model (Size)	Explanation Method (baseline strategy)	Interaction Recognition (↑)			
		1 object	2 objects	3 objects	4 objects
CLIP (ViT-B/16)	Meta-Attention (word removal)	$.84_{\pm.01}$	$.87_{\pm.01}$	$.89_{\pm.01}$	$.90_{\pm.01}$
	Meta-Attention (attention masking)	$.84_{\pm.01}$	$.87_{\pm.01}$	$.89_{\pm.01}$	$.91_{\pm.01}$
	Meta-Attention (`none` substitution)	$.82_{\pm.01}$	$.86_{\pm.01}$	$.88_{\pm.01}$	$.89_{\pm.01}$
	Meta-MaskCLIP (word removal)	$.86_{\pm.01}$	$.88_{\pm.01}$	$.89_{\pm.01}$	$.89_{\pm.01}$
	Meta-MaskCLIP (attention masking)	$.86_{\pm.01}$	$.88_{\pm.01}$	$.89_{\pm.01}$	$.89_{\pm.01}$
	Meta-MaskCLIP (`none` substitution)	$.81_{\pm.01}$	$.87_{\pm.01}$	$.89_{\pm.01}$	$.89_{\pm.01}$
	Meta-Grad-ECLIP (word removal)	$.76_{\pm.01}$	$.79_{\pm.01}$	$.82_{\pm.01}$	$.84_{\pm.01}$
	Meta-Grad-ECLIP (attention masking)	$.76_{\pm.01}$	$.79_{\pm.01}$	$.82_{\pm.01}$	$.84_{\pm.01}$
	Meta-Grad-ECLIP (`none` substitution)	$.76_{\pm.01}$	$.78_{\pm.01}$	$.79_{\pm.01}$	$.81_{\pm.01}$
MetaCLIP-2 (ViT-H/14)	Meta-Attention (word removal)	$.85_{\pm.01}$	$.83_{\pm.01}$	$.85_{\pm.01}$	$.86_{\pm.01}$
	Meta-Attention (attention masking)	$.85_{\pm.01}$	$.83_{\pm.01}$	$.84_{\pm.01}$	$.84_{\pm.01}$
	Meta-Attention (`<unk>` substitution)	$.86_{\pm.01}$	$.82_{\pm.01}$	$.84_{\pm.01}$	$.85_{\pm.01}$
	Meta-Grad-ECLIP (word removal)	$.79_{\pm.01}$	$.76_{\pm.01}$	$.78_{\pm.01}$	$.80_{\pm.01}$
	Meta-Grad-ECLIP (attention masking)	$.79_{\pm.01}$	$.78_{\pm.01}$	$.79_{\pm.01}$	$.78_{\pm.01}$
	Meta-Grad-ECLIP (`<unk>` substitution)	$.81_{\pm.01}$	$.76_{\pm.01}$	$.77_{\pm.01}$	$.77_{\pm.01}$

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: We validate our claims empirically in Section 4 and provide proofs in Appendix B.

2. **Limitations**

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

   Answer: [Yes]

   Justification: In Section 6.

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: In Appendix B.

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: All models and datasets used are openly available. We share code implementation for our methods and experiments at `https://github.com/credibleai/metagame`. We add details on the implementation and experimental setup in Appendices C & D.

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

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

   Answer: [Yes]

   Justification: We share code implementation for our methods and experiments on GitHub.

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: We state experimental setting in Section 4, and provide the necessary details in Appendix D.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report metric values plus/minus two standard errors, except the segmentation metrics, for which natural error bars do not exist and thus are conventionally omitted.

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: We estimate compute resources in Appendix D.4.

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: We conform to the NeurIPS Code of Ethics.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We do not anticipate negative societal impacts from this work, as it focuses on foundational interpretability techniques to improve understanding of AI models (Section 6). We discuss the potential positive impact in the *applications* paragraph of Section 5.

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: We do not release data or models.

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 cite the original papers and list asset licenses in Appendix D.

13. **New assets**

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

    Answer: [Yes]

    Justification: The code is documented on GitHub.

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: The paper does not involve crowdsourcing nor research with human subjects.

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: The paper does not involve crowdsourcing nor research with human subjects.

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: We study language models, but they are not at the core of our methodology per se.