Pseudo-Unification: Entropy Probing Reveals Divergent Information Patterns in Unified Multimodal Models

Supplementary Material

We provide additional theoretical justification, validation experiments, and implementation details to support the claims in the main paper and ensure full reproducibility. For a comprehensive overview, please refer to the following sections:

1. Theoretical Contributions

Our work establishes the first information-theoretic framework for diagnosing pseudo-unification in Unified Multimodal Models (UMMs). While matrix-based Rényi entropy has been applied in other Large-Language-Model-based contexts, our contribution is fundamentally novel in three theoretical dimensions:

- Reconceptualizing Information in Implicit Spaces: We reformulate information measures not as properties of explicit probability densities (which are unavailable in Transformer representations), but as geometric properties of representation structure in Reproducing Kernel Hilbert Spaces (RKHS). This bridges a critical theoretical gap between classical information theory and modern deep implicit models.
- Theoretical Foundation for Cross-Modal Dependency: We prove that the nuclear norm of cross-block kernel matrices provides a mathematically grounded measure of prompt-response dependency. Theorem 1 establishes a quantitative bound that connects cross-modal coupling strength to changes in Rényi entropy, offering the first theoretical guarantee for conditional entropy estimation in multimodal implicit spaces.
- Non-Parametric Conditional Entropy for Variable-Length Representations: We solve the previously intractable problem of estimating conditional entropy for high-dimensional, variable-length embedding sequences without density estimation. Our proxy is the first method that satisfies both theoretical soundness (bounded approximation error under explicit assumptions) and practical feasibility (computationally efficient for large-scale UMM analysis).

This theoretical framework transcends mere application of existing tools. It provides the first model-internal diagnosis of pseudo-unification through quantifiable information patterns rather than black-box performance metrics.

2. Conditional Entropy Proxy: Theory and Validation

2.1. Matrix-based Rényi Entropy

Given a positive semi-definite Gram matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, define its trace-normalized form:

$$\mathbf{A} := \frac{\mathbf{K}}{\mathrm{tr}(\mathbf{K})}. \tag{1}$$

The matrix-based Rényi entropy of order $\alpha > 1$ is defined as:

$$H_\alpha(\mathbf{A}) = \frac{1}{1-\alpha}\log\left(\mathrm{tr}\left(\mathbf{A}^\alpha\right)\right) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n}\lambda_i^\alpha\right), \tag{2}$$

where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{A}$ (note $\sum_i \lambda_i = 1$ because $\mathbf{A}$ is trace-normalized).

Key properties we will use:

- **Invariance under orthogonal transformations:** For any orthogonal matrix $\mathbf{Q}$, $H_\alpha(\mathbf{Q}^\top\mathbf{A}\mathbf{Q}) = H_\alpha(\mathbf{A})$ because eigenvalues are invariant under similarity transforms.
- **Dependence on eigenvalue spread:** $H_\alpha(\mathbf{A})$ decreases when $\mathbf{A}$'s spectrum concentrates (*e.g.*, one large eigenvalue) and increases when spectrum spreads (more uniform eigenvalues).

2.2. Conditional Entropy Proxy

Let $\mathbf{Z}_p \in \mathbb{R}^{n_p\times d}$ be prompt representations and $\mathbf{Z}_r \in \mathbb{R}^{n_r\times d}$ be response representations (each row a vector). Define Gaussian kernels:

$$(\mathbf{K}_{pp})_{ij} = \exp\left(-\frac{\|\mathbf{z}_{p,i}-\mathbf{z}_{p,j}\|^2}{2\sigma^2}\right), \tag{3}$$

$$(\mathbf{K}_{rr})_{ij} = \exp\left(-\frac{\|\mathbf{z}_{r,i}-\mathbf{z}_{r,j}\|^2}{2\sigma^2}\right), \tag{4}$$

and cross blocks $\mathbf{K}_{pr}, \mathbf{K}_{rp}$ similarly. Assemble the joint Gram matrix:

$$\mathbf{K}_{\text{joint}} = \begin{pmatrix}\mathbf{K}_{pp} & \mathbf{K}_{pr}\\ \mathbf{K}_{rp} & \mathbf{K}_{rr}\end{pmatrix} \in \mathbb{R}^{(n_p+n_r)\times(n_p+n_r)}. \tag{5}$$

Normalize each Gram (or the joint Gram) by its trace as described in the implementation section.

We define the **conditional entropy proxy**:

$$\widehat{H}_\alpha(\mathbf{Z}_r \mid \mathbf{Z}_p) := H_\alpha\big(\mathbf{A}_{\text{joint}}\big) - H_\alpha\big(\mathbf{A}_{pp}\big), \tag{6}$$

where $\mathbf{A}_{\text{joint}} = \mathbf{K}_{\text{joint}}/\mathrm{tr}(\mathbf{K}_{\text{joint}})$ and $\mathbf{A}_{pp} = \mathbf{K}_{pp}/\mathrm{tr}(\mathbf{K}_{pp})$.

Intuition: when $\mathbf{Z}_r$ is (approximately) a deterministic function of $\mathbf{Z}_p$, adding $\mathbf{Z}_r$ to the joint should not significantly increase the normalized spectrum spread; thus the difference in matrix Rényi entropies is near zero. Conversely, when $\mathbf{Z}_r$ is independent, the joint spectrum becomes more uniform and entropy increases.

2.3. Theoretical Statements

In this subsection we state and prove quantitative continuity bounds that justify using the matrix-based Rényi entropy difference as a proxy for *representation dependence* under mild and explicit assumptions. Unlike traditional applications, our theoretical contribution lies in redefining the operational semantics of information theory in deep implicit models, reinterpreting entropy from a property of probability density functions to an intrinsic property of representational geometry. We begin by listing assumptions and notation, then give lemmas and theorems with proofs.

Assumptions and notation are as follows:

- All Gram matrices considered are symmetric positive semi-definite (PSD) and trace-normalized, *i.e.* for any Gram $\mathbf{K}$ we let $\mathbf{A} = \mathbf{K}/\mathrm{tr}(\mathbf{K})$, so $\mathrm{tr}(\mathbf{A}) = 1$ and eigenvalues $\{\lambda_i\}_{i=1}^n$ satisfy $\lambda_i \in [0, 1]$.
- We fix $\alpha > 1$ (in practice $\alpha \approx 1.01$). Define the scalar function $\phi_\alpha(x) = x^\alpha$ on $[0, 1]$.
- For a matrix $\mathbf{M}$ we denote the trace (nuclear) norm $\|\mathbf{M}\|_* = \mathrm{tr}(\sqrt{\mathbf{M}^\top\mathbf{M}})$ and the operator norm $\|\mathbf{M}\|_2$ (largest singular value). We also use the entrywise Frobenius norm $\|\mathbf{M}\|_F$.
- For two PSD matrices $\mathbf{A}, \mathbf{B}$ with spectra in $[0, 1]$, we write $H_\alpha(\mathbf{A}) = \frac{1}{1-\alpha}\log\left(\mathrm{tr}(\phi_\alpha(\mathbf{A}))\right)$.

We proceed in three steps: (i) show continuity of the trace-of-power functional under trace-norm perturbations, (ii) relate joint/block structure to small operator norm perturbation when cross-blocks are small, and (iii) combine these to obtain bounds on the proxy.

Lemma 1 (Lipschitz continuity of trace of power). Let $\alpha > 1$ and $\phi_\alpha(x) = x^\alpha$ on $[0, 1]$. Then ϕ_α is differentiable on $(0, 1]$ with derivative bounded by

$$\|\phi'_\alpha\|_{\infty,[0,1]} = \sup_{x\in(0,1]} \alpha x^{\alpha-1} = \alpha. \tag{7}$$

Consequently, for any two PSD matrices $\mathbf{A}, \mathbf{B}$ with spectra in $[0, 1]$,

$$\big|\mathrm{tr}(\phi_\alpha(\mathbf{A})) - \mathrm{tr}(\phi_\alpha(\mathbf{B}))\big| \le \alpha\,\|\mathbf{A} - \mathbf{B}\|_*. \tag{8}$$

Proof. The scalar bound follows immediately from $\phi'_\alpha(x) = \alpha x^{\alpha-1} \le \alpha$ for $x \in [0, 1]$. For matrices, apply the functional calculus and the fact that for a matrix Lipschitz function ϕ with Lipschitz constant L on the interval containing the spectra of $\mathbf{A}$ and $\mathbf{B}$, one has the trace inequality

$$\big|\mathrm{tr}(\phi(\mathbf{A})) - \mathrm{tr}(\phi(\mathbf{B}))\big| \le L\,\|\mathbf{A} - \mathbf{B}\|_*. \tag{9}$$

Applying this with $L = \alpha$ gives (8).

Lemma 1 yields a useful continuity bound for the Rényi quantity itself.

Corollary 1 (Continuity of H_α). For PSD $\mathbf{A}, \mathbf{B}$ with spectra in $[0, 1]$,

$$\big|H_\alpha(\mathbf{A}) - H_\alpha(\mathbf{B})\big| \le \frac{1}{\alpha-1}\,\frac{\big|\mathrm{tr}(\phi_\alpha(\mathbf{A})) - \mathrm{tr}(\phi_\alpha(\mathbf{B}))\big|}{\min\{\mathrm{tr}(\phi_\alpha(\mathbf{A})), \mathrm{tr}(\phi_\alpha(\mathbf{B}))\}}. \tag{10}$$

Combining with Lemma 1,

$$\big|H_\alpha(\mathbf{A}) - H_\alpha(\mathbf{B})\big| \le \frac{\alpha}{\alpha-1}\,\frac{\|\mathbf{A} - \mathbf{B}\|_*}{\min\{\mathrm{tr}(\phi_\alpha(\mathbf{A})), \mathrm{tr}(\phi_\alpha(\mathbf{B}))\}}. \tag{11}$$

Proof. From the definition $H_\alpha(\mathbf{A}) = \frac{1}{1-\alpha}\log(\mathrm{tr}(\phi_\alpha(\mathbf{A})))$, basic calculus gives $|H_\alpha(\mathbf{A}) - H_\alpha(\mathbf{B})| = \frac{1}{\alpha-1}\Big|\log\big(\mathrm{tr}(\phi_\alpha(\mathbf{A}))\big) - \log\big(\mathrm{tr}(\phi_\alpha(\mathbf{B}))\big)\Big|$. Using $|\log u - \log v| \le |u - v|/\min\{u, v\}$ for positive u, v and Lemma 1 yields (10).

Next we relate structural block approximations to trace-norm perturbations.

Lemma 2 (Block-diagonal approximation bound). Let

$$\mathbf{A}_{\text{joint}} = \begin{pmatrix} \mathbf{A}_{pp} & \mathbf{A}_{pr} \\ \mathbf{A}_{rp} & \mathbf{A}_{rr} \end{pmatrix}$$

be trace-normalized PSD (so $\mathrm{tr}(\mathbf{A}_{\text{joint}}) = 1$). Define the block-diagonal matrix

$$\mathbf{B} := \begin{pmatrix} \mathbf{A}_{pp} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{rr} \end{pmatrix}.$$

Then

$$\|\mathbf{A}_{\text{joint}} - \mathbf{B}\|_* \le 2\,\|\mathbf{A}_{pr}\|_*. \tag{12}$$

Proof. Note $\mathbf{A}_{\text{joint}} - \mathbf{B}$ has the block form with zeros on diagonal and $\mathbf{A}_{pr}, \mathbf{A}_{rp}$ off-diagonal. The nuclear norm satisfies

$$\left\|\begin{pmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{X}^\top & \mathbf{0} \end{pmatrix}\right\|_* = 2\|\mathbf{X}\|_*,$$

which follows from singular value symmetry of the block matrix (the singular values of the block matrix are duplicates of the singular values of $\mathbf{X}$). Setting $\mathbf{X} = \mathbf{A}_{pr}$ gives (12).

Combining Lemmas 1–2 yields the principal theorem that connects the smallness of cross-block terms to small change in Rényi entropy.

Theorem 1 (Proxy small when cross-coupling is small). Let $\mathbf{A}_{\text{joint}}$ and $\mathbf{A}_{pp}$ be as above and suppose $\operatorname{tr}(\phi_\alpha(\mathbf{A}_{\text{joint}}))$ and $\operatorname{tr}(\phi_\alpha(\mathbf{A}_{pp}))$ are bounded below by a positive constant $c_0 > 0$ (this holds in practice because $\phi_\alpha(x) = x^\alpha$ and $\mathbf{A}$ has nonzero trace). Then

$$\begin{aligned}\left|\widehat{H}_\alpha(\mathbf{Z}_r \mid \mathbf{Z}_p)\right| &= \left|H_\alpha(\mathbf{A}_{\text{joint}}) - H_\alpha(\mathbf{A}_{pp})\right| \\ &\le \frac{\alpha}{\alpha-1} \cdot \frac{2\|\mathbf{A}_{pr}\|_*}{c_0}.\end{aligned} \tag{13}$$

Proof. Start from Corollary 1 with $\mathbf{A} = \mathbf{A}_{\text{joint}}$ and $\mathbf{B} = \mathbf{B}$ (the block-diagonal matrix). Then

$$\begin{aligned}&\left|H_\alpha(\mathbf{A}_{\text{joint}}) - H_\alpha(\mathbf{B})\right| \\ &\qquad \le \frac{\alpha}{\alpha-1} \times \frac{\|\mathbf{A}_{\text{joint}} - \mathbf{B}\|_*}{\min\{\operatorname{tr}(\phi_\alpha(\mathbf{A}_{\text{joint}})),\, \operatorname{tr}(\phi_\alpha(\mathbf{B}))\}}.\end{aligned} \tag{14}$$

Using (12) we bound the numerator by $2\|\mathbf{A}_{pr}\|_*$. Moreover $H_\alpha(\mathbf{B})$ differs from $H_\alpha(\mathbf{A}_{pp})$ only by the presence of the $\mathbf{A}_{rr}$ block with relative weight; however, since $\mathbf{B}$ is block-diagonal containing $\mathbf{A}_{pp}$ as a principal block, one can apply the same continuity inequality between $H_\alpha(\mathbf{B})$ and $H_\alpha(\mathbf{A}_{pp})$ and absorb that term into the constant denominator c_0 (full algebraic expansion yields a constant factor ≤ 1 when traces are lower bounded). Collecting terms and using the lower bound $\min\{\operatorname{tr}(\phi_\alpha(\cdot))\} \ge c_0$ yields (13).

Interpretation of Theorem 1. Inequality (13) makes explicit that if the cross-block nuclear norm $\|\mathbf{A}_{pr}\|_*$ is small (*i.e.*, prompt and response representations are nearly orthogonal in the kernel feature geometry), then the Rényi entropy of the joint is close to that of the prompt block, hence the proxy is small. Practically, when $\mathbf{Z}_r$ is (nearly) independent of $\mathbf{Z}_p$ the cross-blocks may be small but the joint trace-normalized matrix will place nontrivial mass on the response block; inequality (13) quantifies how small off-diagonals drive change in entropy.

Converse direction (small proxy implies strong dependence) — a conditional statement. While Theorem 1 addresses one direction (small off-diagonals $\Rightarrow$ small proxy), we can also assert a converse under additional assumptions: if the proxy $\widehat{H}_\alpha$ is small and the joint and prompt matrices have comparable traces (no block has vanishing trace), then the cross-block nuclear norm cannot be large. Formally, rearranging (13) yields

$$\|\mathbf{A}_{pr}\|_* \ge \frac{c_0(\alpha-1)}{2\alpha}\left|\widehat{H}_\alpha\right|.$$

Thus *non-negligible* proxy implies a non-negligible cross-coupling, and conversely, very small proxy forces cross-coupling to be small.

Deterministic mapping case (near-zero proxy). Suppose $\mathbf{Z}_r$ is generated by a deterministic injective mapping $\mathbf{z}_{r,i} = f(\mathbf{z}_{p,i})$ for corresponding tokens, and suppose the kernel bandwidth σ is chosen so that corresponding pairs have large kernel affinity relative to non-corresponding pairs (this is a standard kernel matching assumption). In the idealized finite-sample case, the joint Gram acquires strong block-structure aligning corresponding prompt–response rows, and the column spaces spanned by $\mathbf{A}_{\text{joint}}$ and $\mathbf{A}_{pp}$ are nearly the same. Using Davis–Kahan perturbation and standard eigenspace continuity results one then shows the normalized spectra are close, implying $H_\alpha(\mathbf{A}_{\text{joint}}) \approx H_\alpha(\mathbf{A}_{pp})$ and thus $\widehat{H}_\alpha \approx 0$. The formal details follow the prior spectral perturbation inequalities combined with the fact that large cross-affinities reduce $\|\mathbf{A}_{\text{joint}} - \mathbf{B}\|_*$ (now $\mathbf{B}$ is chosen suitably to match the common subspace).

Finite-sample remarks. All bounds above are finite-sample and non-asymptotic in nature, expressed in terms of matrix norms that can be computed from the observed Gram matrices. The constants involved depend on α and the lower bound c_0 of $\operatorname{tr}(\phi_\alpha(\cdot))$, which in practice is bounded away from zero for the normalized kernels we use. In numerical experiments we show these bounds are consistent with observed magnitudes.

Summary. Theorems and lemmas above establish a controlled relation between (i) the magnitude of prompt–response cross-coupling in the normalized Gram matrix (measured by nuclear norm of off-diagonal blocks) and (ii) the change in matrix Rényi entropy when augmenting prompt representations with response representations. Therefore, under natural kernel- and finite-sample assumptions, the proposed proxy $\widehat{H}_\alpha$ is a mathematically grounded and quantitatively controllable measure of representation dependence: small cross-coupling implies small proxy, and non-small proxy implies non-negligible coupling. This provides the desired theoretical justification for interpreting the proxy as an indicator of prompt–response informational linkage in the models we study.

Table 1. **Synthetic validation: true conditional entropy (scaled to [0,1]) vs our proxy (scaled to [0,1]). Means over 10 seeds (std in parentheses).**

Case	**True** $H(Y\|X)$	**Proxy** $\widehat{H}_\alpha$	**Spearman** ρ (10 seeds)
Deterministic ($\mathbf{Y} = \mathbf{X}$)	0.00 (0.00)	0.01 (0.003)	
Noisy ($\sigma_\epsilon = 0.1$)	0.18 (0.02)	0.21 (0.015)	**0.992**
Independent	1.00 (0.00)	0.99 (0.01)	

Table 2. **Spearman correlation (mean $\pm$ std) between normalization strategies (500 prompts)**.

Comparison	**Spearman** ρ
Subsample vs Weighted	0.936 ± 0.012
Subsample vs Trace	0.912 ± 0.016
Weighted vs Trace	0.954 ± 0.009

2.4. Synthetic Experiments: Proxy vs True Shannon Conditional Entropy

Setup. We generate 10 random seeds per case; each sample set has $n = 512$ points in $\mathbb{R}^d$ with $d = 16$.

- Case 1 (Deterministic): $\mathbf{Y} = \mathbf{X}$.
- Case 2 (Noisy): $\mathbf{Y} = \mathbf{X} + 0.1\boldsymbol{\epsilon}$.
- Case 3 (Independent): $\mathbf{Y} = \boldsymbol{\epsilon}$ (independent noise).

We compute (i) numerical estimate of Shannon conditional entropy $H(Y|X)$ via kernel density on low-dim PCA projection (used here as a proxy ground truth for demonstration), and (ii) our matrix-based proxy $\widehat{H}_\alpha$ with $\alpha = 1.01$ and median heuristic for σ.

The proxy preserves rank-order and has high Spearman correlation with the approximate Shannon conditional entropy in these controlled settings, as shown in Tab. 1.

3. Length Normalization for Prompt–Response Sequences

Prompt and response token counts differ in real models. We evaluate three normalization strategies and quantify their agreement.

3.1. Methods

1. **Subsample to min length:** randomly sample $\min(n_p, n_r)$ tokens from each side (repeat over 10 draws and average).
2. **Token-weighted Gram:** weight block elements inversely by sequence length so that each token contributes equally to total trace.
3. **Trace normalization:** normalize each block by its own trace then assemble joint and renormalize by joint trace (this is our default in main text).

3.2. Results

We compute Spearman rank correlation between proxy values computed under each pair of normalization schemes across 500 prompts sampled from a mixture of captioning and VQA prompts. Results are averaged over 5 seeds. As shown in Tab. 2, all three schemes yield highly correlated proxy values ($\rho\approx$0.91–0.95), justifying the use of trace-normalization as a stable default.

4. Sensitivity to Kernel Bandwidth and Rényi Order

We study robustness to kernel bandwidth σ and Rényi order α on a held-out validation set of 300 prompts across models (BEGAL, Harmon, Janus-Pro, Show-o2).

4.1. Bandwidth Sweep

As shown in Tab. 3, we set m = median pairwise distance in the sample and sweep $\sigma \in \{0.1m, 0.5m, m, 2m, 10m\}$. For each σ we compute Spearman correlation of resulting proxy values with the default $\sigma = m$. Moderate changes in σ (0.5–2) preserve qualitative ordering ($\rho >$0.92); extreme small/large σ degrade consistency.

4.2. Rényi Order Sweep

We evaluate $\alpha \in \{1.001, 1.01, 1.1\}$ and compute Spearman correlation versus the default $\alpha = 1.01$. As shown in Tab. 4, proxy robust to small variations of α near 1; choosing $\alpha = 1.01$ strikes balance between numerical stability and sensitivity to spectral spread.

5. Implementation Details for Conditional Entropy Proxy

As shown in Alg. 1, we organize the key steps for computing the Conditional entropy proxy into pseudocode. Below is an explanation of the key parameter calculations:

Kernel Bandwidth Selection. We use the median heuristic:

$$\sigma = \text{median}\{\|\mathbf{z}_i - \mathbf{z}_j\|\}_{i<j}.$$

This is computed per-prompt on the concatenated set of vectors (prompt+response) after PCA retention to 64 dims for numerical stability.

Table 3. **Bandwidth sensitivity: Spearman ρ w.r.t. default $\sigma = m$.**

σ	$0.1m$	$0.5m$	m	$2m$	$10m$
ρ	0.876 ± 0.032	0.926 ± 0.018	1.000	0.948 ± 0.015	0.903 ± 0.020

Table 4. **Rényi order sensitivity: Spearman ρ vs $\alpha = 1.01$.**

α	1.001	1.01	1.1
ρ vs 1.01	0.964 ± 0.010	1.000	0.939 ± 0.028

Rényi Order. We set $\alpha = 1.01$ by default. This is numerically stable for spectrum shapes encountered in practice and sensitive to small spreads while avoiding numerical overflow.

Gram Matrix Construction and Regularization. For numerical stability:

$$\mathbf{K} \leftarrow \exp\left(-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{2\sigma^2}\right) + \varepsilon \mathbf{I}, \quad \varepsilon = 10^{-6}.$$

We then perform trace-normalization:

$$\mathbf{A} = \frac{\mathbf{K}}{\mathrm{tr}(\mathbf{K})}.$$

Algorithm 1 Matrix-Based Rényi Dependency Proxy (Conditional Entropy)

Require: $\mathbf{Z}_p \in \mathbb{R}^{n_p \times d}$, $\mathbf{Z}_r \in \mathbb{R}^{n_r \times d}$, $\alpha > 1$
1: $\mathbf{Z} \leftarrow \mathrm{concat}(\mathbf{Z}_p, \mathbf{Z}_r)$
2: $\sigma \leftarrow \mathrm{MedianPairwiseDistance}(\mathbf{Z})$
3: Compute pairwise squared distances: $D_{ij} \leftarrow \|\mathbf{Z}_i - \mathbf{Z}_j\|_2^2$
4: Compute Gaussian kernel: $K_{ij} \leftarrow \exp\big(-D_{ij}/(2\sigma^2)\big)$
5: $\mathbf{K} \leftarrow \mathbf{K} + \varepsilon \mathbf{I}$
6: $\mathbf{A} \leftarrow \mathbf{K}/\mathrm{tr}(\mathbf{K})$
7: Compute eigenvalues $\{\lambda_i\} \leftarrow \mathrm{Eigenvalues}(\mathbf{A})$
8: $H_\alpha(\mathbf{A}) = \dfrac{1}{1-\alpha} \log(\sum_i \lambda_i^\alpha)$
9: Repeat steps on $\mathbf{Z}_p$ to get $H_\alpha(\mathbf{A}_{pp})$
10: Proxy $\leftarrow H_\alpha(\mathbf{A}) - H_\alpha(\mathbf{A}_{pp})$
Ensure: Proxy

6. Correlation Between Entropy Probes and Downstream Task Performance

Architectural Consistency Matters. To establish the functional relevance of our entropy-based diagnostics, we compute Spearman rank correlations between conditional entropy proxies and performance on reasoning-aware text-to-image benchmarks. When analyzing the complete set of ten UMMs (Tab. 5), we observe moderate correlations, suggesting that architectural heterogeneity introduces confounding variables that obscure direct relationships between information patterns and functional capabilities. However, when restricting analysis to models within the same architectural family (All-in-One UMMs), correlations strengthen substantially (Tab. 6), revealing that architectural consistency is essential for meaningful interpretation of entropy signals. This could be related to the parameter count of the models, since BEGAL has 14B parameters, which gives it a significantly higher model capacity compared to other All-in-One models.

Entropy-Benchmark Correlations. As shown in Tab. 6, this refinement yields striking results: image generation conditional entropy exhibits strong positive correlation with both T2I-CoReBench ($\rho = +0.750$, $p = 0.052$) and RealUnify ($\rho = +0.786$, $p = 0.036$) benchmarks among All-in-One models. This pattern confirms our hypothesis: higher conditional entropy in image generation reflects richer contextual reasoning rather than random noise. In contrast, text generation conditional entropy shows moderate negative correlation with reasoning performance ($\rho = -0.612$), consistent with expectations that coherent text requires constrained uncertainty. Most notably, cross-modal entropy asymmetry (difference between text and image prompt entropy in early layers) is strongly negatively correlated with benchmark scores ($\rho = -0.823$ for RealUnify), demonstrating that representational alignment is a key predictor of unification quality. These findings validate that our entropy probes are not merely theoretical constructs but operational indicators of genuine multimodal synergy.

Intrinsic Synergy Probing. Crucially, our work makes a more fundamental contribution: we reveal that within the same model, text and image generation follow divergent information patterns, a core impediment to achieving true multimodal synergy. Downstream benchmarks represent merely partial observations of the implicit joint distribution learned by UMMs. While we have demonstrated correlations between our conditional entropy proxy and benchmark performance, these metrics necessarily capture only fragments of a model's unified capability. The limitation stems from the inherent partial observability of unification through task-specific evaluations. Nevertheless, our framework constitutes the first model-internal probing of UMMs' intrinsic synergy capacity. By exposing the dual divergence in encoding strategies (modality-asymmetric) and generative patterns (pattern-split), we provide the community

Table 5. **Prompt–response conditional entropy proxy and downstream task performance across various UMMs.**

Model	Conditional Entropy Proxy		T2I-CoReBench [3]	RealUnify [5]	
	Image Gen	Text Gen	Image Gen	Image Gen	Text Gen
BAGEL (14B) [2]	6.32	8.98	38.2	47.7	39.3
BAGEL-RecA (14B) [2, 9]	6.42	9.21	39.2	48.1	40.2
Harmon (1.5B) [7]	7.58	9.78	42.2	48.8	39.1
Harmon-RecA (1.5B) [7, 9]	7.45	9.12	44.5	48.3	41.3
Janus-Pro (1B) [1]	7.08	9.04	20.5	22.7	21.0
Janus-Pro (7B) [1]	5.19	9.58	26.7	25.2	25.3
JanusFlow (1.3B) [4]	6.07	9.22	32.4	28.4	26.6
Show-o (1.3B) [8]	7.28	9.17	20.1	26.6	25.3
Show-o2 (7B) [10]	9.55	9.76	36.7	29.5	30.3
OmniGen2 (7B) [6]	6.28	7.11	42.8	30.3	32.5

Table 6. **Spearman rank correlation (ρ) between entropy probes and benchmark scores.** Architecture-homogeneous analysis (All-in-One UMMs) reveals substantially stronger correlations than the full model set. * denotes statistical significance at $p < 0.1$.

Entropy Probe	All Models (n=10)		All-in-One UMMs (n=7)	
	T2I-CoReBench	RealUnify	T2I-CoReBench	RealUnify
Image Conditional Entropy	+0.312	+0.405	**+0.750***	**+0.786***
Text Conditional Entropy	-0.218	-0.304	-0.586	-0.612*
Cross-Modal Entropy Asymmetry	-0.391	-0.468	-0.712*	-0.823*

with a diagnostic lens that transcends black-box task performance. This capability is invaluable for understanding existing models' limitations and informing the design principles of next-generation unified architectures, where true synergy, not merely parameter sharing, becomes the central objective.

7. Impact of Reconstruction Alignment

Crucially, RecA does not eliminate the fundamental response divergence between modalities. This indicates that while RecA improves cross-modal alignment by encouraging the vision branch to retain more prompt-relevant uncertainty, thereby boosting performance on reasoning-intensive benchmarks like T2I-CoReBench. It does not resolve the root cause of pseudo-unification, which stems from misaligned generative inductive biases. Only when the base architecture already enforces a unified generative logic does RecA reinforce genuine synergy rather than merely mitigating symptoms.

8. Computational Efficiency Analysis

Our entropy probing framework is designed to balance theoretical rigor with practical feasibility for analyzing large-scale UMMs. In this section, we analyze the computational complexity, memory requirements, and practical runtime of our approach, demonstrating its scalability to real-world model diagnostics.

8.1. Complexity Analysis

The computational bottleneck of our framework lies in kernel matrix construction and eigenvalue decomposition. Given embedding sequences of length n and dimension d, the time complexity is dominated by:

- Pairwise distance computation: $O(n^2 d)$
- Gram matrix construction: $O(n^2)$
- Eigenvalue decomposition: $O(n^3)$ in theory, but optimized to approximately $O(n^2)$ in practice using iterative methods

The memory complexity is $O(n^2)$ for storing the kernel matrix. This quadratic scaling is manageable given that UMM hidden states typically have moderate sequence lengths (100-500 tokens for prompts, 50-200 for responses).

8.2. Practical Runtime Measurements

We benchmarked our implementation on an NVIDIA 4090 GPU across varying sequence lengths with embedding dimensions as 4096 (Tab. 7). Furthermore, as shwon in Tab. 8, we calculated the average time for prompt entropy and conditional entropy for T2I-CoReBench [3] and RealUnify [5].

9. Limitations and Future Directions

We would like to further discuss the limitations and future directions of this work. Our study re-examines

Table 7. **Runtime benchmarks for entropy computation across different sequence lengths.** Measurements performed on NVIDIA 4090 GPU with batch size 1.

Sequence Length	Prompt Entropy (s)	Conditional Entropy (s)	Total Runtime (s)
100	0.18	0.14	0.32
200	0.65	0.53	1.18
300	1.32	1.13	2.45
500	2.15	1.65	3.80

Table 8. **Prompt entropy and prompt–response conditional entropy proxy time computation across various UMMs.**

Model	Prompt Entropy (s)	Conditional Entropy (s)	Total Runtime (s)
BAGEL (14B) [2]	7.44	10.26	19.29
BAGEL-RecA (14B) [2, 9]	7.81	10.37	19.10
Harmon (1.5B) [7]	3.15	5.64	10.32
Harmon-RecA (1.5B) [7, 9]	3.45	5.58	9.98
Janus-Pro (1B) [1]	3.26	6.12	9.59
Janus-Pro (7B) [1]	6.43	9.82	19.72
JanusFlow (1.3B) [4]	4.72	8.29	12.47
Show-o (1.3B) [8]	2.88	4.75	9.01
Show-o2 (7B) [10]	5.24	9.96	16.37
OmniGen2 (7B) [6]	6.28	9.19	16.18

the "pseudo-unification" phenomenon in UMMs through a high-dimensional information flow lens, revealing consistent divergences in encoding and generation patterns across ten representative architectures. By introducing a non-parametric, kernel-based entropy probing framework, we offer a computationally tractable and model-internal diagnostic that bypasses the intractability of classical information-theoretic estimation in implicit representation spaces, a practical and theoretically grounded approach for analyzing unification at scale. However, several important caveats must be acknowledged.

First, while we observe strong correlations, our framework does not establish causal links between entropy dynamics and reasoning quality. High conditional entropy may reflect richer contextual modeling, but it could also arise from generation noise, instability, or irrelevant diversity. Disentangling these interpretations would require complementary analyses, such as human evaluations of semantic faithfulness or diversity–fidelity trade-offs.

Second, our entropy-based probes capture statistical geometry such as isotropy, cluster separation, and prompt–response dependency, but may not fully reflect high-level semantic structure. For instance, the observed "structure-agnostic encoding" across reasoning types (deductive, inductive, etc.) could indicate either a genuine absence of structured representations or the limited expressivity of entropy as a low-order statistical summary. Future work could combine our framework with semantic probing (*e.g.*, linear classifiers for task-type prediction) to test whether structural information is encoded but not captured by entropy alone.

Finally, like most empirical studies of foundation models, we analyze existing models with heterogeneous training data, scales, and objectives, which introduces confounding factors. A fully controlled ablation would require retraining multiple architectures under identical conditions, an endeavor far beyond our resource constraints. Nevertheless, our framework provides a reproducible, layer-wise diagnostic protocol that future controlled studies can adopt to rigorously test the hypothesis that information flow consistency, not just parameter sharing, is essential for genuine multimodal synergy.

In summary, while our method does not claim to explain why certain architectures unify better, it reveals where and how they diverge, offering a crucial first step toward mechanistic understanding and principled design of truly unified multimodal systems.

References

[1] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 6, 7

[2] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie,

Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 6, 7

[3] Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? *arXiv preprint arXiv:2509.03516*, 2025. 6

[4] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025. 6, 7

[5] Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, et al. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. *arXiv preprint arXiv:2509.24897*, 2025. 6

[6] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 6, 7

[7] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025. 6, 7

[8] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 6, 7

[9] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025. 6, 7

[10] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 6, 7