A Comprehensive Information-Decomposition Analysis of Large Vision-Language Models

Lixin Xiu[1], Xufang Luo[2*], Hideki Nakayama[1*]

[1] The University of Tokyo, [2] Microsoft Research

* Corresponding authors

Project Link

ICLR 2026

Outline

1. Background and motivation

2. Methodology: proposed pipeline and experimental settings

3. Findings and discussion

Background and motivation

Current interpretability works on large vision-language models (LVLMs) are mainly:

1. A “micro-scope” focus; 2. Ad hoc metrics.

When an LVLM succeeds, is it due to robust visual understanding, sophisticated language processing from LLM, or a true, synergistic fusion of both?

We turn to partial information decomposition (PID)!

Background of PID

There are 3 r.v.s. X_1, X_2 and Y over $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{Y}$.

Aim: We want to quantify the information of (X_1, X_2) provided about Y.

Problem: Interative information $I(X_1; X_2; Y)$ could be negative.

Motivation: Can we decompose information into some non-negative quantities?

3 types of information:

- **Redundancy:** R
- **Uniqueness of X_1 and X_2:** U_1 **and** U_2 **respectively**
- **Synergy:** S

Definition

Let Δ be the set of all joint distributions of X_1, X_2 and Y. Define

$$\Delta_P = \{Q \in \Delta : Q(x_i, y) = P(x_i, y), \forall\, x_i \in \mathcal{X}_i, y \in \mathcal{Y}, i \in \{1,2\}\},$$

where Δ_P is marginal-matching distributions of true distribution P.

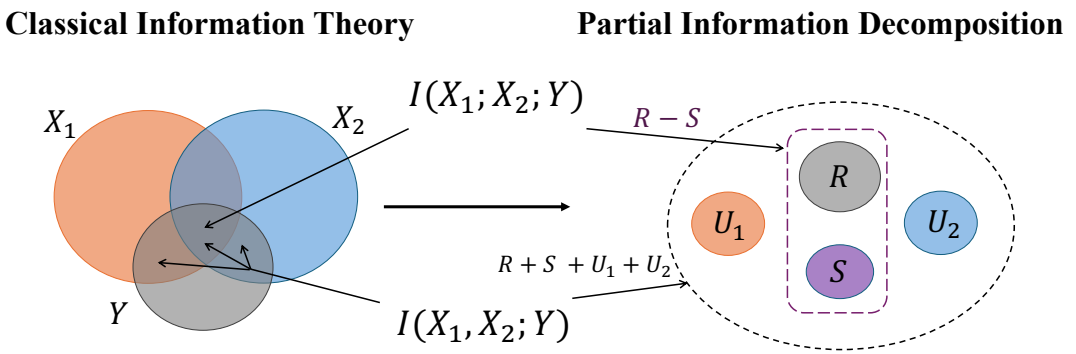
The formal definition of 4 PID quantities is as follows:

$$\tilde{R} = \max_{Q \in \Delta_P} I_Q(X_1; X_2; Y)$$

$$\widetilde{U_1} = \min_{Q \in \Delta_P} I_Q(X_1; Y | X_2)$$

$$\widetilde{U_2} = \min_{Q \in \Delta_P} I_Q(X_2; Y | X_1)$$

$$\tilde{S} = I_P(X_1, X_2; Y) - \min_{Q \in \Delta_P} I_Q(X_1, X_2; Y)$$



PID estimation pipeline

Task: Multiple-choice VQA, to fix $|Y|$ explicitly.

Inputs to BATCH estimator:

X_1: the **average vector** of visual token embeddings;

X_2: the **average vector** of textual token embeddings;

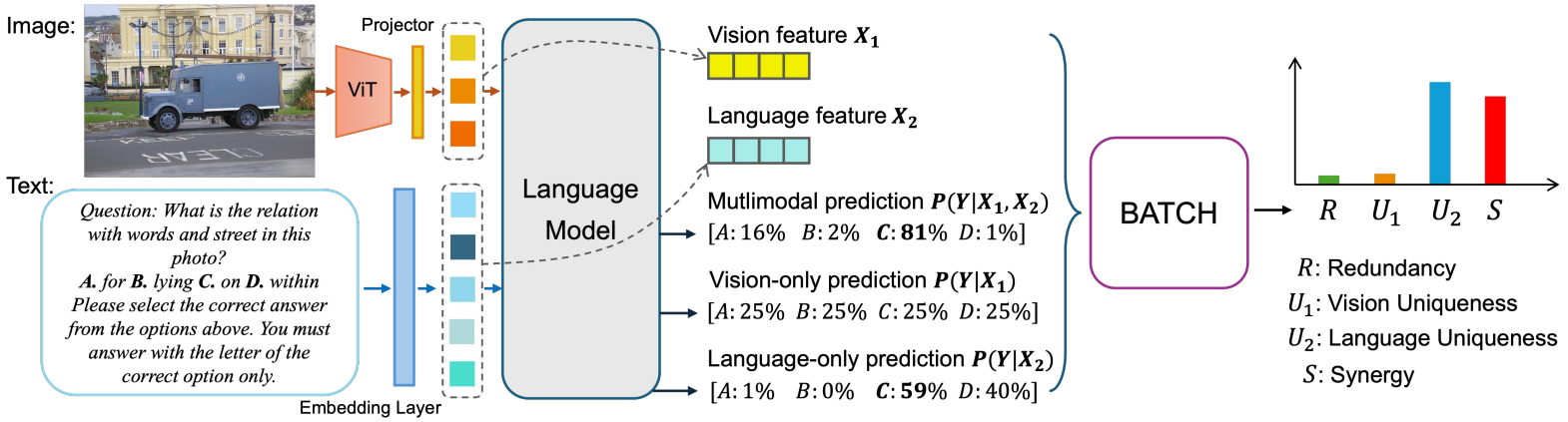
$P(Y|X_1, X_2)$: multimodal prediction **(two input modalities)**;

$P(Y|X_1)$: vision unimodal prediction **(only vision modality)**;

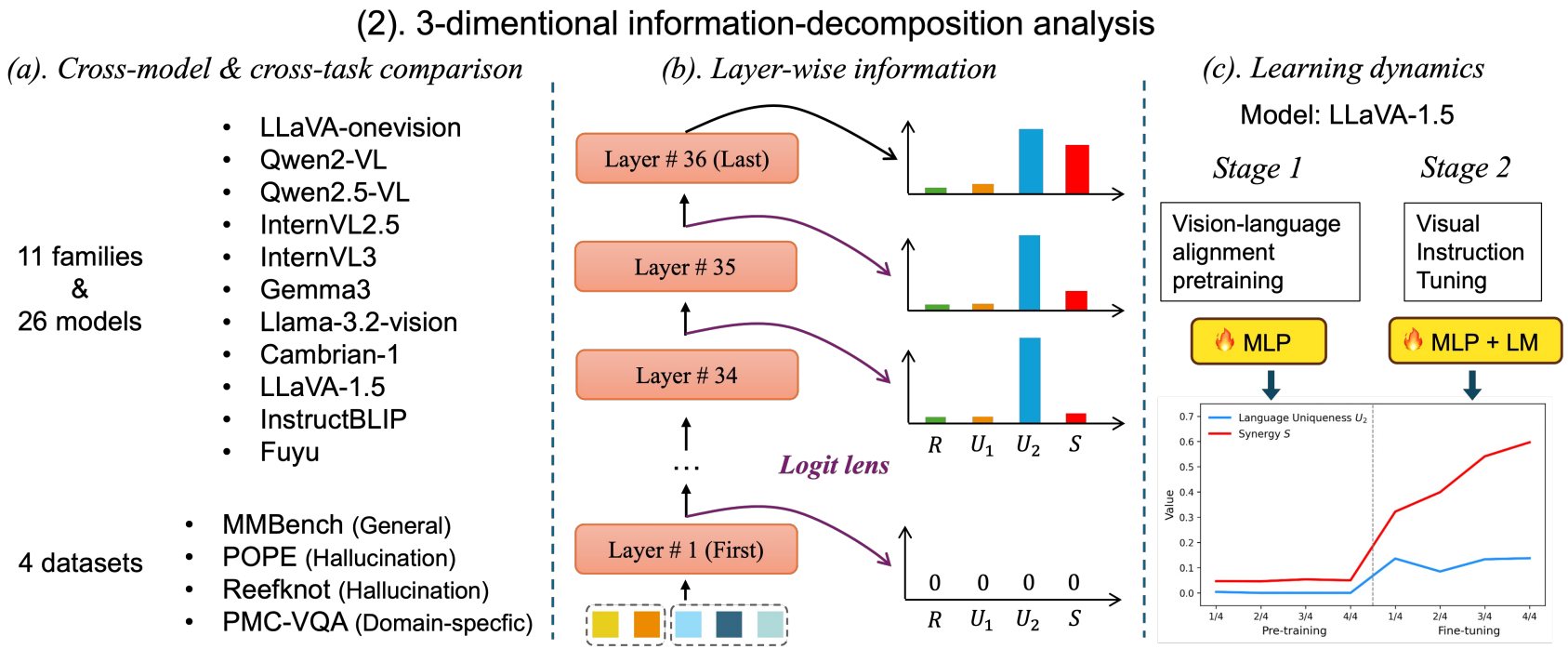
$P(Y|X_2)$: language unimodal prediction **(only language modality)**;

We adapt BATCH for more accurate PID estimation in LVLM scenarios!

(1). Proposed framework for PID estimation in LVLM scenario



Experimental settings

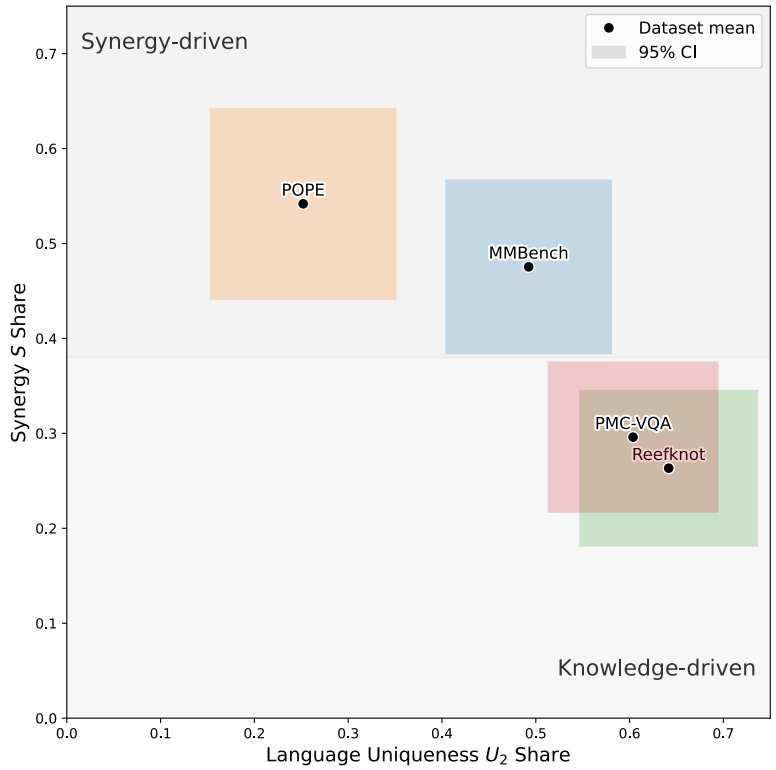


Analysis is conducted on 3 dimensions: **breadth (cross-model and cross-task)**, **depth (layer-wise)** and **time (learning dynamics)**.

To our knowledge, this is the first comprehensive LVLM analysis through the lens of information decomposition.

Main finding 1: Task regimes

We plot the dataset-level mean shares of S and U_2 averaged across all models for each dataset, with 95% CIs.

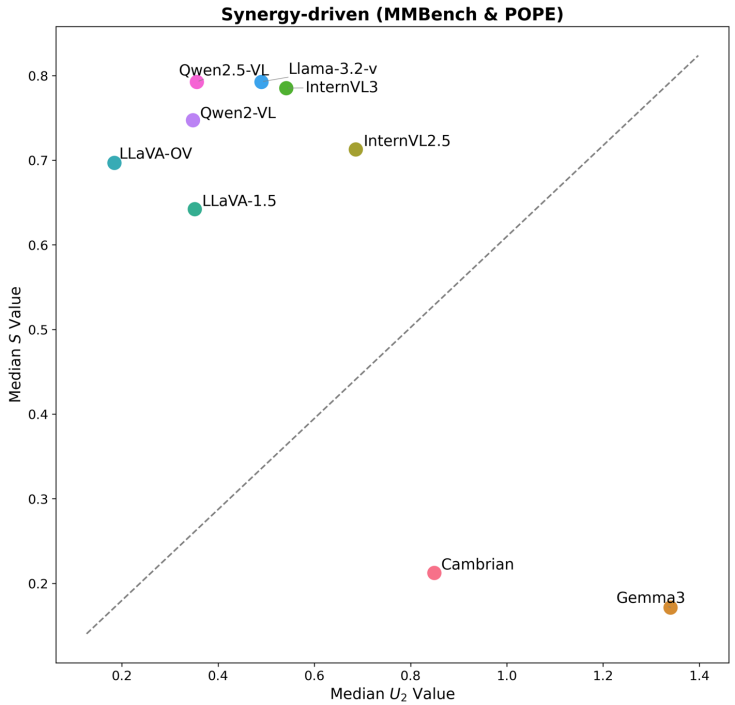


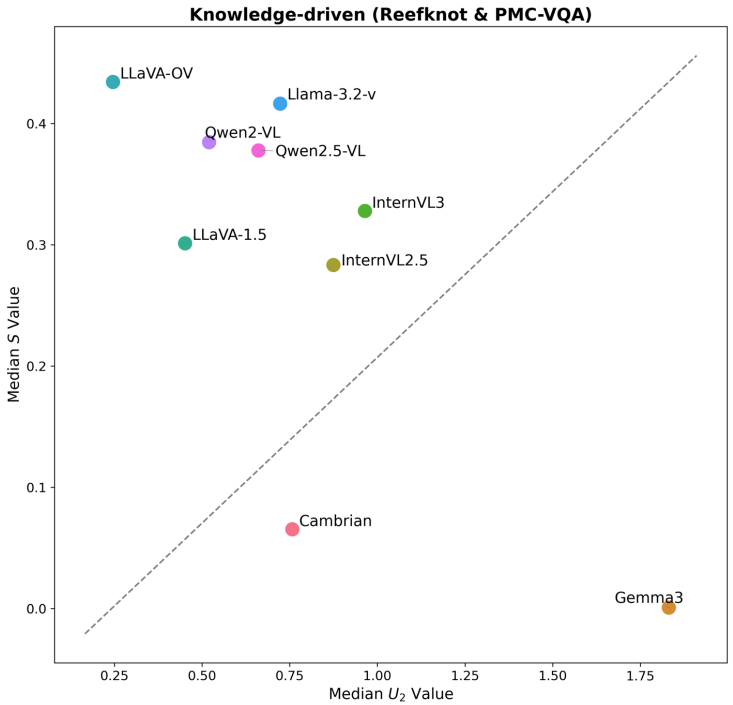
LVLM behavior is governed by two distinct information-use regimes. A **synergy-driven** regime characterizes general multimodal reasoning tasks that reward strong cross-modal fusion. In contrast, a **knowledge-driven** regime characterizes specialized tasks where model's language-side knowledge and priors fundamentally constrain performance.

Main finding 2: Model strategies

Do model families lean toward combining inputs or language knowledge—and does that preference hold across settings?

We summarize each family's behavior by its median S and U_2 within each regime:

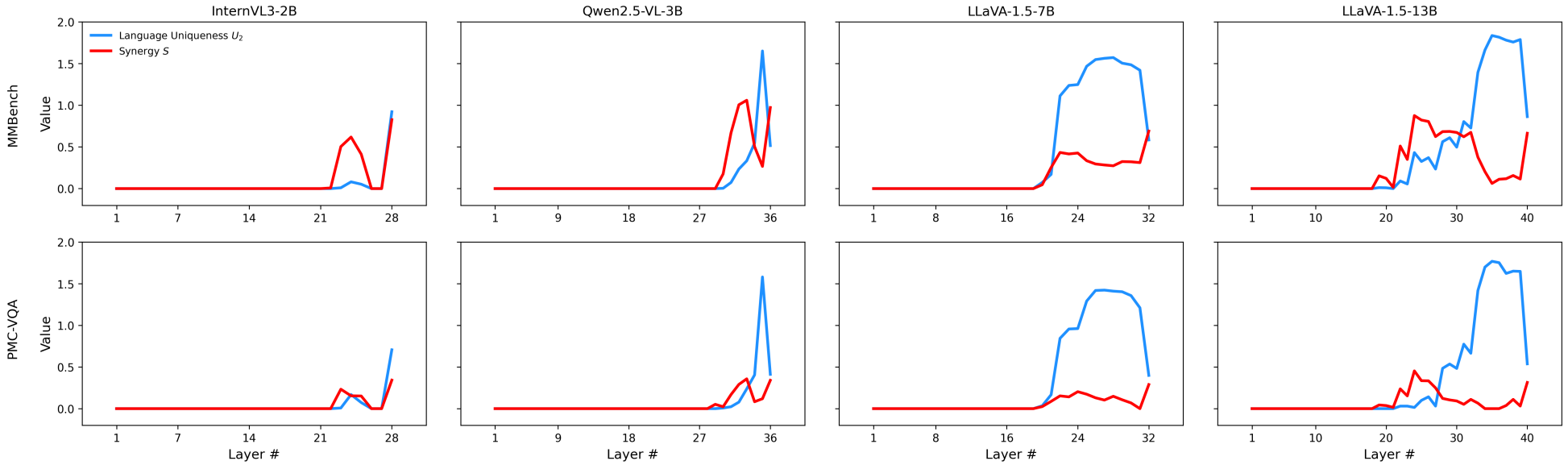




LVLM families adopt one of two stable strategies. Fusion-centric families consistently prioritize cross-modal reasoning S, while language-centric families rely more heavily on language priors U_2. This strategic identity persists across task regimes.

Main finding 3: Where does fusion emerge?

The next question is where S and U_2 arises in the stack of layers:



The layer-wise dynamics reveal a three-phase reasoning process. Information emerges in the middle-to-late layers, then moves from language-based representation building in the later layers to a decisive, synergistic fusion of modalities in the final layer.

Discussion

Higher accuracy does not necessarily imply stronger multimodal fusion.

PID complements accuracy-only evaluation by revealing whether success comes from true cross-modal interaction or language-side priors.

Looking ahead, $(U1, U2, S)$ can serve as diagnostic signals during scaling and instruction tuning, and potentially as auxiliary objectives to balance fusion and language priors. PID-based analyses can also guide the construction of benchmarks that explicitly require high synergy S or isolate language priors U_2.