Strong Helps Weak: Directional Cross-Modal Alignment Transfer in Multi-modal LLMs

Hoigi Seo[1*] **Byung Hyun Lee**[1*] **Minjun Kim**[1*]
Dohyun Mah[1] **Jongho Lee**[1,2] **Se Young Chun**[1,2,3†]
[1]Dept. of ECE [2]INMC & AIIS [3]IPAI, Seoul National University, Republic of Korea
{seohoiki3215, ldlqudgus756, minjoony, dohyun0505}
{jonghoyi, sychun}@snu.ac.kr

Abstract

Multi-modal large language models (MLLMs) achieve strong modality understanding by pairing a large language model (LLM) with an encoder for a target modality such as vision, video, or audio. However, improving an MLLM's capability for a given modality typically requires additional training on large modality-specific datasets, incurring substantial data collection and compute costs. Model merging offers an alternative, but it is often infeasible for data-scarce, large per-sample size, or domain-specific modalities (*e.g.*, audio and video), where same-modality model variants are rarely available. In this work, we characterize an intriguing asymmetric phenomenon: merging a well-aligned, data-rich source-modality MLLM into a data-scarce target-modality MLLM substantially improves the target on its own benchmarks. Our theoretical and empirical analyses show that this gain stems from enhanced alignment between modality-specific and textual tokens, induced by the stronger donor modality. Specifically, we derive a mutual-information lower bound that is monotonic in alignment-related quantities and strongly correlated with downstream MLLM performance. Building on this principle, we propose Directional Cross-modal Alignment Transfer (DCAT), a novel framework that transfers textual alignment from a strong, well-aligned source (donor) modality to a weak target (recipient) modality, boosting target-modality performance without further fine-tuning. We further show that the alignment-enhancing objective admits a closed-form weight-space solution computed from only a small calibration set. DCAT outperforms existing model-merging methods, offering an efficient path toward cross-modal alignment transfer. Project page with code is available at `https://seohoiki3215.github.io/DCAT_project_page`

1 Introduction

Large language models (LLMs) [3, 41, 54, 67, 70] achieve strong performance across textual tasks through scaling in data and model size. Most existing multi-modal LLMs (MLLMs) [12, 42, 81] have advanced visual reasoning, video understanding, or audio recognition by pairing a strong LLM backbone with a single modality encoder such as vision [52, 68, 88], video [35, 39, 46] or audio [53, 59, 62] encoders, and a projector [36, 42] that aligns the modality with the textual space. However, enhancing the target-modality capability of an MLLM typically requires additional training on large modality-specific data, incurring substantial data collection and compute costs.

Model merging [9, 17, 28, 48, 75, 82] offers a promising alternative for improving MLLMs by combining same-architecture models with little data. For vision MLLMs, multiple same-backbone variants and abundant modality data are often available and can be merged to enhance visual capabilities.

* equal contribution, † corresponding author.

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

However, such data abundance and same-modality variants are much scarcer for other domain-specific modalities, where training multiple variants is costly or infeasible. While existing studies have mostly focused on merging within a single modality, merging different-modality MLLMs remains largely unexplored. We argue that MLLMs trained for different modalities may still share the same base LLM, making cross-modality weight merging possible. Thus, we ask a more targeted question: can a source-modality MLLM (donor) transfer useful alignment to a target-modality MLLM (recipient), thereby improving the recipient's target-modality performance without additional training?

In this paper, we identify and study an intriguing and recurring phenomenon: merging a donor and a recipient via existing merging methods can yield substantial performance gains over the original recipient on its associated benchmark. We investigate this phenomenon along two complementary axes. On the theoretical side, we analyze it through the lens of mutual information (MI) [31, 61]. Under a mild assumption, we prove that a lower bound on the MI between text and the other modality is monotonic in the product of two quantities: the Subspace Energy Overlap (SEO), which measures the extent to which each modality's token energy lies within the text-aligned subspace, and the Spectral Diversity (SD) [56], which measures how uniformly that energy is distributed within this subspace. We refer to this product as the *Alignment* score. On the empirical side, across hundreds of model checkpoints spanning multiple architectures, we conduct a series of statistically tested analyses and find that the Alignment score is strongly and significantly correlated with MLLM performance, and that merged models exhibiting performance gains also show a statistically significant increase in the Alignment. Taken together, these theoretical and empirical results support our MI-lower-bound-based Alignment score, as a plausible explanatory mechanism for the observed benefits of model merging.

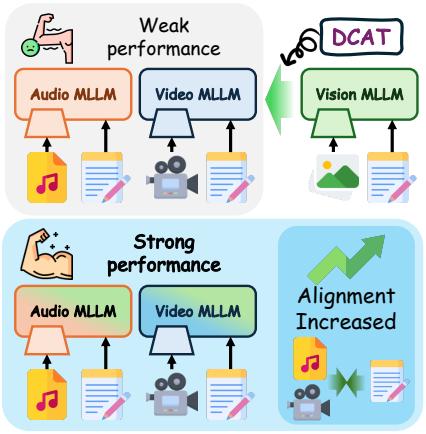

Figure 1: **Conceptual diagram of DCAT.** DCAT transfers alignment from a strong donor modality model to a relatively weak recipient modality model, thereby improving the recipient model's modality-specific performance.

Building on these analyses, we propose Directional Cross-modal Alignment Transfer (DCAT), a novel framework that transfers alignment from a high-alignment donor MLLM to a recipient MLLM using only a few hundred calibration samples, thereby improving the recipient's target-modality performance. To our knowledge, we are the first to formulate and analyze this phenomenon; accordingly, the most natural baselines for DCAT are the existing model-merging methods through which the phenomenon is initially observed. As illustrated in Fig. 1, we instantiate this directional transfer by using vision MLLMs, which generally exhibit high alignment, as donors and audio or video MLLMs as recipients. Across a diverse suite of benchmarks, including MMAU [69], AIR-Bench [84], Video-MME [16], and MVBench [37], DCAT achieves 29.55% improvement for audio tasks and 24.92% for video tasks on average over the original recipient models.

Our contributions can be summarized as follows:

- We investigate an asymmetric phenomenon: model merging between MLLMs trained on different modalities improves target-modality performance. We introduce a theory-motivated alignment metric that explains this phenomenon, supported by strong empirical analysis.
- Building on this analysis, we propose Directional Cross-modal Alignment Transfer (DCAT), a method that leverages a small amount of data to perform a closed-form weight update, thereby transferring alignment from a well-aligned modality to a relatively weaker one.
- We conduct extensive experiments across multiple MLLMs on various benchmarks. DCAT consistently outperforms existing model merging methods and achieves average performance gains of 29.55% over the base model on audio tasks and 24.92% on video tasks.

2 Related Work

Multi-modal large language models. Recent advances in LLMs have spurred extensive work on MLLMs, which attach modality-specific encoders and projectors to an LLM to generate text

Table 1: **Quantitative comparison between modality-specific models and merged models.** We report **(Left)** the MMAU [69] performance obtained by merging an audio-only model with an audio-vision merged model (A+V), and **(Right)** the Video-MME [16] performance obtained by merging a video-only model with a video-vision merged model (Vd+V). Results are reported for Task Arithmetic (TA), ISO-C (ISO), TSV, and OptMerge (Opt.). In all cases, merging is applied only to the LLM and the Alignment score was measured at the 21st layer of the Vicuna model. We set $\gamma = 0.8$.

	Music	Sound	Speech	Overall	Align_γ
Audio only	54.79	60.66	31.23	48.90	0.0091
A+V (TA)	55.69	64.56	39.34	53.20	0.0103
A+V (ISO)	58.68	66.67	34.83	53.40	0.0950
A+V (TSV)	55.99	66.67	35.44	52.70	0.0100
A+V (Opt.)	59.58	64.26	35.44	53.10	0.0099

	Short	Medium	Long	Overall	Align_γ
Video only	39.89	34.22	30.89	35.00	0.0248
Vd+V (TA)	44.67	34.78	31.89	37.11	0.0287
Vd+V (ISO)	43.67	38.22	33.44	38.44	0.0269
Vd+V (TSV)	46.44	36.67	34.00	39.04	0.0278
Vd+V (Opt.)	44.33	37.00	33.56	38.30	0.0267

from inputs such as vision [2, 4, 7, 34, 35, 39, 42, 46, 60, 73, 93] and audio [11, 12, 23, 24, 26, 80]. Training these models is compute-intensive, often requiring optimization of the LLM together with newly added components. Moreover, modality-specific data, such as video, requires far more storage than text and is costly or difficult to collect, making modality-specific performance gains expensive. These challenges are especially severe for rare or high-volume modalities with limited supervision, motivating training-free and data-efficient alternatives for improving MLLM capabilities.

Model merging. As neural networks scale, deploying separate task specialists becomes prohibitive, motivating model merging as a training-free way to consolidate independently fine-tuned models. TA [28] enables model editing via task vectors, defined as weight differences between fine-tuned and pretrained models. Subsequent methods mitigate interference among task-specific updates: TIES [82] trims redundant changes and resolves sign conflicts; ISO [48] separates task-specific subspaces; TSV [17] decomposes updates into singular directions for low-dimensional merging; and WUDI [9] identifies interfering vectors for robust merging. OptMerge [75] extends this paradigm across heterogeneous tasks and modalities. However, existing studies have not systematically examined cross-modal merging from a directional transfer perspective, where a source-modality MLLM is used to improve a designated target-modality MLLM on its own modality-specific benchmarks.

3 Analysis

3.1 Model merging increases modality performance

While task-wise model merging, which combines LLMs fine-tuned on different downstream tasks into one model, has been extensively studied, comparatively little attention has been paid to merging MLLMs trained for different input modalities. Tab. 1 reports MMAU [69] scores for vision–audio MLLM merges and Video-MME [16] scores for vision–video MLLM merges. All checkpoints are from OptMerge [75] which are Vicuna [10]-based. Despite modality heterogeneity, merging with a vision MLLM improves audio and video benchmark performance by up to 18.45% over the original modality-specific MLLM. This raises a natural question: what drives this striking performance gain?

3.2 On the lower bound of Mutual Information

We interpret this phenomenon through the lens of the mutual information (MI) between text and the other modality since it provides a principled way to quantify how much modality-specific information is accessible from text representations. As a first step toward making this perspective concrete, we introduce a quantity that we call the Subspace Energy Overlap (SEO), defined as follows.

Definition 1 (Subspace Energy Overlap)**.** *Let $Y_{\text{text}} \in \mathbb{R}^{N_t \times d}$ denote the text-token activation matrix, and let $U_k \in \mathbb{R}^{k \times d}$ be the top-k right singular vectors of the centered text activations. For modality-token activations $Y_{\text{mod}} = \{y_i\}_{i=1}^{N_m}$, the* Subspace Energy Overlap *(SEO) is defined as*

$$\textit{SEO} = \frac{\|Y_{\text{mod}} U_k^\top\|_F^2}{\|Y_{\text{mod}}\|_F^2}. \tag{1}$$

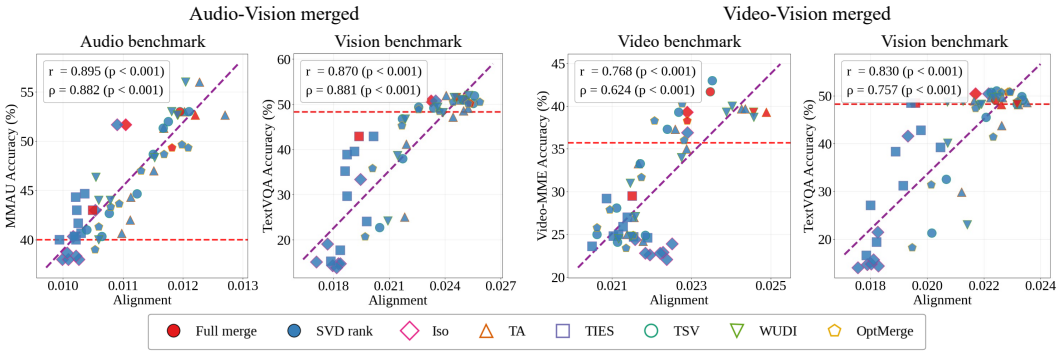

Figure 2: **Empirical analysis of our alignment metric and performance.** We analyze the correlation between the Alignment score and modality performance across a total of $N = 114$ model checkpoints. We use MMAU [69] for audio, Video-MME [16] for video, and TextVQA [63] for vision performance. Checkpoints are obtained from a range of merging methods [9, 17, 28, 48, 75, 82], combined with several variations of the merged weights. The red dotted line refers to the performance of original model. For all experiments, we chose $k = 32$ and $\gamma = 0.8$, and measure performance on 300 questions per benchmark across categories. Text embeddings are extracted from the questions.

SEO quantifies the fraction of modality-token energy captured by the principal text subspace, with higher values indicating that modality representations are more concentrated within this subspace.

Next, we introduce a second quantity, which we call the Spectral Diversity (SD), defined as follows.

Definition 2 (Spectral Diversity)**.** *Let $Z = Y_{\text{mod}} U_k^\top \in \mathbb{R}^{N_m \times k}$ and Σ_Z be the empirical covariance matrix of the projected representations. Denote the eigenvalues of Σ_Z by $\{\lambda_j\}_{j=1}^k$, and define $p_j = \frac{\lambda_j}{\sum_{\ell=1}^k \lambda_\ell}$. The* Spectral Diversity *(SD) is defined as*

$$SD = \frac{1}{k} \exp\left(-\sum_{j=1}^{k} p_j \log p_j \right). \tag{2}$$

Intuitively, $k \cdot \text{SD}$ corresponds to the effective rank [56] in text-subspace for the projected modality representation. Thus, the effective-rank view makes SD directly interpretable in the text-subspace: higher SD indicates that modality information is broadly distributed within that subspace. Then, the following theorem lower-bounds the mutual information between modality tokens and text.

Theorem 1 (Weighted Alignment-MI Lower Bound)**.** *Let $z = yU_k^\top$ where U_k contains the top-k right singular vectors of the centered text activation matrix. Suppose that $z \sim \mathcal{N}(\mu_Z, \Sigma_Z)$, $z|x \sim \mathcal{N}(\mu_z(x), \Sigma_{z|x})$, and $p_j \geq \frac{\delta}{k}$ for some sufficiently large $\delta \in (0, 1]$. Then, for any $\gamma \in (0, 1)$:*

$$I(y; x) \;\geq\; \kappa(\gamma) \cdot \log\big(SEO^\gamma \cdot SD^{1-\gamma}\big) + C_\delta, \quad \kappa(\gamma) = \frac{k}{2\min(\gamma, 1-\gamma)} \tag{3}$$

with $C_\delta = \frac{k}{2} \log(tr(\Sigma_y)/k) - \frac{1}{2} \log\det(\bar{\Sigma}_{z|x}) + \frac{k}{2} \log \delta$.

The proof of Theorem 1 is in Sec. B.1 and we empirically verify that δ is not vanishingly small in Sec. B.5. Although the lower bound in this theorem is maximized at $\gamma = 0.5$, the most suitable value of γ may vary across models and modalities; we therefore consider a more flexible, looser bound.

It is widely recognized that increasing the MI between modalities is essential for improving performance in multi-modal models [1, 51, 65]. This implies that maximizing the lower bound established in Theorem 1 provides a principled means of enhancing performance. Motivated by the bound in Eq. 3, we define the Alignment score as a weighted geometric mean of SEO and SD:

$$\text{Align}_\gamma := \text{SEO}^\gamma \cdot \text{SD}^{1-\gamma}. \tag{4}$$

That is, it measures how much modality-token energy enters the text principal subspace (SEO) and how broadly that energy is distributed across the effective directions of the text subspace (SD). In the following section, we leverage the proposed Align_γ to analyze the effect of various weight choices and present a statistically grounded empirical study examining whether the performance gains from model merging can be attributed to an increase in the MI lower bound for each modality.

3.3 Empirical analysis on Alignment score

We measure the Align_γ for each model before and after merging, with the Align. column of Tab. 1. The alignment increases significantly under model merging ($p \ll 10^{-10}$, Wilcoxon signed-rank test), suggesting that merging may improve performance by raising the MI lower bound.

To test whether this relationship is global, we evaluate 114 Vicuna-based MLLM checkpoints generated by multiple merging methods, low-rank approximations of weight differences, and module-restricted merging (*e.g.*, attention or FFN). For each checkpoint, we plot modality-wise performance against Align_γ at the 21st layer (Fig. 2), where Pearson (r) and Spearman (ρ) coefficients show significant ($p < 10^{-3}$) strong positive correlations across all modalities. Audio and video models improve clearly, while the vision model shows only marginal improvement or degradation; thus, we refer to relatively degraded models as *donor models* and improved models as *recipient models*.

We selected the 21st layer, the mid-to-late layer, to see this trend. In Fig. 3, we plot the Pearson coefficient between Align_γ and performance at each layer. Beyond the 19th layer, we consistently observe a high correlation across all modality merge pairs and performance. This finding is consistent with prior works [22, 29, 60, 64, 71, 79, 85, 86] showing that MLLMs exhibit high, meaningful attention from the middle layers onward, which has been interpreted as the point where alignment between text and the corresponding modality begins. In other words, the phenomenon captured here parallels observations made in earlier studies using different metrics, suggesting that our Align_γ metric not only functions as a lower bound on MI but also quantifies the alignment between textual and non-textual modalities. To ensure robustness, we adopt the 21st layer in our analysis, as it exhibits the lowest standard deviation. Experimental details and broader analysis of alignment across different models and benchmarks are provided in the Appendix Sec. D and Sec. E.

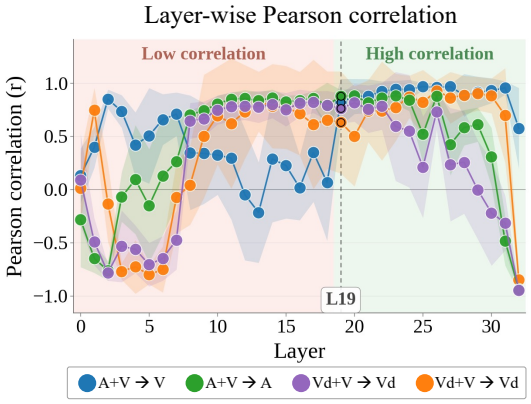

Figure 3: **Layer-wise Pearson correlation between Alignment score and performance across MLLM combinations and modalities.** We measure alignment in a layer-wise manner for each modality across multiple modality pairs. A consistently strong positive correlation between the Alignment score and performance emerges from the middle layers onward.

4 Method

Building on these analyses, we propose **Directional Cross-modal Alignment Transfer (DCAT)**, which transfers donor alignment to recipient models via a closed-form solution using minimal calibration inputs. DCAT exploits cross-modal merging benefits by targeting the alignment-related terms in Eq. (3). We further incorporate C_δ to preserve the overall modality activation scale.

Notation. Specifically, let θ_0 be the pretrained base weight (*e.g.*, base LLM weight) for a given layer, and let $\tau_r = \theta_r - \theta_0$ and $\tau_d = \theta_d - \theta_0$ denote the recipient and donor task vectors. We denote recipient-modality calibration inputs and text calibration inputs as X_{mod} and X_{text}, respectively. For a candidate task vector τ, the corresponding modality-token and text-token activations are given by $X_{\text{mod}}(\theta_0 + \tau)$ and $X_{\text{text}}(\theta_0 + \tau)$. Then recipient activations are computed as:

$$Y_{\text{text}} = X_{\text{text}}(\theta_0 + \tau_r), \quad Y_{\text{mod}} = X_{\text{mod}}(\theta_0 + \tau_r). \tag{5}$$

We also define the orthogonal projector of U_k and its orthogonal complement projector as $P_k (= U_k^\top U_k)$ and $P_k^\perp (= I - P_k)$, respectively. Here, U_k represents the principal text subspace, so P_k and $P_k^\perp$ decompose an activation into the text-subspace component and its residual component.

Cross-modal alignment improvement. Motivated by the former analysis, DCAT first introduces a donor-guided cross-modal alignment term on text inputs:

$$\mathcal{L}_{\text{cross}}(\tau) = \beta \cdot \|X_{\text{text}}(\tau - \tau_r)\|_F^2 + (1 - \beta) \cdot \|X_{\text{text}}(\tau - \tau_d)\|_F^2. \tag{6}$$

The recipient anchor preserves the recipient model's original behavior and prevents the update from overwriting modality-specific capabilities. The donor anchor is complementary: from the perspective of linear mode connectivity [15, 77], donor weights on text inputs define directions along which DCAT can improve modality-text alignment while remaining in a text-compatible region of weight space. Thus, the donor-side term does not merely pull the solution toward the donor model; it transfers donor-induced cross-modal alignment directions through text space, allowing the recipient model to benefit from the donor while suppressing unexpected drift of C_δ.

Activation-space target construction. Beyond donor-guided cross-modal alignment transfer, DCAT further improves the recipient model's self-alignment by constructing an activation-space target for the recipient modality. Since U_k defines the principal text subspace, we project the recipient representation as $X_{\text{mod}}(\theta_0 + \tau_r)U_k^T$. We compute its SVD as $A\ \text{diag}(\sigma_1, \ldots, \sigma_k)B^\top$ and define $\rho = \|X_{\text{mod}}(\theta_0 + \tau_r)\|_F/\sqrt{k}$, which preserves the total output energy across all k text-subspace directions. DCAT then reshapes the original spectrum into a more balanced projected target T_α:

$$T_\alpha = c_\alpha \cdot A\ \text{diag}\left(\sigma_j^\alpha \rho^{1-\alpha}\right)_{j=1}^k B^\top U_k, \qquad c_\alpha = \frac{\|X_{\text{mod}}(\theta_0 + \tau_r)\|_F}{\left(\sum_{j=1}^k \sigma_j^{2\alpha}\rho^{2(1-\alpha)}\right)^{1/2}}. \tag{7}$$

Note that T_α reflects favorable properties for increasing the lower bound in Theorem 1 such that:

$$T_\alpha P_k^\perp = 0, \qquad \|T_\alpha\|_F = \|X_{\text{mod}}(\theta_0 + \tau_r)\|_F. \tag{8}$$

Derivation for Eq. 8 is provided in Appendix Sec. B.2. The first property makes T_α contained in the principal text subspace, while the second prevents the target from changing the scale of the recipient activations. Thus, T_α improves self-alignment by pulling the recipient modality toward its own text subspace while preserving the recipient's activation energy, which prevents increasing text-subspace overlap by merely shrinking the modality activation, undesirable for the trace component in C_δ.

The role of α is to control the trade-off between preserving the recipient representation within the text subspace and increasing spectral diversity, while the target remains norm-preserving for all α. Specifically, keeping the original singular values $\{\sigma_j\}$ would preserve the recipient geometry but may leave the projected representation spectrally concentrated. On the other hand, replacing them with a fully isotropic spectrum would maximize diversity but may distort the recipient representation.

Unified self-alignment matching. Given the target T_α, DCAT improves self-alignment by matching the recipient-modality activations to this target:

$$\begin{aligned}\mathcal{L}_{\text{self}}(\tau) &= \underbrace{\|X_{\text{mod}}(\theta_0 + \tau)P_k^\perp\|_F^2}_{\textbf{Encouraging SEO}} + \underbrace{\|X_{\text{mod}}(\theta_0 + \tau)P_k - T_\alpha\|_F^2}_{\textbf{Encouraging SD}} \\ &= \|X_{\text{mod}}(\theta_0 + \tau) - T_\alpha\|_F^2,\end{aligned} \tag{9}$$

where the equality follows from the orthogonality of P_k (derived in Appendix Sec. B.3). Note that the first term reduces activation energy outside the principal text subspace, while the second shapes the projected representation within that subspace. Therefore, once the target T_α is constructed inside the text subspace with a balanced spectrum, $\mathcal{L}_{\text{self}}(\tau)$ simultaneously encourages SEO and SD.

Closed-form DCAT update. Combining the cross-modal alignment improvement term $\mathcal{L}_{\text{cross}}(\tau)$ and the self-alignment matching term $\mathcal{L}_{\text{self}}(\tau)$, the final objective is:

$$\min_\tau\ \mathcal{L}_{\text{self}}(\tau) + \eta \cdot \mathcal{L}_{\text{cross}}(\tau). \tag{10}$$

That is, $\mathcal{L}_{\text{cross}}(\tau)$ transfers donor-induced cross-modal alignment directions while preserving the recipient model's original text behavior, and $\mathcal{L}_{\text{self}}(\tau)$ further reshapes the recipient-modality activations toward the norm-preserving target T_α. The donor enters through the cross-modal alignment improvement term, which biases the closed-form solution toward donor-compatible text-aligned directions without overwriting the recipient modality representation.

Since Eq. 10 is quadratic in τ, DCAT admits a closed-form layer-wise solution without gradient-based optimization (see Appendix Sec. B.4). To avoid mismatch from applying all layer-wise solutions at once, since T_α is constructed using pre-update recipient activations, DCAT progressively merges each layer's solution and recomputes calibration activations for the next layer using the partially updated model. This reduces accumulated activation drift while preserving closed-form efficiency.

Scheduling α. The degree of alignment can vary across layers, as well as across the linear projections within each layer. Accordingly, α should adaptively balance SEO and SD by adjusting the spectrum of the target T_α inside the principal text subspace. To this end, we apply a simple yet effective adaptive schedule for α as follows:

$$\alpha = \text{Sigmoid}\left(\frac{\log \text{SEO}_r}{\log \text{SD}_r + \epsilon}\right), \tag{11}$$

where $\epsilon > 0$ is for stability. This schedule makes the target more conservative when the recipient has weak SEO, and more isotropic when the projected representation is spectrally concentrated. Accordingly, larger α preserves more of the recipient's projected spectrum, while smaller α encourages SD. To support this formulation, we analyze the effective rank of textual and modality-specific token representations across models and modalities; details are provided in Appendix Sec. E.

5 Experiment

5.1 Experimental settings

Baselines. Although first observed in model merging, DCAT has a distinct goal: transferring stronger donor alignment to a recipient rather than equipping one MLLM with multiple tasks or modalities. Because both settings combine models at the weight level, we compare DCAT with representative merging baselines: Task Arithmetic [28] (TA), TIES-merging [82] (TIES), ISO-C [48] (ISO), Task Singular Vector [17] (TSV), WUDI [9], and OptMerge [75]. Experiments cover Audio+Vision and Video+Vision settings on two widely used backbones, Vicuna [10] and LLaMA 3.1 [19]. For Vicuna, we use the weights released by OptMerge. For LLaMA 3.1, we use in-the-wild models that share the same architecture but are independently trained with different recipes, demonstrating DCAT's robustness and practical utility. Full baseline details are in the Appendix Sec. D.2.

Benchmarks. We extensively compare DCAT's alignment transfer performance across audio and video benchmarks. For audio, we evaluate on MMAU [69] (1K samples), AIR-Bench Foundation [84] (25K samples), MuCho-music [74] (1.2K samples), Clotho-AQA [40] (1.4K samples), and VocalSound [18] (3.6K samples). For video, we evaluate on Video-MME [16] (2.7K samples), TempCompass [44] (5.5K samples), MVBench [37] (4K samples), and MLVU [91] (2.2K samples). Subcategory-level results are provided in the Appendix Sec. F.2.

Implementation details. DCAT progressively merges models layer by layer from the first to the last layer. At each layer, our closed-form update is applied to all linear projections, aligning internal representations across computational modules. For calibration, we use AVQA [83], which supports both audio and video: audio tracks calibrate audio MLLMs, and video tracks calibrate video MLLMs, using 256 samples each. Remaining hyperparameters, including the text subspace dimension k, cross-modal alignment weight η, and anchor parameter β, are detailed in Appendix C.

5.2 Quantitative comparison

To evaluate how effectively DCAT boosts modality performance, we compare it against various model merging methods. The Audio+Vision results are shown in the top of Tab. 2, and the Video+Vision results in the bottom. For the LLaMA-based model, the default merging coefficient $\lambda = 1.0$ used by most methods consistently degrades performance; for a fair comparison, we therefore additionally report each method's performance at $\lambda = 0.2$, which yields its best results. On audio, DCAT improves over the baseline by 7.62% on comprehensive benchmarks such as MMAU, and by up to 100.92% on MuCho-Music. On video, DCAT delivers gains of 16.37% on Video-MME and 23.83% on MVBench. Together with our analysis, these results demonstrate DCAT's effectiveness.

5.3 Does DCAT increase SEO and SD? Yes.

We theoretically showed that increasing SEO and SD can raise a lower bound on the MI, and on this basis we proposed DCAT. To verify that DCAT indeed increases SEO and SD in practice, we measured both quantities at every layer of each Vicuna-based model under the Audio+Vision and Video+Vision settings. In addition, to check whether the first term of C_δ in Theorem 1, $\frac{k}{2}\log(\text{tr}(\Sigma_y/k))$, is preserved, we measured the token norm ratio $\|Y_{\text{mod}}^{\text{aft.}}\|_F / \|Y_{\text{mod}}^{\text{bef.}}\|_F$ between before and after applying

Table 2: **Quantitative comparison of DCAT.** We evaluate DCAT against various model merging methods across benchmarks, using a vision model as donor and an audio/video model as recipient. **(Top)** In Audio+Vision, DCAT transfers vision-model alignment to the audio model, achieving gains up to **100.92%** (MuCho-Music) over the original. **(Bottom)** Video+Vision shows a similar trend, with improvements up to **53.88%** (MLVU). (*Since our LLaMA 3.1-based audio model does not tune the language backbone, several merging methods reduce to TA, yielding identical outputs.)

			MMAU				AIR-Bench				MuCho	Clotho	Vocal	Average
			Music	Sound	Speech	Overall	Music	Sound	Speech	Overall				
Vicuna-based		Original	54.79	60.66	31.23	48.90	38.44	46.76	44.62	43.17	26.96	61.23	48.98	45.85
		TA [28]	55.69	64.56	39.34	53.20	45.64	57.48	47.03	49.09	50.04	81.00	44.81	55.63
		TIES [82]	54.49	62.46	36.04	51.00	44.86	55.92	47.51	48.68	38.42	74.97	50.93	52.80
		ISO [48]	58.68	66.67	34.83	53.40	45.82	60.77	46.72	49.79	52.82	78.57	44.31	55.78
		TSV [17]	55.99	66.67	35.44	52.70	47.03	58.51	48.69	50.51	50.80	81.21	47.54	56.55
		WUDI [9]	55.99	65.17	39.94	53.70	45.84	58.16	46.85	49.23	49.45	80.72	46.98	56.02
		OptMerge [75]	59.58	64.26	35.44	53.10	47.02	58.24	48.83	50.50	50.21	80.24	49.07	56.62
		DCAT (Ours)	60.22	69.06	41.74	56.20	48.46	59.51	49.36	52.44	54.17	82.96	51.25	59.40
LLaMA 3.1-based		Original	52.40	57.66	56.46	55.50	33.53	36.35	53.00	42.86	52.57	55.48	31.02	47.49
	$\lambda = 1.0$	TA*	48.50	54.65	54.65	52.60	35.54	32.97	46.23	39.68	49.62	57.07	27.57	45.31
		TIES	52.10	56.16	59.76	56.00	37.43	36.11	49.39	42.43	54.93	57.56	30.49	48.28
		ISO	55.39	57.96	59.46	57.60	36.15	34.69	47.01	40.63	55.94	58.67	29.35	48.44
		TSV*	48.50	54.65	54.65	52.60	35.54	32.97	46.23	39.68	49.62	57.07	27.57	45.31
		WUDI	49.10	55.56	56.16	53.60	35.87	32.53	45.97	39.56	49.96	57.28	26.90	45.46
		OptMerge*	48.50	54.65	54.65	52.60	35.54	32.97	46.23	39.68	49.62	57.07	27.57	45.31
	$\lambda = 0.2$	TA*	55.99	58.56	57.06	57.20	35.73	35.57	51.75	42.81	54.51	57.56	31.22	48.66
		TIES	57.49	58.26	57.06	57.60	35.26	36.01	53.10	43.37	54.59	56.66	30.58	48.56
		ISO	52.99	58.56	56.76	56.10	34.55	36.43	53.67	43.50	53.24	57.49	30.19	48.10
		TSV*	55.99	58.56	57.06	57.20	35.73	35.57	51.75	42.81	54.51	57.56	31.22	48.66
		WUDI	54.49	57.06	55.56	55.70	35.49	35.45	52.14	42.88	52.74	57.49	29.91	47.74
		OptMerge*	55.99	58.56	57.06	57.20	35.73	35.57	51.75	42.81	54.51	57.56	31.22	48.66
		DCAT (Ours)	58.56	59.16	59.76	59.16	37.96	37.28	54.25	44.93	58.30	61.56	31.94	51.18

			Video-MME				TempCompass						MVBench	MLVU	Average
			Short	Medium	Long	Overall	Action	Attr.	Dir.	Order	Speed	Overall			
Vicuna-based		Original	39.89	34.22	30.89	35.00	78.48	54.00	48.53	47.56	45.85	55.46	38.48	32.59	40.38
		TA [28]	44.67	34.78	31.89	37.11	78.16	53.13	46.89	49.09	46.99	55.38	43.68	44.84	45.25
		TIES [82]	31.44	27.78	25.67	28.30	76.53	50.78	46.45	47.66	48.82	54.62	32.78	33.15	37.21
		ISO [48]	43.67	38.22	33.44	38.44	79.38	53.81	46.45	46.44	47.77	55.38	43.00	45.31	45.53
		TSV [17]	46.44	36.67	34.00	39.04	82.72	55.76	47.75	49.29	47.69	57.24	44.58	45.18	46.51
		WUDI [9]	44.56	34.22	30.89	36.56	76.94	53.71	46.11	48.68	46.72	54.93	43.30	44.38	44.79
		OptMerge [75]	44.33	37.00	33.56	38.30	80.68	55.57	47.40	47.26	48.38	56.47	44.83	43.29	45.72
		DCAT (Ours)	47.96	40.46	36.17	41.53	83.80	58.42	50.98	50.83	53.52	59.71	47.65	46.34	48.81
LLaMA 3.1-based		Original	47.33	35.44	31.78	38.19	79.87	44.53	45.33	42.28	40.44	51.28	41.00	31.79	40.57
	$\lambda = 1.0$	TA	48.22	36.33	32.11	38.89	77.83	49.51	43.17	45.63	46.29	53.11	43.88	45.22	45.28
		TIES	50.78	38.67	35.00	41.48	74.74	44.63	41.70	37.40	34.06	47.22	46.88	47.12	45.68
		ISO	50.67	38.67	33.33	40.89	44.01	32.03	26.56	19.41	20.17	28.85	46.03	45.28	40.26
		TSV	51.78	39.11	34.67	41.85	83.29	52.83	46.63	47.26	50.31	56.77	47.00	48.08	48.43
		WUDI	50.78	36.67	33.00	40.15	19.48	11.91	10.73	3.86	5.24	10.53	45.35	47.51	35.89
		OptMerge	52.44	39.89	36.22	42.85	5.79	4.00	4.33	1.93	1.57	3.59	47.23	49.24	35.73
	$\lambda = 0.2$	TA	52.56	38.11	38.11	42.93	63.49	38.09	36.33	35.87	39.04	43.15	43.18	40.94	42.55
		TIES	46.56	37.33	35.78	39.89	65.53	37.79	37.20	36.18	44.54	44.92	45.43	38.58	42.21
		ISO	45.67	36.44	36.44	39.52	83.21	52.34	47.23	46.85	52.58	57.19	44.80	44.48	46.50
		TSV	49.22	37.67	37.44	41.44	79.05	47.56	43.43	43.90	49.17	53.36	45.65	40.42	45.22
		WUDI	51.56	37.00	38.33	42.30	85.74	53.52	47.75	48.78	52.66	58.44	44.55	43.81	47.28
		OptMerge	47.78	37.78	36.22	40.59	83.94	51.86	46.71	47.05	53.10	57.30	46.48	41.83	46.55
		DCAT (Ours)	53.67	40.33	39.33	44.44	86.37	55.64	50.47	50.19	56.02	59.88	51.18	50.15	51.41

DCAT, while the second term $\det(\bar{\Sigma}_{z|x})$ is intractable to estimate directly. The results, presented in Fig. 4, show that DCAT yields a significant increase in both SEO and SD while keeping the norm ratio essentially unchanged, suggesting that DCAT successfully raises the lower bound on the MI.

5.4 Ablation study

To assess the contribution of each component of DCAT, we conduct an ablation study summarized in Tab. 3. Experiments use Vicuna-based models: the audio+vision setting employs the full MMAU benchmark, while the video+vision setting uses 900 randomly sampled examples from Video-MME. A2–A5 ablate the components of DCAT's objective and the adaptive α schedule and the results show that each component contributes meaningfully to performance and confirm both the validity of design choices. Further ablation study including hyperparameter sweep, calibration dataset size and projection dimension k is provided in Appendix Sec. F.8.

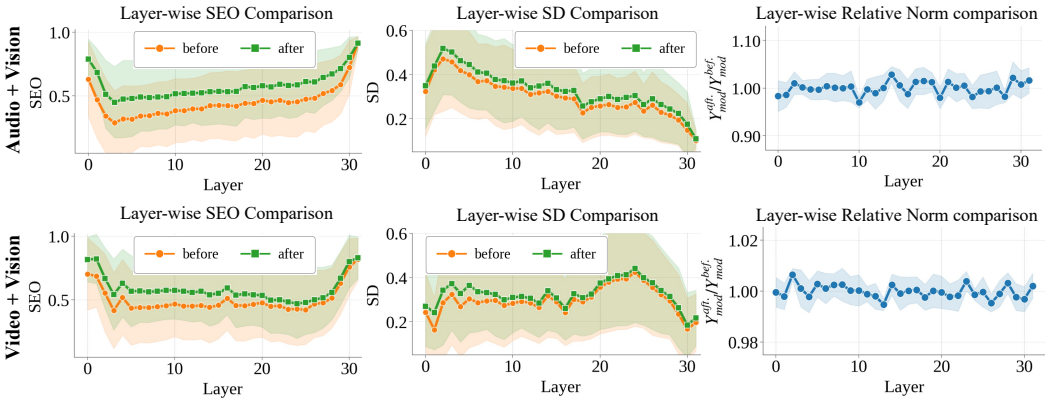

Figure 4: **SEO, SD and token norm ratio before and after applying DCAT.** We compare SEO, SD, and the norm ratio of modality tokens before and after applying DCAT. As depicted in first and second column, SEO and SD increase to a statistically significant degree ($p < 0.05$), while the relative norm remains close to 1.0, confirming that our DCAT formulation behaves as intended.

Table 3: **Ablation study on DCAT.** We ablate the individual components of DCAT. (A2–A5) isolate the contribution of each term in the DCAT objective by either dropping a term or fixing the α scheduler, while (A6–A9) examine the sensitivity of DCAT to its key hyperparameters.

ID	Study subject	MMAU				Video-MME			
		Music	Sound	Speech	Overall	Short	Medium	Long	Overall
A0	Original	54.79	60.66	31.23	48.90	38.33	33.67	30.67	34.22
A1	Base (DCAT)	60.22	69.06	41.74	56.20	46.33	39.67	34.33	40.11
A2	w/o SEO encouraging term in $\mathcal{L}_{\text{self}}(\tau)$	57.49	66.97	34.83	53.10	45.33	34.00	30.33	36.56
A3	w/o SD encouraging term in $\mathcal{L}_{\text{self}}(\tau)$	55.09	64.26	34.53	51.30	35.67	25.00	25.00	28.56
A4	w/o $\mathcal{L}_{\text{cross}}(\tau)$	25.45	27.63	24.62	25.90	3.33	4.67	2.67	3.56
A5	Fixed α scheduling	55.99	61.86	36.34	51.40	36.67	31.33	27.67	31.89

6 Discussion

As shown in Fig. 2, we consistently observe that vision models suffer a performance drop when merging is applied, and this trend holds across multiple models and architectures. Accordingly, most of our experiments employ a vision model as the donor and MLLMs of other modalities as recipients for alignment transfer. We offer the following conjecture as to why the donor has consistently been a vision model: because visual data is easier to obtain and far more abundant than audio or video data, existing vision MLLMs are likely to be more thoroughly aligned than their counterparts in other modalities. Under this hypothesis, DCAT is expected to leverage models trained on common modalities to enhance the performance of MLLMs built for relatively rare and data-scarce modalities.

7 Conclusion

In this paper, we report the phenomenon that inter-modality model merging can improve performance, and we interpret this effect from a Mutual Information (MI) perspective by deriving a lower bound on the MI and proposing a monotonic Alignment metric. An extensive empirical and statistical analysis demonstrates that this Alignment exhibits a strong correlation with downstream performance. Building on these theoretical and empirical findings, we propose Directional Cross-Modal Alignment Transfer (DCAT), which transfers alignment from a donor modality model to a recipient modality model in closed form. Across multiple audio and video MLLMs, we validate DCAT on a wide range of benchmarks, achieving average relative improvements of 29.55% on audio and 24.92% on video over the original models, thereby substantiating the soundness and the effectiveness of DCAT.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government(MSIT) [No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), No.RS-2025-02314125, Effective Human-Machine Teaming With Multimodal Hazy Oracle Models], the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-02263628), Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [26CR1100, Development of molecular endoscopy for real-time precision tumor analysis], and Korea Health Industry Development Institute(RS-2024-00439677). Also, the authors acknowledge financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.

References

[1] Antonio Almudévar, José Miguel Hernández-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. Aligning multimodal representations through an information bottleneck. *ICML*, 2025.

[2] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, volume 34, pages 7432–7439, 2020.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[9] Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. *ICML*, 2025.

[10] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[11] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[12] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

[13] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, pages 2924–2936, 2019.

[14] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[15] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *ICML*, 2020.

[16] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, pages 24108–24118, 2025.

[17] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *CVPR*, pages 18695–18705, 2025.

[18] Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155, 2022.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[20] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[22] Ming-Hao Hsu, Xueyao Zhang, Xiaohai Tian, Jun Zhang, and Zhizheng Wu. Anatomy of the modality gap: Dissecting the internal states of end-to-end speech llms. *arXiv preprint arXiv:2603.01502*, 2026.

[23] Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavllm: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572, 2024.

[24] Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*, 2025.

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019.

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[27] Dang Huu-Tien, Takeshi Suzuki, Le-Minh Nguyen, et al. Regmean++: Enhancing effectiveness and generalization of regression mean for model merging. *TMLR*, 2025.

[28] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *ICLR*, 2023.

[29] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *CVPR*, pages 25004–25014, 2025.

[30] Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *ICLR*, 2023.

[31] J Kreer. A question of terminology. *IRE Transactions on Information Theory*, 3(3):208–208, 1957.

[32] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *EMNLP*, pages 785–794, 2017.

[33] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models, March 2024. Bo Li, Peiyuan Zhang, Kaichen Zhang, and Fanyi Pu contributed equally.

[34] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *TMLR*, 2025.

[35] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[37] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024.

[38] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023.

[39] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pages 5971–5984, 2024.

[40] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE, 2022.

[41] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.

[43] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, pages 216–233. Springer, 2024.

[44] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *ACL*, pages 8731–8772, 2024.

[45] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.

[46] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, pages 12585–12602, 2024.

[47] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 36:46212–46244, 2023.

[48] Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost Van De Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. *ICML*, 2025.

[49] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, pages 2381–2391, 2018.

[50] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[51] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[53] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[55] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95, 2011.

[56] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[57] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[58] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *EMNLP-IJCNLP*, pages 4463–4473, 2019.

[59] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[60] Hoigi Seo, Dong Un Kang, Hyunjin Cho, Joohoon Lee, and Se Young Chun. On epistemic uncertainty of visual tokens for object hallucinations in large vision-language models. *NeurIPS*, 2025.

[61] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[62] Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

[63] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.

[64] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *ICML*, 2025.

[65] Weijie Su, Xizhou Zhu, Chenxin Tao, Lewei Lu, Bin Li, Gao Huang, Yu Qiao, Xiaogang Wang, Jie Zhou, and Jifeng Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. In *CVPR*, pages 15888–15899, 2023.

[66] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL*, pages 4149–4158, 2019.

[67] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[68] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[69] Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *ICLR*, 2025.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[71] Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. How visual representations map to language feature space in multimodal llms. *arXiv preprint arXiv:2506.11976*, 2025.

[72] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[73] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[74] Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, György Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models. In *ISMIR*, 2024.

[75] Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Optmerge: Unifying multimodal llm capabilities and modalities via model merging. *arXiv preprint arXiv:2505.19892*, 2025.

[76] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.

[77] Chengyue Wu, Teng Wang, Yixiao Ge, Zeyu Lu, Ruisong Zhou, Ying Shan, and Ping Luo. π-tuning: Transferring multimodal foundation models with optimal multi-task interpolation. In *ICML*, 2023.

[78] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 37:28828–28857, 2024.

[79] Hongxuan Wu, Yukun Zhang, and Xueqing Zhou. How vision becomes language: A layer-wise information-theoretic analysis of multimodal reasoning. *arXiv preprint arXiv:2602.15580*, 2026.

[80] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[81] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, pages 131–147. Springer, 2024.

[82] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *NeurIPS*, 36:7093–7115, 2023.

[83] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *ACMMM*, 2022.

[84] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024.

[85] Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *CVPR*, pages 24972–24982, 2025.

[86] Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.

[87] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *ACL*, pages 4791–4800, 2019.

[88] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023.

[89] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024.

[90] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[91] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[92] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

[93] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ICLR*, 2024.

A Code

We provide the code implementation of DCAT for reproducibility. Please refer to this link for environment setup instructions and execution guidelines.

B Proof

B.1 Proof of Theorem 1

We first derive two auxiliary lemmas that connect the definitions of SEO and SD to the covariance quantities used in the information-theoretic lower bound.

Lemma 1 (SEO bounded by projected covariance ratio)**.** *Let $Y_{\text{mod}} \in \mathbb{R}^{N_m \times d}$ be the modality-token activation matrix, and let $U_k \in \mathbb{R}^{k \times d}$ contain the top-k right singular vectors of the centered text-token activations. Define*

$$\mu_y := \frac{1}{N_m} Y_{\text{mod}}^\top \mathbf{1} \in \mathbb{R}^d, \qquad \widetilde{Y}_{\text{mod}} := Y_{\text{mod}} - \mathbf{1}\mu_y^\top. \tag{A1}$$

Assume that the mean activation is orthogonal to the text principal subspace[1]*:*

$$\mu_y U_k^\top = 0. \tag{A2}$$

Define

$$Z := Y_{\text{mod}} U_k^\top, \qquad \widetilde{Z} := \widetilde{Y}_{\text{mod}} U_k^\top, \tag{A3}$$

and define the unbiased covariance matrices

$$\Sigma_y := \frac{1}{N_m - 1} \widetilde{Y}_{\text{mod}}^\top \widetilde{Y}_{\text{mod}}, \qquad \Sigma_Z := \frac{1}{N_m - 1} \widetilde{Z}^\top \widetilde{Z}. \tag{A4}$$

The Subspace Energy Overlap is defined using the uncentered activations:

$$\text{SEO} = \frac{\|Y_{\text{mod}} U_k^\top\|_F^2}{\|Y_{\text{mod}}\|_F^2}. \tag{A5}$$

If $\operatorname{tr}(\Sigma_y) > 0$, then

$$\text{SEO} \le \frac{\operatorname{tr}(\Sigma_Z)}{\operatorname{tr}(\Sigma_y)}. \tag{A6}$$

Proof. We decompose the modality activations into their centered component and mean component:

$$Y_{\text{mod}} = \widetilde{Y}_{\text{mod}} + \mathbf{1}\mu_y^\top. \tag{A7}$$

Since $\mathbf{1}^\top \widetilde{Y}_{\text{mod}} = 0$, these two components are orthogonal in Frobenius inner product. Therefore,

$$\begin{aligned} \|Y_{\text{mod}}\|_F^2 &= \|\widetilde{Y}_{\text{mod}} + \mathbf{1}\mu_y^\top\|_F^2 \\ &= \|\widetilde{Y}_{\text{mod}}\|_F^2 + \|\mathbf{1}\mu_y^\top\|_F^2 \\ &= (N_m - 1)\operatorname{tr}(\Sigma_y) + N_m \|\mu_y\|_2^2. \end{aligned} \tag{A8}$$

Next, consider the projected activations. By the assumption $\mu_y U_k^\top = 0$, we have

$$\mathbf{1}\mu_y^\top U_k^\top = 0. \tag{A9}$$

Hence,

$$\begin{aligned} \|Y_{\text{mod}} U_k^\top\|_F^2 &= \|\widetilde{Y}_{\text{mod}} U_k^\top + \mathbf{1}\mu_y^\top U_k^\top\|_F^2 \\ &= \|\widetilde{Y}_{\text{mod}} U_k^\top\|_F^2 \\ &= \|\widetilde{Z}\|_F^2 \\ &= (N_m - 1)\operatorname{tr}(\Sigma_Z). \end{aligned} \tag{A10}$$

[1] Our empirical validation on LLaMA-3.1 indicates $\|\mu_y U_k^\top\|_2 / \|\mu_y\|_2$ is 0.09726 ± 0.02166 which is sufficiently small.

Substituting Eq. (A8) and Eq. (A10) into the definition of SEO gives

$$\begin{aligned} \mathrm{SEO} &= \frac{\|Y_{\mathrm{mod}} U_k^\top\|_F^2}{\|Y_{\mathrm{mod}}\|_F^2} \\ &= \frac{(N_m - 1)\,\mathrm{tr}(\Sigma_Z)}{(N_m - 1)\,\mathrm{tr}(\Sigma_y) + N_m \|\mu_y\|_2^2}. \end{aligned} \tag{A11}$$

Since $N_m \|\mu_y\|_2^2 \geq 0$, we have

$$(N_m - 1)\,\mathrm{tr}(\Sigma_y) + N_m \|\mu_y\|_2^2 \geq (N_m - 1)\,\mathrm{tr}(\Sigma_y). \tag{A12}$$

Therefore,

$$\begin{aligned} \mathrm{SEO} &= \frac{(N_m - 1)\,\mathrm{tr}(\Sigma_Z)}{(N_m - 1)\,\mathrm{tr}(\Sigma_y) + N_m \|\mu_y\|_2^2} \\ &\leq \frac{(N_m - 1)\,\mathrm{tr}(\Sigma_Z)}{(N_m - 1)\,\mathrm{tr}(\Sigma_y)} \\ &= \frac{\mathrm{tr}(\Sigma_Z)}{\mathrm{tr}(\Sigma_y)}. \end{aligned} \tag{A13}$$

□

Lemma 2 (SD in terms of determinant and trace of Σ_Z)**.** *Let $\Sigma_Z \succ 0$ be the empirical covariance matrix of the projected modality-token activations $Z = Y_{\mathrm{mod}} U_k^\top$, and let $\lambda_1, \ldots, \lambda_k > 0$ denote its eigenvalues. Define*

$$p_j = \frac{\lambda_j}{\sum_{\ell=1}^k \lambda_\ell}, \qquad j = 1, \ldots, k. \tag{A14}$$

The Spectral Diversity is

$$\mathrm{SD} = \frac{1}{k} \exp\left(-\sum_{j=1}^k p_j \log p_j \right). \tag{A15}$$

Define the determinant-normalized spectral volume as

$$\mathrm{SD}_{\mathrm{det}} := \frac{\det(\Sigma_Z)^{1/k}}{\mathrm{tr}(\Sigma_Z)/k}. \tag{A16}$$

Assume that the projected spectrum satisfies the condition

$$p_j \geq \frac{\delta}{k}, \qquad \forall j \in \{1, \ldots, k\}, \tag{A17}$$

for some sufficiently large $\delta \in (0, 1]$. Then

$$\mathrm{SD}_{\mathrm{det}} \geq \delta\, \mathrm{SD}. \tag{A18}$$

Equivalently,

$$\log \mathrm{SD}_{\mathrm{det}} \geq \log \mathrm{SD} + \log \delta. \tag{A19}$$

Proof. First, by the entropy upper bound of a discrete distribution over k outcomes,

$$-\sum_{j=1}^k p_j \log p_j \leq \log k. \tag{A20}$$

Therefore, from Eq. (A15),

$$0 < \mathrm{SD} \leq 1. \tag{A21}$$

Next, using $\sum_j \lambda_j = \mathrm{tr}(\Sigma_Z)$ and $p_j = \lambda_j / \sum_{\ell=1}^k \lambda_\ell$, we have

$$\lambda_j = p_j\, \mathrm{tr}(\Sigma_Z). \tag{A22}$$

Thus,

$$\begin{aligned}
\mathrm{SD}_{\det} &= \frac{\det(\Sigma_Z)^{1/k}}{\mathrm{tr}(\Sigma_Z)/k} \\
&= \frac{\left(\prod_{j=1}^{k} \lambda_j\right)^{1/k}}{\mathrm{tr}(\Sigma_Z)/k} \\
&= \frac{\left(\prod_{j=1}^{k} p_j \,\mathrm{tr}(\Sigma_Z)\right)^{1/k}}{\mathrm{tr}(\Sigma_Z)/k} \\
&= k \left(\prod_{j=1}^{k} p_j\right)^{1/k} \\
&= \left(\prod_{j=1}^{k} k p_j\right)^{1/k}.
\end{aligned} \tag{A23}$$

By the condition in Eq. (A17), we have $kp_j \geq \delta$ for all j. Hence,

$$\mathrm{SD}_{\det} = \left(\prod_{j=1}^{k} k p_j\right)^{1/k} \geq \delta. \tag{A24}$$

Since $\mathrm{SD} \leq 1$, it follows that

$$\mathrm{SD}_{\det} \geq \delta \geq \delta\,\mathrm{SD}. \tag{A25}$$

Taking logarithms gives

$$\log \mathrm{SD}_{\det} \geq \log \mathrm{SD} + \log \delta. \tag{A26}$$

□

We empirically validate the use of δ in Sec. B.5. Based on the results of Lemma 1 and Lemma 2, we prove the main theorem.

Theorem 1 (Weighted Alignment-MI Lower Bound)**.** *Let $z = yU_k^\top$, where $U_k \in \mathbb{R}^{k\times d}$ contains the top-k right singular vectors of the centered text activation matrix. Let $\Sigma_y = \mathrm{Cov}(y)$ and $\Sigma_Z = \mathrm{Cov}(z)$. For each text-token representation x, let*

$$\Sigma_{z|x} := \mathrm{Cov}(z|x), \qquad \bar{\Sigma}_{z|x} := \mathbb{E}_x\left[\Sigma_{z|x}\right]. \tag{A27}$$

With assumption in Lemma 2, we also supposed that $\Sigma_Z \succ 0$, $\bar{\Sigma}_{z|x} \succ 0$, and:

- *(A1) $z \sim \mathcal{N}(\mu_Z, \Sigma_Z)$;*
- *(A2) $z|x \sim \mathcal{N}(\mu_z(x), \Sigma_{z|x})$;*

Then, for any $\gamma \in (0,1)$,

$$I(y;x) \geq \kappa(\gamma) \log\left(\mathrm{SEO}^{\gamma}\mathrm{SD}^{1-\gamma}\right) + C_\delta, \qquad \kappa(\gamma) = \frac{k}{2\min(\gamma, 1-\gamma)}, \tag{A28}$$

where

$$C_\delta = \frac{k}{2}\log\left(\frac{\mathrm{tr}(\Sigma_y)}{k}\right) - \frac{1}{2}\log\det(\bar{\Sigma}_{z|x}) + \frac{k}{2}\log\delta. \tag{A29}$$

Proof. Since $z = yU_k^\top$ is a deterministic function of y, the Markov chain

$$x \longleftrightarrow y \longrightarrow z \tag{A30}$$

holds. By the data processing inequality,

$$I(y;x) \geq I(z;x). \tag{A31}$$

It therefore suffices to lower-bound $I(z;x)$.

By the definition of mutual information,

$$I(z;x) = h(z) - h(z|x). \tag{A32}$$

Since $z \sim \mathcal{N}(\mu_Z, \Sigma_Z)$ and $\Sigma_Z \succ 0$, its marginal entropy satisfies:

$$h(z) = \frac{1}{2}\log\det(2\pi e\,\Sigma_Z). \tag{A33}$$

Under the conditional Gaussian assumption $z|x \sim \mathcal{N}(\mu_z(x), \Sigma_{z|x})$, we have

$$h(z|x) = \mathbb{E}_x\left[\frac{1}{2}\log\det(2\pi e\,\Sigma_{z|x})\right]. \tag{A34}$$

Since $\log\det(\cdot)$ is concave on the positive definite cone, Jensen's inequality gives

$$\mathbb{E}_x\left[\log\det(\Sigma_{z|x})\right] \le \log\det\left(\mathbb{E}_x[\Sigma_{z|x}]\right) = \log\det(\bar{\Sigma}_{z|x}). \tag{A35}$$

Therefore,

$$h(z|x) \le \frac{1}{2}\log\det(2\pi e\,\bar{\Sigma}_{z|x}). \tag{A36}$$

Combining the preceding displays yields

$$\begin{aligned} I(z;x) &= h(z) - h(z|x) \\ &\ge \frac{1}{2}\log\det(2\pi e\,\Sigma_Z) - \frac{1}{2}\log\det(2\pi e\,\bar{\Sigma}_{z|x}) \\ &= \frac{1}{2}\log\det(\Sigma_Z) - \frac{1}{2}\log\det(\bar{\Sigma}_{z|x}). \end{aligned} \tag{A37}$$

We now express the determinant term using SEO and SD. By the definition of $\mathrm{SD}_{\det}$,

$$\det(\Sigma_Z)^{1/k} = \frac{\mathrm{tr}(\Sigma_Z)}{k}\mathrm{SD}_{\det}. \tag{A38}$$

By Lemma 1,

$$\frac{\mathrm{tr}(\Sigma_Z)}{k} \ge \frac{\mathrm{tr}(\Sigma_y)}{k}\mathrm{SEO}. \tag{A39}$$

Hence,

$$\begin{aligned} \log\det(\Sigma_Z) &= k\log\left(\det(\Sigma_Z)^{1/k}\right) \\ &= k\log\left(\frac{\mathrm{tr}(\Sigma_Z)}{k}\mathrm{SD}_{\det}\right) \\ &\ge k\log\left(\frac{\mathrm{tr}(\Sigma_y)}{k}\mathrm{SEO}\cdot\mathrm{SD}_{\det}\right). \end{aligned} \tag{A40}$$

By Lemma 2, the condition in Eq. A17 implies

$$\log\mathrm{SD}_{\det} \ge \log\mathrm{SD} + \log\delta. \tag{A41}$$

Therefore,

$$\log\det(\Sigma_Z) \ge k\log\left(\frac{\mathrm{tr}(\Sigma_y)}{k}\right) + k\log\mathrm{SEO} + k\log\mathrm{SD} + k\log\delta. \tag{A42}$$

Substituting this into the mutual-information lower bound gives

$$\begin{aligned} I(z;x) &\ge \frac{k}{2}\log\mathrm{SEO} + \frac{k}{2}\log\mathrm{SD} + \frac{k}{2}\log\left(\frac{\mathrm{tr}(\Sigma_y)}{k}\right) - \frac{1}{2}\log\det(\bar{\Sigma}_{z|x}) + \frac{k}{2}\log\delta \\ &= \frac{k}{2}\left(\log\mathrm{SEO} + \log\mathrm{SD}\right) + C_\delta. \end{aligned} \tag{A43}$$

Finally, we convert the unweighted log-product into the weighted alignment score. Let

$$a := \log\mathrm{SEO}, \qquad b := \log\mathrm{SD}. \tag{A44}$$

Since $0 < \mathrm{SEO} \leq 1$ and $0 < \mathrm{SD} \leq 1$, we have $a \leq 0$ and $b \leq 0$. For any $\gamma \in (0, 1)$, let

$$m_\gamma := \min(\gamma, 1-\gamma). \tag{A45}$$

Since $\gamma \geq m_\gamma$ and $1-\gamma \geq m_\gamma$, and since $a, b \leq 0$, we have

$$\gamma a \leq m_\gamma a, \qquad (1-\gamma)b \leq m_\gamma b. \tag{A46}$$

Adding both inequalities gives

$$\gamma a + (1-\gamma)b \leq m_\gamma (a+b). \tag{A47}$$

Equivalently,

$$a + b \geq \frac{1}{m_\gamma} \left(\gamma a + (1-\gamma)b\right). \tag{A48}$$

Since

$$\gamma a + (1-\gamma)b = \log\left(\mathrm{SEO}^\gamma \mathrm{SD}^{1-\gamma}\right), \tag{A49}$$

we obtain

$$\log \mathrm{SEO} + \log \mathrm{SD} \geq \frac{1}{\min(\gamma, 1-\gamma)} \log\left(\mathrm{SEO}^\gamma \mathrm{SD}^{1-\gamma}\right). \tag{A50}$$

Consequently,

$$I(z; x) \geq \frac{k}{2\min(\gamma, 1-\gamma)} \log\left(\mathrm{SEO}^\gamma \mathrm{SD}^{1-\gamma}\right) + C_\delta. \tag{A51}$$

Since $I(y; x) \geq I(z; x)$, we conclude that

$$I(y; x) \geq \frac{k}{2\min(\gamma, 1-\gamma)} \log\left(\mathrm{SEO}^\gamma \mathrm{SD}^{1-\gamma}\right) + C_\delta. \tag{A52}$$

This completes the proof. □

B.2 Derivation of Eq. 8

In the main text, for the top-k right singular vectors U_k of the text activations, we defined the corresponding orthogonal projector and its orthogonal complement as $P_k\,(= U_k^\top U_k)$ and $P_k^\perp\,(= I - P_k)$, respectively. We further defined the uniform singular value that preserves the total output energy:

$$\rho \;=\; \frac{\|X_{\mathrm{mod}}(\theta_0 + \tau_r)\|_F}{\sqrt{k}}.$$

In addition, we used the following target for the modality-token activations after the weight update:

$$T_\alpha = c_\alpha \cdot A\,\mathrm{diag}\left(\sigma_j^\alpha \rho^{1-\alpha}\right)_{j=1}^k B^\top U_k, \qquad c_\alpha = \frac{\|X_{\mathrm{mod}}(\theta_0 + \tau_r)\|_F}{\left(\sum_{j=1}^k \sigma_j^{2\alpha} \rho^{2(1-\alpha)}\right)^{1/2}}$$

We now derive the following two properties of T_α:

$$T_\alpha P_k^\perp = 0, \qquad \|T_\alpha\|_F = \|X_{\mathrm{mod}}(\theta_0 + \tau_r)\|_F.$$

For the first one, T_α can be written in the form $T_\alpha = QU_k$ for some matrix Q. Hence, we obtain:

$$\begin{aligned} T_\alpha P_k^\perp &= QU_k(I - P_k) \\ &= QU_k(I - U_k^\top U_k) \\ &= QU_k - QU_k U_k^\top U_k \\ &= QU_k - QU_k \\ &= 0, \end{aligned}$$

since $U_k U_k^\top = I_k$. For the second one, we denote D_α as:

$$D_\alpha \;:=\; \mathrm{diag}\big(\sigma_j^\alpha\, \rho^{1-\alpha}\big)_{j=1}^k.$$

Then, starting from the definition of T_α:

$$\|T_\alpha\|_F^2 \;=\; c_\alpha^2 \;\cdot\; \big\|A\,D_\alpha\,B^\top U_k\big\|_F^2.$$

Expanding the Frobenius norm via $\|M\|_F^2 = \mathrm{tr}(M^\top M)$:

$$\begin{aligned}\big\|A\,D_\alpha\,B^\top U_k\big\|_F^2 &= \mathrm{tr}\Big(U_k^\top\,B\,D_\alpha\,\underbrace{A^\top A}_{=I_k}\,D_\alpha\,B^\top U_k\Big)\\ &= \mathrm{tr}\Big(U_k^\top\,B\,D_\alpha^2\,B^\top U_k\Big).\end{aligned} \tag{A53}$$

Applying the cyclic property of the trace, $\mathrm{tr}(PQRS) = \mathrm{tr}(SPQR)$:

$$\begin{aligned}\mathrm{tr}\Big(U_k^\top\,B\,D_\alpha^2\,B^\top U_k\Big) &= \mathrm{tr}\Big(\underbrace{U_k\,U_k^\top}_{=I_k}\,B\,D_\alpha^2\,B^\top\Big)\\ &= \mathrm{tr}\Big(B\,D_\alpha^2\,B^\top\Big).\end{aligned} \tag{A54}$$

Since B is orthogonal,

$$\mathrm{tr}\Big(B\,D_\alpha^2\,B^\top\Big) = \mathrm{tr}(D_\alpha^2) \;=\; \sum_{j=1}^{k}\sigma_j^{2\alpha}\,\rho^{2(1-\alpha)}. \tag{A55}$$

Substituting (A53)–(A55) back and inserting the definition of c_α:

$$\begin{aligned}\|T_\alpha\|_F^2 &= c_\alpha^2\cdot\sum_{j=1}^{k}\sigma_j^{2\alpha}\,\rho^{2(1-\alpha)}\\ &= \frac{\|X_{\mathrm{mod}}(\theta_0+\tau_r)\|_F^2}{\displaystyle\sum_{j=1}^{k}\sigma_j^{2\alpha}\,\rho^{2(1-\alpha)}}\;\cdot\;\sum_{j=1}^{k}\sigma_j^{2\alpha}\,\rho^{2(1-\alpha)}\\ &= \|X_{\mathrm{mod}}(\theta_0+\tau_r)\|_F^2.\end{aligned}$$

B.3 Derivation of Eq. 9

In Eq. 9 of the main text, we adopted the following objective to enhance alignment:

$$\begin{aligned}\mathcal{L}_{\mathrm{self}}(\tau) &= \underbrace{\|X_{\mathrm{mod}}(\theta_0+\tau)P_k^\perp\|_F^2}_{\textbf{Encouraging SEO}} + \underbrace{\|X_{\mathrm{mod}}(\theta_0+\tau)P_k - T_\alpha\|_F^2}_{\textbf{Encouraging SD}}\\ &= \|X_{\mathrm{mod}}(\theta_0+\tau) - T_\alpha\|_F^2.\end{aligned}$$

Then, the second equality can be derived as follows. From the previous derivation, T_α satisfies $T_\alpha P_k^\perp = 0$, which is equivalent to:

$$T_\alpha P_k = T_\alpha. \tag{A56}$$

For notational convenience, let us write $Y := X_{\mathrm{mod}}(\theta_0+\tau)$. Since $P_k + P_k^\perp = I$, we can decompose Y and T_α as:

$$Y = YP_k + YP_k^\perp, \qquad T_\alpha = T_\alpha P_k + T_\alpha P_k^\perp = T_\alpha + 0 = T_\alpha,$$

where the second equality uses (A56). Therefore:

$$Y - T_\alpha = (YP_k - T_\alpha) + YP_k^\perp. \tag{A57}$$

Expanding the squared Frobenius norm of (A57):

$$\|Y - T_\alpha\|_F^2 = \|YP_k - T_\alpha\|_F^2 + \|YP_k^\perp\|_F^2 + 2\,\mathrm{tr}\Big((YP_k - T_\alpha)^\top YP_k^\perp\Big).$$

It remains to show that the cross term vanishes. We split it into two parts:

$$\operatorname{tr}\Big((Y P_k - T_\alpha)^\top Y P_k^\perp\Big) = \underbrace{\operatorname{tr}\Big(P_k\, Y^\top Y\, P_k^\perp\Big)}_{\text{(I)}} \;-\; \underbrace{\operatorname{tr}\Big(T_\alpha^\top\, Y\, P_k^\perp\Big)}_{\text{(II)}}. \tag{A58}$$

For the term (I), the cyclic property of the trace gives us:

$$\operatorname{tr}\Big(P_k\, Y^\top Y\, P_k^\perp\Big) = \operatorname{tr}\Big(\underbrace{P_k^\perp\, P_k}_{=0}\, Y^\top Y\Big) = 0, \tag{A59}$$

where we used $P_k^\perp P_k = 0$. For the term (II), since $T_\alpha = Q\, U_k$ by its definition, we have $T_\alpha^\top = U_k^\top Q^\top$. By the cyclic property of the trace:

$$\operatorname{tr}\Big(T_\alpha^\top\, Y\, P_k^\perp\Big) = \operatorname{tr}\Big(U_k^\top\, Q^\top\, Y\, P_k^\perp\Big) = \operatorname{tr}\Big(\underbrace{P_k^\perp\, U_k^\top}_{=0}\, Q^\top\, Y\Big) = 0, \tag{A60}$$

where we used

$$P_k^\perp\, U_k^\top = (I - U_k^\top U_k)\, U_k^\top = U_k^\top - U_k^\top \underbrace{U_k\, U_k^\top}_{=I_k} = U_k^\top - U_k^\top = 0.$$

B.4 Derivation of closed form for DCAT

Proposition 1. *Let $X_m := X_{\text{mod}}$, $X_t := X_{\text{text}}$ for brevity. Assume that*

$$X_m^\top X_m + \eta X_t^\top X_t \succ 0. \tag{A61}$$

Then, for a given layer with pretrained weight $\theta_0 \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, the DCAT objective is defined as:

$$\min_{\tau}\; \mathcal{L}_{\text{self}}(\tau) \;+\; \eta \cdot \mathcal{L}_{\text{cross}}(\tau), \tag{A62}$$

with

$$\mathcal{L}_{\text{self}}(\tau) = \|X_{\text{m}}(\theta_0 + \tau) - T_\alpha\|_F^2,$$
$$\mathcal{L}_{\text{cross}}(\tau) = \beta\, \|X_{\text{t}}(\tau - \tau_r)\|_F^2 + (1-\beta)\, \|X_{\text{t}}(\tau - \tau_d)\|_F^2.$$

Then, the objective (A62) *has a unique minimizer as:*

$$\tau^\star = \big(X_m^\top X_m + \eta\, X_t^\top X_t\big)^{-1}\Big[X_m^\top\big(T_\alpha - X_m\theta_0\big) \;+\; \eta\, X_t^\top X_t\big(\beta\tau_r + (1-\beta)\tau_d\big)\Big]. \tag{A63}$$

Proof. The total objective $f(\tau) := \mathcal{L}_{\text{self}}(\tau) + \eta\, \mathcal{L}_{\text{cross}}(\tau)$ is a quadratic function of τ, so we proceed by computing its gradient and setting it to zero. First, $\mathcal{L}_{\text{self}}(\tau)$ can be represented as:

$$\begin{aligned}\mathcal{L}_{\text{self}}(\tau) &= \|X_{\text{m}}(\theta_0 + \tau) - T_\alpha\|_F^2 \\ &= \operatorname{tr}\big(\tau^\top X_m^\top X_m\, \tau\big) \;+\; 2\operatorname{tr}\big(\tau^\top X_m^\top (X_m\theta_0 - T_\alpha)\big) \;+\; \text{const}.\end{aligned}$$

Using the identities $\nabla_\tau \operatorname{tr}(\tau^\top M \tau) = (M + M^\top)\tau = 2M\tau$ for symmetric M, and $\nabla_\tau \operatorname{tr}(\tau^\top C) = C$, we obtain

$$\nabla_\tau\, \mathcal{L}_{\text{self}}(\tau) = 2\, X_m^\top X_m\, \tau \;+\; 2\, X_m^\top (X_m\theta_0 - T_\alpha) = 2\, X_m^\top\big(X_m(\theta_0 + \tau) - T_\alpha\big). \tag{A64}$$

Similarly, the gradient of $\mathcal{L}_{\text{cross}}$ can be represented as:

$$\begin{aligned}\nabla_\tau\, \mathcal{L}_{\text{cross}}(\tau) &= 2\beta\, X_t^\top X_t(\tau - \tau_r) \;+\; 2(1-\beta)\, X_t^\top X_t(\tau - \tau_d) \\ &= 2\, X_t^\top X_t\, \tau \;-\; 2\, X_t^\top X_t\big(\beta\tau_r + (1-\beta)\tau_d\big).\end{aligned} \tag{A65}$$

Then, $\nabla_\tau\, \mathcal{L}_{\text{self}} + \eta\, \nabla_\tau\, \mathcal{L}_{\text{cross}}$ from (A64) and (A65) is represented as:

$$X_m^\top X_m\, \tau \;+\; X_m^\top (X_m\theta_0 - T_\alpha) \;+\; \eta\, X_t^\top X_t\, \tau \;-\; \eta\, X_t^\top X_t\big(\beta\tau_r + (1-\beta)\tau_d\big) \;=\; 0. \tag{A66}$$

Rearranging Eq. A66 gives the linear system:

$$\underbrace{\big(X_m^\top X_m + \eta\, X_t^\top X_t\big)}_{=:\,A}\, \tau \;=\; \underbrace{X_m^\top(T_\alpha - X_m\theta_0) \;+\; \eta\, X_t^\top X_t\big(\beta\tau_r + (1-\beta)\tau_d\big)}_{=:\,B}.$$

Since we stack sufficiently large tokens to construct X_m and X_t, we assume $X_m^\top X_m + \eta\, X_t^\top X_t \succ 0$, the closed-form solution for τ is:

$$\tau^\star = \big(X_m^\top X_m + \eta\, X_t^\top X_t\big)^{-1}\Big[X_m^\top\big(T_\alpha - X_m\theta_0\big) \;+\; \eta\, X_t^\top X_t\big(\beta\tau_r + (1-\beta)\tau_d\big)\Big].$$

□

Table A1: **Empirical validation on SD, $\mathrm{SD}_{\mathrm{det}}$, and δ_{eff}.** To assess the tightness of the bound, we compute the entropy-based spectral diversity SD, the proof-side spectral diversity $\mathrm{SD}_{\mathrm{det}}$, and spectral conversion factor δ_{eff} for Vicuna-based models in Fig. 2. The results show that δ_{eff} remains close to one, indicating that the entropy-based effective-rank SD introduces only a mild constant-level relaxation in the MI lower bound.

Audio+Vision			Video+Vision		
SD	$\mathrm{SD}_{\mathrm{det}}$	$\delta_{\mathrm{eff}} \leq$	SD	$\mathrm{SD}_{\mathrm{det}}$	$\delta_{\mathrm{eff}} \leq$
$0.2839_{\pm 0.0273}$	$0.2752_{\pm 0.0228}$	$0.9704_{\pm 0.0224}$	$0.2015_{\pm 0.0178}$	$0.2226_{\pm 0.0197}$	$1.1046_{\pm 0.0140}$

B.5 Empirical validation of δ

The proof of Theorem 1 involves a spectral non-degeneracy factor that connects the proof-side spectral term $\mathrm{SD}_{\mathrm{det}}$ with the entropy-based effective-rank measure SD used in the main text. In particular, the relevant condition can be expressed as the existence of a constant $\delta > 0$ such that

$$\mathrm{SD}_{\mathrm{det}} \geq \delta\, \mathrm{SD}. \tag{A67}$$

Equivalently, the largest empirical value supported by a given spectrum is

$$\delta_{\mathrm{eff}} := \frac{\mathrm{SD}_{\mathrm{det}}}{\mathrm{SD}}. \tag{A68}$$

Importantly, δ_{eff} is a spectral conversion factor between two spectral-spread measures, not a per-eigenvalue probability lower bound of the form $p_j \geq \delta/k$. Therefore, δ_{eff} is not constrained to be at most one, and values larger than one do not contradict the probability constraint $\sum_j p_j = 1$.

For empirical validation, we compute both SD and $\mathrm{SD}_{\mathrm{det}}$ using the features examined in our analysis, and report their ratio δ_{eff} in Table A1. As shown in the table, δ_{eff} remains close to one in both settings: 0.9704 for Audio+Vision and 1.1046 for Video+Vision. These results indicate that the conversion from $\mathrm{SD}_{\mathrm{det}}$ to the entropy-based SD incurs only a mild constant-level relaxation in the MI lower bound. Thus, although the main text adopts the more interpretable entropy-based effective-rank measure SD, the empirical gap between SD and $\mathrm{SD}_{\mathrm{det}}$ is small in our analyzed model families. This supports the use of SD as a practical and meaningful spectral diversity metric for the proposed Alignment score.

C Implementation Details

C.1 Models and computation

For the Vicuna, the setting allows a controlled comparison because the three MLLMs are trained from the same base under the same recipe except for the modality in OptMerge [75]. To complement this controlled setting, we additionally evaluate DCAT on independently trained models that nevertheless share the LLaMA-3.1-8B-Instruct backbones for audio, video, and vision modalities. For the two backbones, the vision MLLM is used as the donor in every experiment, since vision MLLMs are the modality on which standard merging tends to leave performance unchanged or slightly degraded while consistently lifting the other modalities. All experiments run on a single NVIDIA L40S 48 GB GPU and the DCAT solver over all projections completes in about 25-30 minutes per setting.

C.2 Calibration and alignment-target data

For all four settings, calibration data is drawn from the publicly available AVQA dataset [83] since the evaluation benchmarks (MMAU, Video-MME, etc.) are entirely disjoint from AVQA so that calibration and evaluation never share data. AVQA provides paired audio and video streams aligned to the same set of question/answer prompts, which lets us instantiate DCAT in a uniform way across audio and video recipients without changing the underlying source of X_{text}. For audio recipients, we feed the *audio track* in AVQA to the recipient MLLM as the modality input; for video recipients, we instead feed the *video clip* of the same sample with the same prompt used for the audio track. To solve the objective in Eq. 10, we draw small set of AVQA dataset (256 samples for each modality).

Table A2: **Full implementation configuration of DCAT for the four headline settings.** A blank entry under *Evaluation seed* indicates that the benchmark is small enough to be evaluated in full and therefore does not require subsampling.

Component	Vicuna+Audio	Vicuna+Video	LLaMA+Audio	LLaMA+Video
Backbone	Vicuna-7B-v1.5	Vicuna-7B-v1.5	LLaMA-3.1-8B-Instruct	LLaMA-3.1-8B-Instruct
Calibration source	AVQA (audio track)	AVQA (video track)	AVQA (audio track)	AVQA (video track)
# calibration samples	256	256	256	256
k (text subspace dim)	128	128	256	256
β	0.30	0.30	0.95	0.95
η	1×10^5	1×10^5	1×10^6	1×10^6

C.3 Progressive, dependency-aware merging

Although the closed-form update in Eq. 10 could in principle be solved for all projections in parallel, we found it preferable to apply the update *progressively* along two axes—across layers and across projection groups within each layer. Concretely, we process layers sequentially, and within each layer we partition the seven projections into four dependency-aware groups that mirror the activation-flow structure of a transformer block: `{query,key,value}` and `{out}` projections in the attention layers, `{up,gate}` and `{down}` projections in the attention projections and the feed-forward networks. The four groups are merged in the listed order, and after every group merge we re-collect X_{mod} and X_{text} at the relevant point so that the next group is solved against activations that already reflect the previous update. In this way, errors that would arise from treating projections as independent are absorbed iteratively, and the alignment accumulated across layers is corrected as it propagates.

C.4 Hyperparameter values

In DCAT we should control three hyperparameters: the text-subspace dimension k, the recipient/donor balance β in the cross-modal alignment term, and the cross-modal alignment weight η. For example, we use a shared text-subspace dimension within each backbone family: $k = 128$ for Vicuna and $k = 256$ for LLaMA 3.1. The complete configuration is reported in Table A2.

C.5 Closed-form solve and engineering notes

The DCAT objective in Eq. 10 is quadratic in τ and admits a layer-wise SPD system $A\tau = B$ with

$$A \;=\; X_{\text{mod}}^{\top} X_{\text{mod}} \;+\; \eta\, X_{\text{text}}^{\top} X_{\text{text}}.$$

We solve this system in float64 using a Cholesky-based linear solver, with a small ridge $\epsilon_{\text{solve}} = 10^{-6}\,\text{tr}(A)/d$ added on the diagonal for numerical stability. Activation matrices used by the solver are collected in float32, and the resulting weight delta is applied to the recipient checkpoint in its native bfloat16 precision. No gradient-based optimization is involved at any point.

D Experimental Details

D.1 Analysis details

Alignment-accuracy Correlation Analysis. To examine whether the proposed alignment metric is correlated with model performance, we conduct a correlation analysis between alignment and benchmark accuracy using both Pearson and Spearman correlation coefficients. The analysis is performed on Audio+Vision and Video+Vision settings, under Vicuna [10] and LLaMA 3.1 [19] backbones. For model merging, we include Task Arithmetic [28] (TA), TIES-merging [82] (TIES), ISO-C [48] (ISO), Task Singular Vector [17] (TSV), WUDI [9], and OptMerge [75].

For each merging method, we first construct a fully merged model by merging the Vision model with either the Audio or Video model. Because correlation analysis requires multiple evaluation points rather than a single merged checkpoint, we further generate additional model variants by applying SVD-based low-rank decomposition to the fully merged network parameters. Specifically, starting from each full-merge checkpoint, we decompose the merged weights and reconstruct a series

of low-rank approximations with ranks 1, 2, 4, 8, 16, 32, 64, and 128. We then evaluate both the full-merge model and these SVD-derived variants, and compute Pearson and Spearman correlations between their Alignment scores and benchmark accuracies. This allows us to analyze how changes in the merged parameter space are reflected in both alignment and task performance.

The evaluation is conducted separately on the benchmark corresponding to each constituent modality. For example, in the Audio+Vision setting, we measure performance on both the audio benchmark and the vision benchmark. We use TextVQA [63] for vision, MMAU [69] for audio, and Video-MME [16] for video. In addition, we examine a broader set of benchmarks, whose results are reported in Sec. E.1 For each benchmark, we evaluate 300 samples that are equally sampled across categories to ensure balanced coverage. For video evaluation, we use tokens from only the first, middle, and last frames of each sample for computational efficiency, instead of processing all frames.

To provide a reference for the effect of merging itself, we additionally evaluate the original modality-specific models, i.e., the models trained on a single modality without merging, on the corresponding benchmarks. These reference results allow us to determine whether merging improves or degrades performance relative to the original specialized models.

For experiments, the proposed alignment metric is computed between modality tokens and text tokens at each backbone layer with $k = 32$ and $\gamma = 0.8$, and the alignment values are averaged over all samples in the benchmark dataset to obtain a layer-wise alignment profile. Since alignment varies substantially across layers, we analyze the results in a layer-wise manner, as shown in Fig. 3. Consistent with prior observations that cross-modal alignment tends to be strongest in middle layers, we select the 21st layer as a representative layer and report the linear regression plots and correlation values based on this layer in Fig. 2.

D.2 Comparison details

D.2.1 Baselines

Task Arithmetic. Task Arithmetic represents the effect of task-specific fine-tuning as a *task vector* in weight space, defined by the difference between the fine-tuned model and the pretrained initialization [28]. Given a pretrained model with parameters θ_0 and modality- or task-specific models $\{\theta_i\}_{i=1}^N$ fine-tuned from the same initialization, each task vector is computed as $\tau_i = \theta_i - \theta_0$. Model merging is then performed by adding a scaled sum of task vectors back to the pretrained weights:

$$\theta_{\mathrm{TA}} = \theta_0 + \lambda \sum_{i=1}^{N} (\theta_i - \theta_0)\,, \tag{A69}$$

where λ controls the strength of the merged update. The original work shows that such vectors can be added or negated to compose or suppress task behaviors, enabling training-free model editing and multi-task merging at the weight level.

TIES-Merging. TIES-Merging addresses parameter interference that arises when merging multiple task-specific models fine-tuned from the same pretrained initialization [82]. Given task vectors $\tau_i = \theta_i - \theta_0$, TIES first trims small-magnitude updates by resetting them to zero, thereby removing redundant parameter changes. It then elects a shared sign for each parameter dimension and merges only the task-vector entries whose signs agree with the elected sign. A simplified form of the resulting merge can be written as

$$\theta_{\mathrm{TIES}} = \theta_0 + \lambda \mathbf{m}, \qquad m_j = \frac{1}{|\mathcal{A}_j|} \sum_{i \in \mathcal{A}_j} \tilde{\tau}_{i,j}, \quad \mathcal{A}_j = \left\{ i : \mathrm{sign}(\tilde{\tau}_{i,j}) = \mathrm{sign}\left(\sum_{k=1}^{N} \tilde{\tau}_{k,j} \right) \right\}, \tag{A70}$$

where $\tilde{\tau}_i$ denotes the trimmed task vector and λ controls the merge strength. By resolving sign disagreement before averaging, TIES-Merging reduces destructive interference and improves training-free multi-task model merging compared to direct weight averaging or task-vector addition.

Isotropic Merging (ISO-C). Isotropic Merging analyzes model merging from the perspective of the singular subspaces of task-specific weight updates [48]. Given layer-wise task matrices $\Delta_i = W_i - W_0$, it first constructs a merged task matrix $\Delta = \sum_{i=1}^{N} \Delta_i$ and decomposes it as

$\Delta = U\ \mathrm{diag}(\sigma) \cdot V^\top$. Instead of preserving the skewed singular value spectrum of the summed update, Iso-C replaces the singular values with a uniform scale, thereby balancing the contribution of different singular directions. Iso-CTS further augments this common subspace with task-specific residual directions, which helps retain underrepresented task information when merging many diverse models. The resulting merge can be written as

$$W_{\mathrm{Iso}} = W_0 + \lambda U_{\mathrm{CTS}}\,(\bar{\sigma} I)\,V_{\mathrm{CTS}}^\top, \qquad \bar{\sigma} = \frac{1}{r}\sum_{j=1}^{r}\sigma_j, \tag{A71}$$

where U_{CTS} and V_{CTS} denote the orthogonalized bases obtained from common and task-specific subspaces, and λ controls the merge strength.

Task Singular Vectors. Task Singular Vectors (TSV) analyzes task vectors at the layer level by preserving each task-specific update as a matrix rather than flattening the entire network into a single vector [17]. Given a pretrained layer W_0^ℓ and task-specific layers $\{W_i^\ell\}_{i=1}^N$, TSV first decomposes each layer task matrix $\Delta_i^\ell = W_i^\ell - W_0^\ell$ using SVD and retains only the top-r singular components. The method observes that overlap between singular vectors from different tasks induces task interference, and therefore TSV-Merge reduces such interference by decorrelating the concatenated left and right singular-vector bases through whitening or an equivalent orthogonal Procrustes step. A simplified layer-wise merge can be written as

$$W_{\mathrm{TSV}}^\ell = W_0^\ell + \lambda \hat{U}^\ell \bar{\Sigma}^\ell (\hat{V}^\ell)^\top. \tag{A72}$$

where $\hat{U}^\ell$ and $\hat{V}^\ell$ denote the decorrelated versions of $\bar{U}^\ell$ and $\bar{V}^\ell$, and λ controls the merge strength. By combining low-rank approximation with singular-vector decorrelation, TSV-Merge aims to reduce destructive interference among task-specific updates while keeping the merge training-free.

WUDI-Merging. WUDI-Merging is a data-free model merging method that reduces interference among task-specific experts by exploiting the observation that linear-layer task vectors approximately span the corresponding input subspace [9]. Given a pretrained model θ and task vectors $\tau_i = \theta_i - \theta$, WUDI focuses on the linear layers and directly optimizes the merged task vector $\tau_{m,l}$ to stay compatible with each task vector $\tau_{i,l}$. For each linear layer l, the optimization objective is written as

$$\mathcal{L}_l(\tau_{m,l}) = \sum_i \frac{1}{\|\tau_{i,l}\|_F^2} \left\|(\tau_{m,l} - \tau_{i,l})\tau_{i,l}^\top\right\|_F^2, \quad \theta_m = \theta + \tau_m. \tag{A73}$$

The merged task vectors are optimized independently for each linear layer and then assembled into the final merged model. By using task vectors themselves as guidance signals, WUDI-Merging mitigates destructive interference without requiring additional training data, test-time samples, or extra task-specific modules.

OptMerge. OptMerge studies model merging for MLLMs and proposes a data-free optimization method for combining task- or modality-specific expert models [75]. Given task vectors $\tau_i = \theta_i - \theta_0$, OptMerge observes that directly optimizing the merged task vector can be unstable due to redundant and noisy components in task-specific updates. For each linear layer l, it first applies a low-rank approximation to the centered task matrix, $\tau_{i,l} - \bar{\tau}_l \approx U_{i,k}\Sigma_{i,k}V_{i,k}^\top$, and uses the denoised update $\tilde{\tau}_{i,l} = U_{i,k}\Sigma_{i,k}V_{i,k}^\top + \bar{\tau}_l$ for optimization. The layer-wise merged vector is obtained by minimizing task-vector interactions:

$$\tau_{m,l}^* = \arg\min_{\tau_{m,l}} \sum_{i=1}^{N} \frac{\left\|(\tau_{m,l} - \tilde{\tau}_{i,l})S_{i,l}^\top\right\|_F^2}{\|\tau_{i,l}\|_F^2}, \quad S_{i,l} = \Sigma_{i,k}V_{i,k}^\top, \quad \theta_m = \theta_0 + \tau_m. \tag{A74}$$

For LoRA fine-tuned models, OptMerge further uses direct truncated SVD, mean initialization, and SGD-based optimization to avoid excessive growth of the merged vector norm.

D.2.2 Base models

Vicuna-based MLLMs. We use the Vicuna-based modality-specific MLLMs released by OptMerge [75]. All models share Vicuna-7B-v1.5 as the language backbone, while employing modality-specific encoders and connectors: the vision-language model uses CLIP-ViT-L-336px [52] with an

MLP projector, the audio-language model uses BEATs-Iter3+ [45] with a Q-Former [36] connector, and the video-language model uses LanguageBind [92] with an MLP projector. These models are trained on bi-modal modality-text data for vision, audio, and video, respectively, following a two-stage procedure consisting of connector alignment and subsequent fine-tuning. Following the original modality-merging setup, we preserve each modality-specific encoder and connector, and apply merging only to the shared language-model parameters. This setup allows us to evaluate whether cross-modality merging can transfer alignment across different modality-specific MLLMs while keeping the underlying language backbone fixed.

LLaMA 3.1-based MLLMs. We further evaluate DCAT on LLaMA 3.1-based MLLMs collected from publicly available in-the-wild checkpoints. Specifically, we use LLaVA-LLaMA3.1-8B[2] as the vision-language model, Ultravox-v0.4.1-LLaMA-3.1-8B[3] as the audio-language model, and LLaVA-Video-LLaMA-3.1-8B[4] as the video-language model. Unlike the Vicuna-based models, these checkpoints are independently developed and trained with different data mixtures, training recipes, and modality-specific components, while sharing the same LLaMA 3.1-8B language-model architecture. For each model, we preserve its original modality encoder and connector, and apply merging only to the language-model parameters that are architecturally compatible across models. This setting provides a more practical and less controlled testbed for evaluating whether DCAT can robustly transfer cross-modal alignment across heterogeneous real-world MLLMs.

Considered, but not experimented models. Our merging setup assumes that the modality-specific MLLMs share the same language-model architecture and parameter shapes, since both DCAT and the baseline merging methods operate directly at the weight level. We also considered other in-the-wild MLLMs that appear to share the same backbone family, such as Qwen-Audio [12] and LLaVA-Next [35] with Qwen-1.5 [3] backbones. However, these models often differ in tokenizer size, input embedding size, and LM head size, which prevents direct weight-space merging across all compared methods without additional architectural modification or retraining. Similarly, several recent public MLLMs, including Idefics3-8B-Llama3[5] and Llama-3.1-Nemotron-Nano-VL-8B-V1[6], were considered but excluded because their backbone implementations or parameter shapes are not fully compatible with the other modality-specific models. We did not introduce additional training or ad-hoc resizing procedures to force compatibility, as this would confound the comparison among training-free merging methods. To the best of our knowledge, our experiments cover the publicly available in-the-wild MLLMs that satisfy the required architectural compatibility conditions; we provide a more detailed discussion of this limitation in Sec. G.

D.2.3 Benchmarks

To compare the overall performance across various tasks, we evaluated the baseline and merged models on multiple audio and video question-answering benchmarks. For audio, we used MMAU [69], AIR-Bench-Foundation [84], MuChoMusic [74], Clotho-AQA [40], and VocalSound [18]. For video, we used Video-MME [16], TempCompass [44], MVBench [37], MLVU [91], EgoSchema [47], and LongVideoBench [78].

MMAU and AIR-Bench-Foundation consist of 1,000 and 24,683 multiple-choice questions (MCQs), respectively, paired with audio clips covering natural sounds, speech, and music. MuChoMusic contains 1,187 MCQs designed to evaluate the music understanding capability of models. For Clotho-AQA, we used a subset consisting of 1,442 binary QA questions related to audio. VocalSound contains 3,591 questions for classifying six types of human vocal sounds. We used a custom prompt to prevent the model from generating answers outside the given choices.

Video-MME contains 2,700 MCQs based on 900 videos with diverse durations and categories. For TempCompass, we used the multiple-choice, caption-matching, and yes/no subsets, which contain 1,580, 1,503, and 2,453 questions, respectively. These subsets evaluate the temporal understanding capability of models across five categories: action, attribute change, direction, event order, and speed. MVBench consists of 20 diverse question sets for evaluating temporal perception, with 200 questions

[2] https://huggingface.co/modelscope/llava-llama3.1-8b

[3] https://huggingface.co/fixie-ai/ultravox-v0_4_1-llama-3_1-8b

[4] https://huggingface.co/weizhiwang/LLaVA-Video-Llama-3.1-8B

[5] https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3

[6] https://huggingface.co/nvidia/Llama-3.1-Nemotron-Nano-VL-8B-V1

in each set. MLVU consists of 2,174 MCQs that evaluate long-video understanding ability across various genres. For EgoSchema, we used the subset containing 500 long-form video MCQ pairs. LongVideoBench consists of 1,337 MCQs designed to evaluate long-context video understanding. Overall results for these benchmarks, except EgoSchema and LongVideoBench, are reported in Table 2, and detailed results for specific categories in MVBench and TempCompass are provided in Appendix F.2.

Additionally, we used GQA [25], MMBench [43], MMStar [8], POPE [38], and TextVQA [63] to evaluate the vision capabilities of the models. The GQA testdev-balanced split contains 12,578 real-world visual question-answering examples. MMBench (Dev_EN) and MMStar contain 1,164 and 1,500 questions, respectively, for evaluating various capabilities of vision-language models. POPE consists of 9,000 binary QA pairs designed to evaluate object hallucination. TextVQA (Val) consists of 5,000 questions for evaluating a model's ability to read text in images. Quantitative results for the vision benchmarks are provided in Appendix F.3.

Each benchmark was evaluated using the lmms-eval [33, 89] toolkit. However, some prompts or evaluation methods were customized for our evaluation. Example prompts and evaluation methods are summarized in Table A13.

E Additional Analysis

E.1 Additional correlation analysis

To further assess the generality of the relationship between the proposed Alignment score and modality performance, we extend the correlation analysis to a broader set of benchmarks. For vision, we additionally consider GQA [25], OCR-VQA [50], MathVision [72], and VizWiz [20]; for audio, we use VocalSound [18], MuCho-music [74], Clotho-AQA [40]; and for video, we use TempCompass [44], MVBench [37], and MLVU [91]. The experimental protocol follows Sec. D.1. In particular, we perform the analysis for both Audio+Vision and Video+Vision settings on both Vicuna and LLaMA 3.1 backbones. For each merged setting, we evaluate the benchmarks corresponding to each constituent modality. As in the main analysis, we examine the relationship between Alignment score and modality performance both layer-wise and at the representative middle layer (21st layer).

The layer-wise results are shown in Fig. A1- A2, and the corresponding 21st-layer regression and correlation results are presented in Fig. A3-A4. Across all additional benchmarks and backbone settings, we observe a consistent trend. The strongest positive correlation between the Alignment score and performance appears in the middle layers, in agreement with the observations from the main analysis. Moreover, both Pearson and Spearman correlation coefficients between alignment and accuracy are positive across all cases in the 21st layer, indicating that higher alignment is consistently associated with better benchmark accuracy. These results further support that the proposed Alignment score is not specific to a particular benchmark or backbone, but instead provides a robust indicator of model quality across diverse evaluation settings.

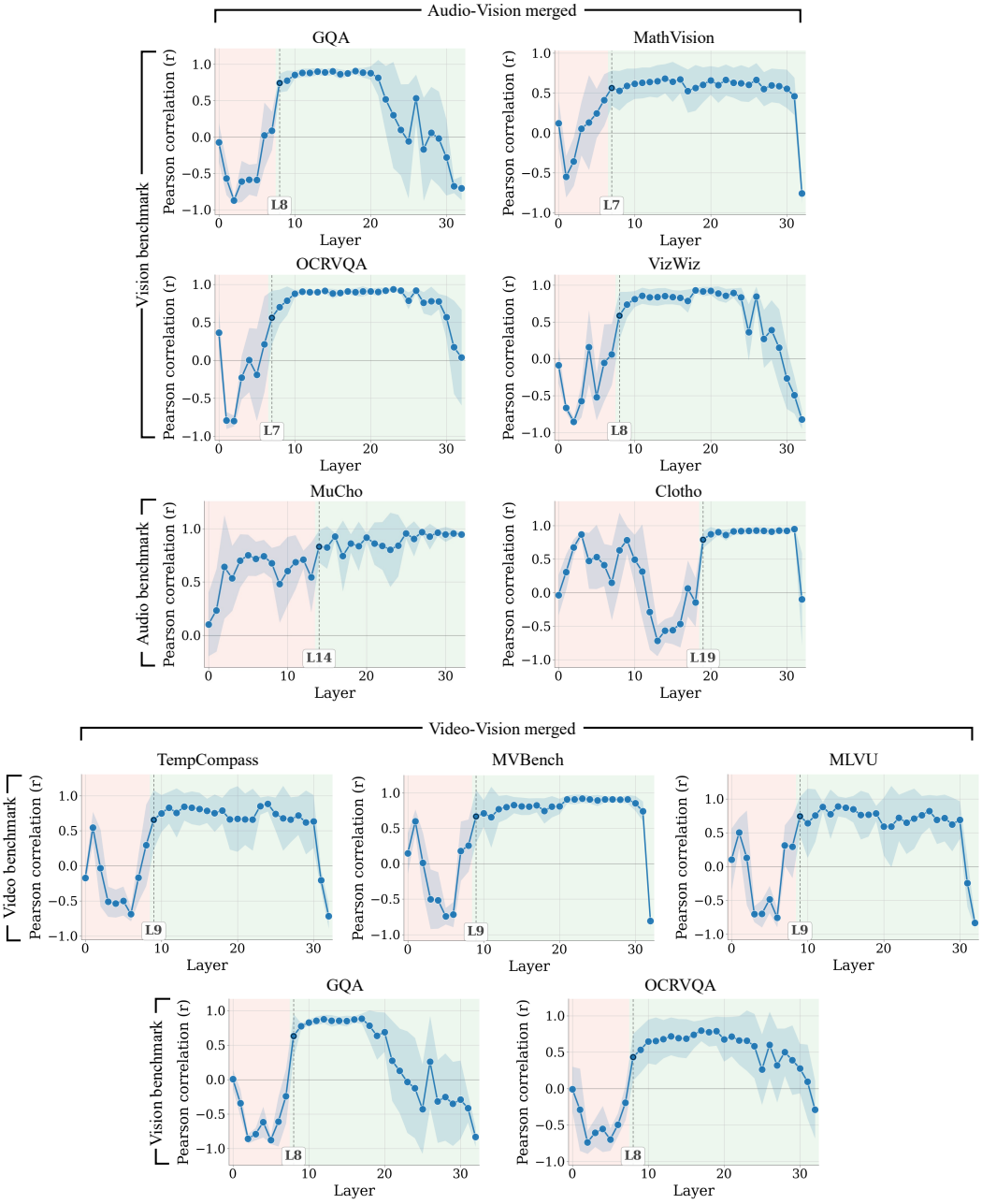

Figure A1: **Layer-wise Pearson correlation between Alignment score and modality performance on Vicuna backbone.** We measure the Alignment score in a layer-wise manner for vision, audio, and video benchmark on Audio+Vision and Video+Vision Vicuna backbone. Across all cases, the correlation between Alignment score and accuracy is strongest in the middle layers.

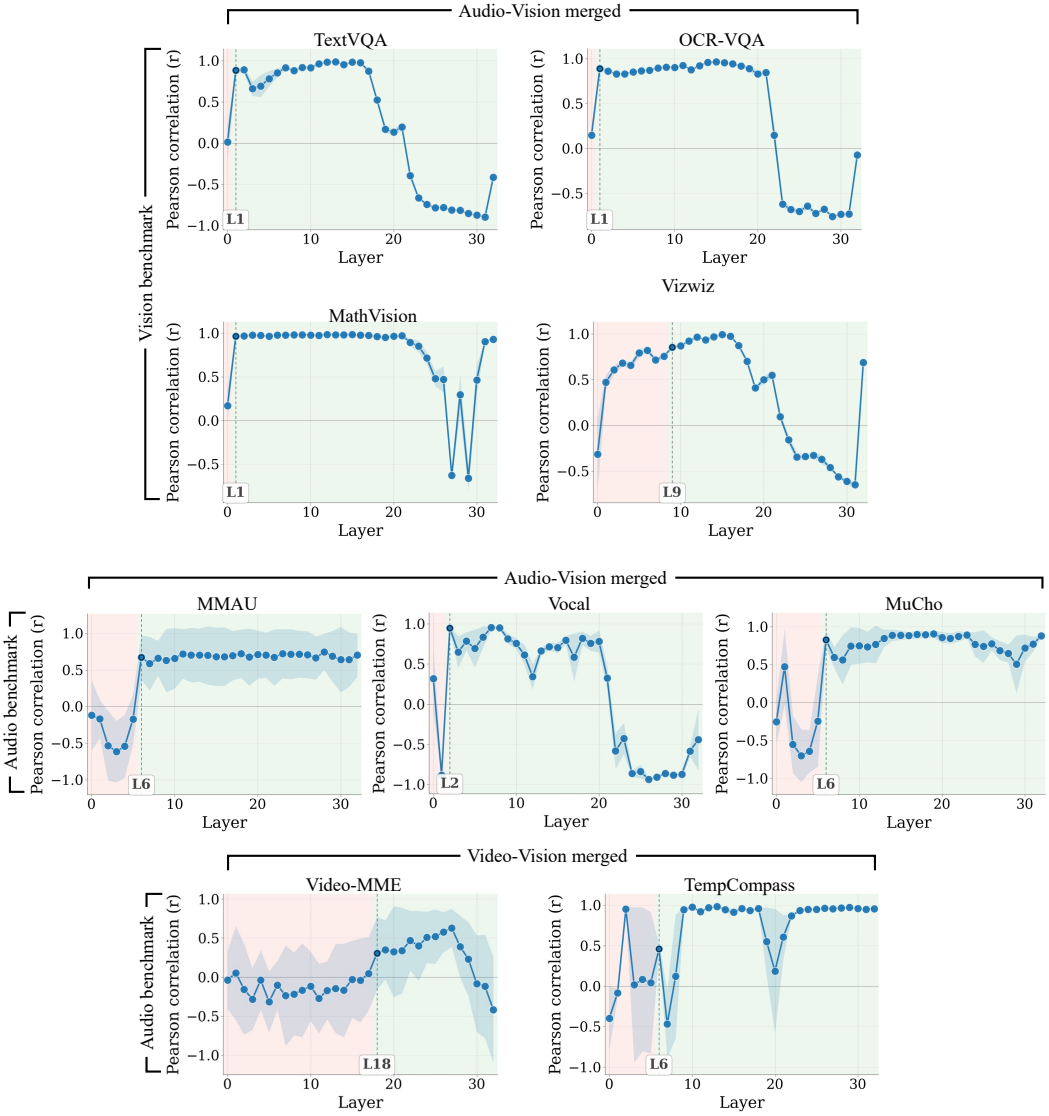

Figure A2: **Layer-wise Pearson correlation between Alignment score and modality performance on LLaMA-3.1.** We measure the Alignment score in a layer-wise manner for vision, audio, and video on Audio+Vision and Video+Vision LLaMa backbone. Across all cases, the correlation between the Alignment score and accuracy is strongest in the middle layers.

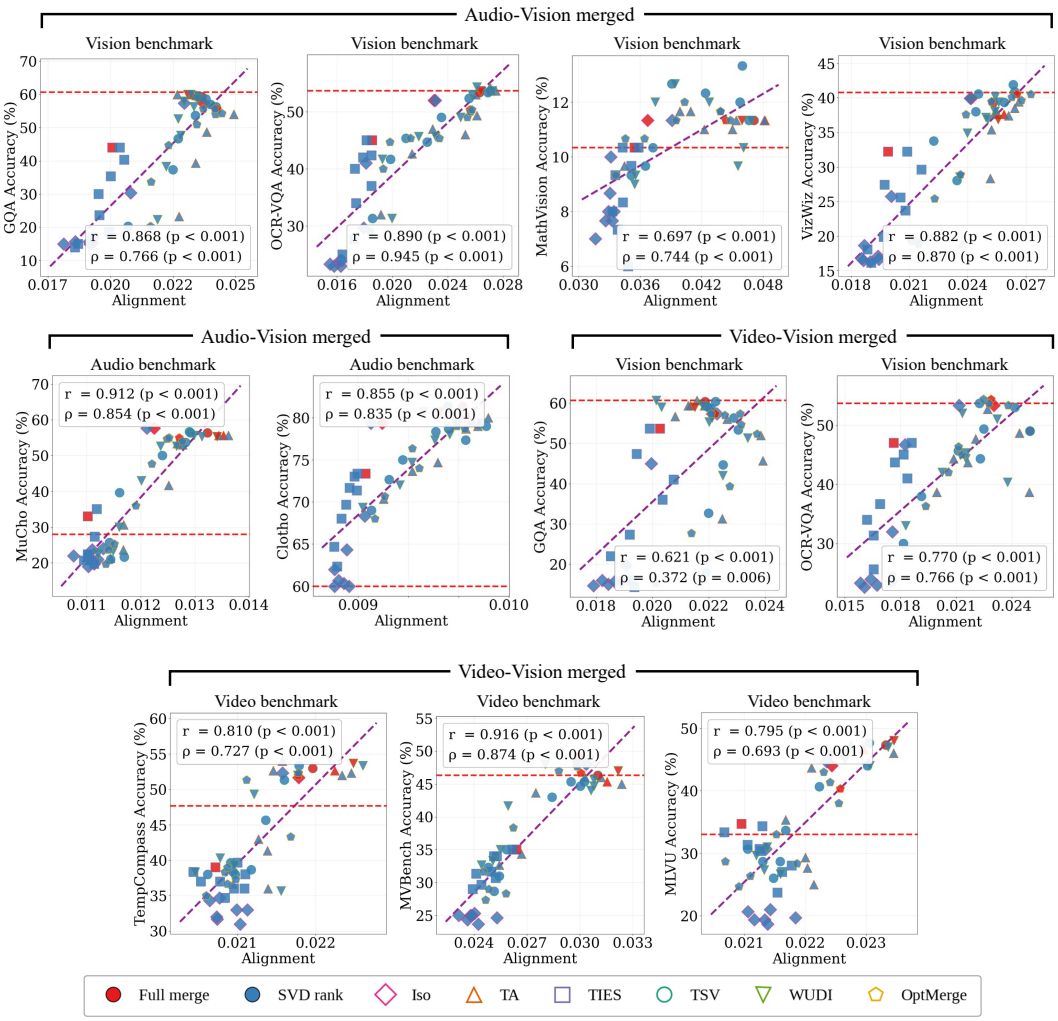

Figure A3: **Correlation between Alignment score and modality performance on the 21st layer of Vicuna backbone.** We analyze the correlation between the Alignment score and modality performance across checkpoints obtained from merging methods [9, 17, 28, 48, 75, 82], and their SVD-derived variants. The red dotted line refers to the performance of original model. Across all results, the Alignment score shows a positive correlation with modality performance under both Pearson and Spearman measures.

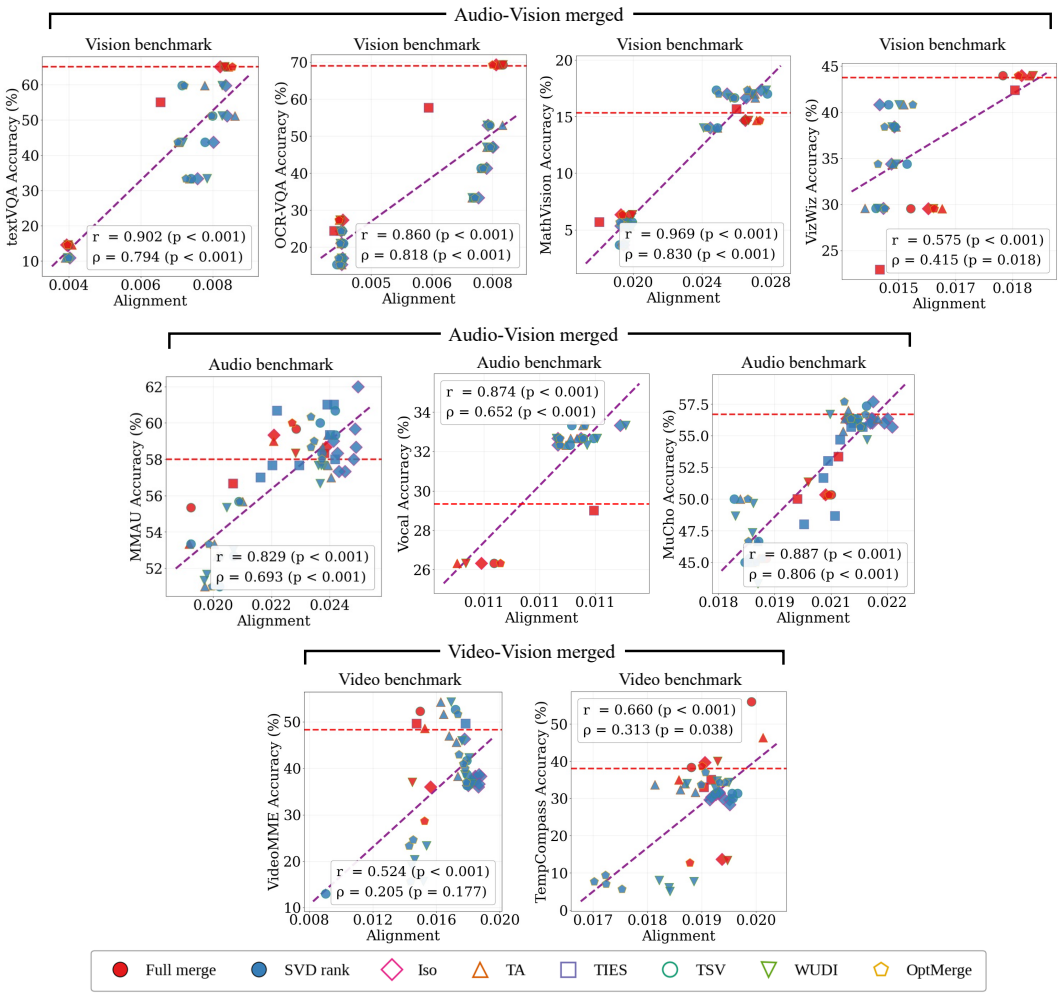

Figure A4: **Correlation between Alignment score and modality performance on the 21st layer of LLaMA-3.1 backbone.** We analyze the correlation between the Alignment score and modality performance across checkpoints obtained from merging methods [9, 17, 28, 48, 75, 82], and their SVD-derived variants. The red dotted line refers to the performance of original model. Across all results, the Alignment score shows a positive correlation with modality performance under both Pearson and Spearman measures.

E.2 Additional layer-wise SEO and SD analysis

To further examine how DCAT affects the internal structure of the merged model, we conduct an additional layer-wise analysis of SEO and SD at the level of individual transformer components. Specifically, for the Vicuna backbone, we compare SEO and SD before and after applying DCAT for each major submodule in both the MLP and attention blocks, including the MLP projections (`gate`, `down`, and `up`) and the attention projections (`q`, `k`, `v`, and `o`). This analysis is performed on the merged Vicuna models in the Audio+Vision and Video+Vision settings.

The results are shown in Fig. A5 and Fig. A6. Across all examined components and across both merged models, applying DCAT consistently increases both SEO and SD. This trend is observed not only at the model level but also throughout the individual layers and their constituent projections, indicating that the effect of DCAT is broadly distributed across the transformer rather than being confined to a small subset of parameters. These results provide additional empirical support for our theoretical motivation, showing that DCAT systematically encourages the structural properties associated with a higher mutual-information lower bound throughout the network.

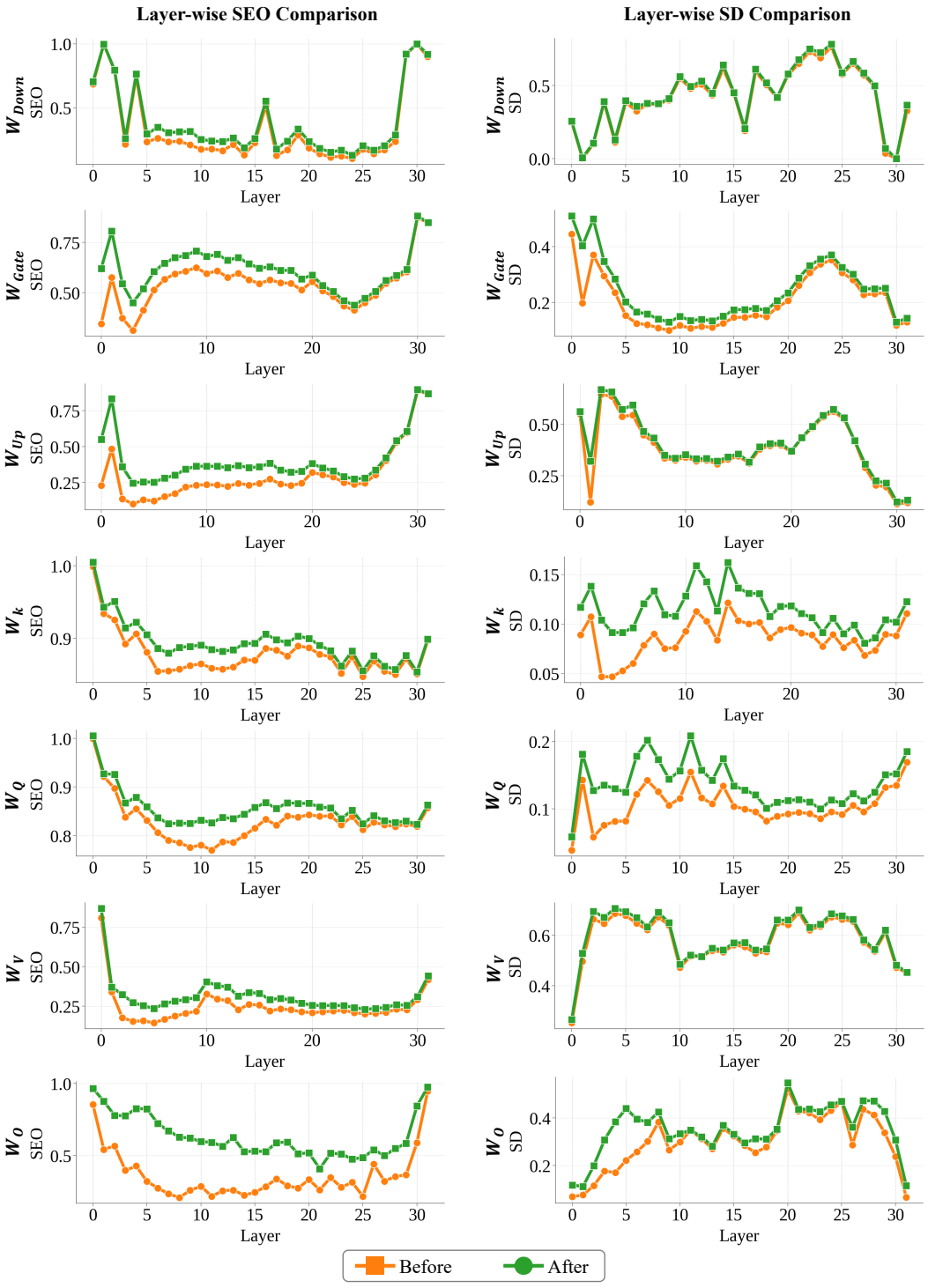

Figure A5: **Layer-wise SEO and SD of individual transformer submodules before and after applying DCAT on the Audio+Vision Vicuna.** We compare SEO and SD for the MLP projections (W_{gate}, W_{down}, and W_{up}) and the attention projections (W_k, W_q, W_v, and W_o) across all Vicuna layers. In all examined components, applying DCAT consistently increases both SEO and SD.

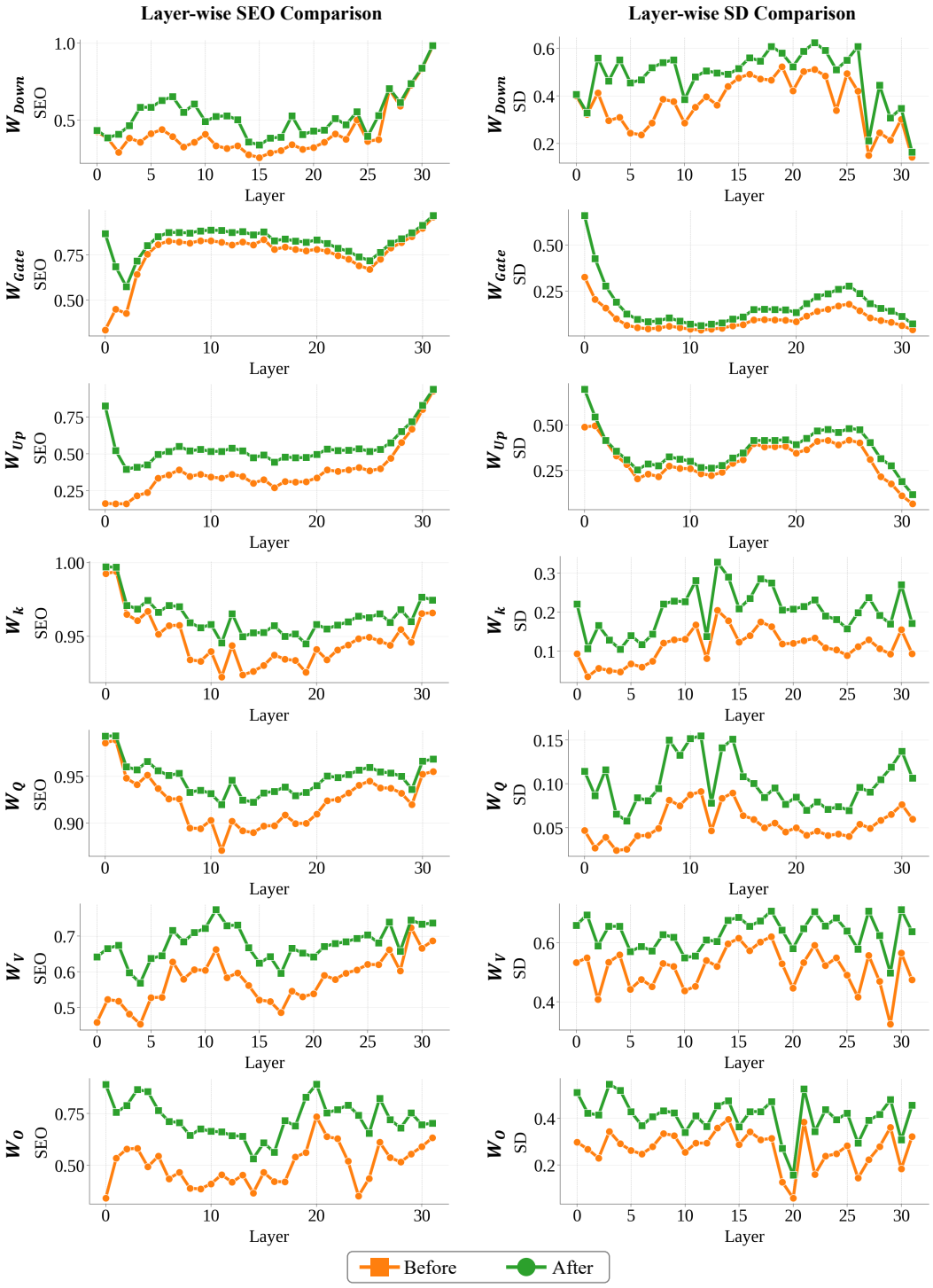

Figure A6: **Layer-wise SEO and SD of individual transformer submodules before and after applying DCAT on the Video+Vision Vicuna.** We compare SEO and SD for the MLP projections (W_{gate}, W_{down}, and W_{up}) and the attention projections (W_k, W_q, W_v, and W_o) across all Vicuna layers. In all examined components, applying DCAT consistently increases both SEO and SD.

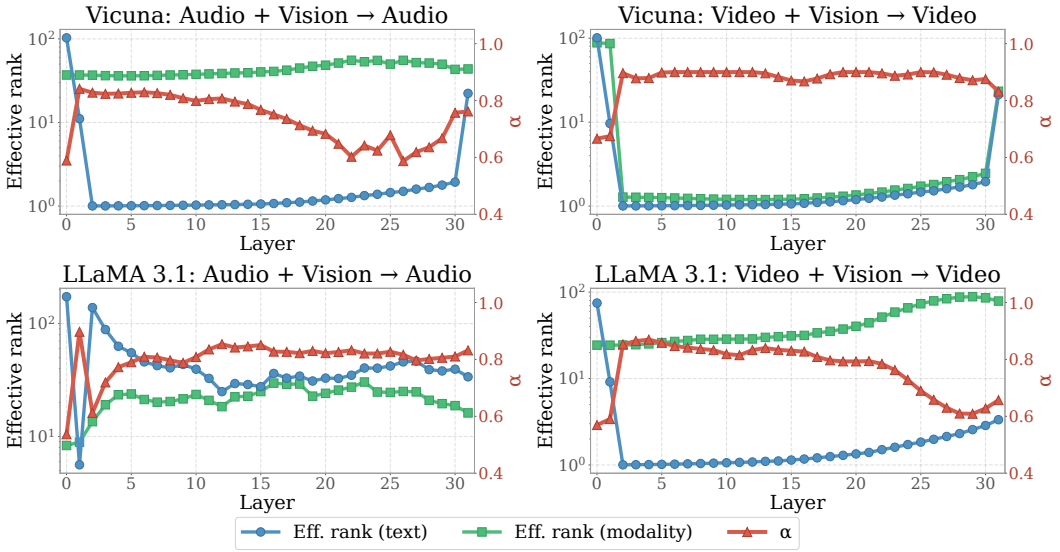

Figure A7: **Layer-wise effective rank and scheduling of α.** DCAT schedules α inversely with the text effective rank, as small text effective rank with large α could introduce noisy alignment between text and modality-token activations.

E.3 Rationale of scheduling α

DCAT adjusts α based on the relative changes in SEO and SD, as described in Eq. 11. This design is motivated by the layer-wise structure of the text subspace. As shown in Fig. A7, the effective rank of text-token activations generally decreases in the middle layers, although the degree of this trend varies across architectures and modalities. When the text subspace is effectively low-rank, aggressively increasing SD may spread modality information over weak or noisy text-space directions, leading to noisy alignment rather than reliable modality-text alignment. To mitigate this risk, DCAT uses the relative behavior of SEO and SD to adaptively schedule α, making the target more conservative when the reliable text subspace is limited.

Empirically, this schedule induces an inverse relationship between the text effective rank and α. In layers where the text effective rank is small, α becomes larger, encouraging DCAT to emphasize SEO and preserve the recipient's projected spectrum rather than over-promoting SD. Conversely, when the text effective rank is larger, α decreases and allows stronger encouragement of SD. Although the LLaMA 3.1-based audio model shows a relatively different effective-rank pattern compared with the other settings, the scheduled α still follows the intended inverse trend with respect to the text effective rank.

F Additional Results

F.1 Statistical significance of presented results

To assess whether the quantitative improvements reported in the main text are statistically significant, we additionally compute the performance gain relative to the original modality-specific model and the corresponding p-value for each benchmark in the Audio+Vision and Video+Vision Vicuna settings. Specifically, for each merging method ([9, 17, 28, 48, 75, 82]), including ours, we report the performance difference in percentage points (Δ, %p) with respect to the original model, together with its statistical significance.

We evaluate statistical significance using McNemar's test, which is appropriate in our setting because all compared models are evaluated on the same paired samples from each benchmark. Since each sample yields a matched pair of outcomes between two models (correct or incorrect), McNemar's test provides a natural way to determine whether the observed difference in accuracy is statistically meaningful beyond aggregate score differences.

Table A3: p-value of DCAT on audio+vision

		MMAU		AIR-Bench		MuCho		Clotho		Vocal	
		Δ (%p)	p-value	Δ (%p)	p-value	Δ (%p)	p-value	Δ (%p)	p-value	Δ (%p)	p-value
Vicuna-based	TA	+4.30	1.71e-02	+5.91	3.62e-61	+23.08	8.57e-29	+19.76	4.14e-36	-4.18	1.75e-07
	TIES	+2.10	1.78e-01	+5.50	3.74e-78	+11.46	7.79e-17	+13.73	1.83e-24	+1.95	1.55e-02
	ISO	+4.50	5.57e-03	+6.62	3.53e-88	+25.86	1.64e-41	+17.34	2.99e-33	-4.68	2.86e-13
	TSV	+3.80	2.80e-02	+7.34	3.15e-112	+23.84	1.65e-33	+19.97	4.21e-37	-1.45	4.50e-02
	WUDI	+4.80	6.74e-03	+6.06	5.07e-64	+22.49	1.35e-27	+19.49	1.09e-34	-2.01	9.10e-03
	OptMerge	+4.20	7.74e-03	+7.33	6.15e-127	+23.25	5.65e-37	+19.00	1.97e-36	+0.08	9.39e-01
	DCAT (Ours)	+7.30	4.12e-04	+9.27	2.42e-116	+27.21	7.82e-39	+21.73	2.27e-43	+2.27	5.83e-03
LLaMA 3.1-based	TA* ($\lambda = 1.0$)	-2.90	8.41e-02	-3.18	2.78e-17	-2.95	1.01e-01	+1.60	2.62e-01	-3.45	3.47e-04
	TIES ($\lambda = 1.0$)	+0.50	7.85e-01	-0.43	2.17e-01	+2.36	1.67e-01	+2.08	5.99e-02	-0.53	4.11e-01
	ISO ($\lambda = 1.0$)	+2.10	1.52e-01	-2.23	5.00e-11	+3.37	4.42e-02	+3.19	1.48e-03	-1.67	1.26e-06
	TSV* ($\lambda = 1.0$)	-2.90	8.41e-02	-3.18	2.78e-17	-2.95	1.01e-01	+1.60	2.62e-01	-3.45	3.47e-04
	WUDI ($\lambda = 1.0$)	-1.90	2.69e-01	-3.30	2.47e-18	-2.61	1.43e-01	+1.80	2.12e-01	-4.12	1.80e-05
	OptMerge* ($\lambda = 1.0$)	-2.90	8.41e-02	-3.18	2.78e-17	-2.95	1.01e-01	+1.60	2.62e-01	-3.45	3.47e-04
	TA* ($\lambda = 0.2$)	+1.70	2.21e-01	-0.04	8.92e-01	+1.94	2.17e-01	+2.08	7.85e-03	+0.19	6.84e-01
	TIES ($\lambda = 0.2$)	+2.10	1.01e-01	+0.51	6.96e-02	+2.02	1.17e-01	+1.18	7.48e-02	-0.45	1.09e-01
	ISO ($\lambda = 0.2$)	+0.60	6.75e-01	+0.64	1.54e-02	+0.67	4.75e-01	+2.01	9.14e-04	-0.84	2.88e-03
	TSV* ($\lambda = 0.2$)	+1.70	2.21e-01	-0.04	8.92e-01	+1.94	2.17e-01	+2.08	7.85e-03	+0.19	6.84e-01
	WUDI ($\lambda = 0.2$)	+0.20	9.39e-01	+0.02	9.57e-01	+0.17	9.50e-01	+2.01	1.49e-02	-1.11	3.11e-02
	OptMerge* ($\lambda = 0.2$)	+1.70	2.21e-01	-0.04	8.92e-01	+1.94	2.17e-01	+2.08	7.85e-03	+0.19	6.84e-01
	DCAT (Ours)	+3.66	5.20e-03	+2.07	3.41e-14	+5.73	1.73e-05	+6.08	2.34e-13	+0.92	5.33e-03

Table A4: p-value of DCAT on video+vision

		Video-MME		TempCompass		MVBench		MLVU	
		Δ (%p)	p-value	Δ (%p)	p-value	Δ (%p)	p-value	Δ (%p)	p-value
Vicuna-based	TA	+2.11	7.31e-02	-0.07	9.47e-01	+5.20	1.30e-07	+12.37	3.29e-20
	TIES	-6.70	7.38e-16	-0.83	1.81e-01	-5.70	2.99e-17	+0.41	6.36e-01
	ISO	+3.44	8.96e-04	-0.07	9.43e-01	+4.53	1.74e-06	+13.06	2.53e-21
	TSV	+4.04	1.33e-04	+1.79	1.13e-02	+6.10	2.25e-10	+12.97	1.32e-28
	WUDI	+1.56	1.81e-01	-0.52	5.39e-01	+4.83	8.67e-07	+12.05	1.21e-19
	OptMerge	+3.30	6.17e-05	+1.01	6.10e-02	+6.35	5.89e-14	+10.35	6.89e-24
	DCAT (Ours)	+6.53	3.06e-08	+4.25	1.11e-06	+9.17	2.93e-20	+13.75	6.69e-39
LLaMA 3.1-based	TA* ($\lambda = 1.0$)	+0.70	3.94e-01	+1.82	1.18e-02	+2.88	1.14e-04	+11.55	5.43e-31
	TIES ($\lambda = 1.0$)	+3.30	1.15e-04	-4.06	4.81e-08	+5.88	1.44e-13	+13.75	1.74e-39
	ISO ($\lambda = 1.0$)	+2.70	2.30e-03	-22.44	1.88e-166	+5.03	3.54e-09	+11.96	1.18e-29
	TSV* ($\lambda = 1.0$)	+3.67	8.75e-06	+5.49	6.86e-15	+6.00	5.26e-15	+14.54	1.27e-45
	WUDI ($\lambda = 1.0$)	+1.96	1.49e-02	-40.75	0.00e+00	+4.35	3.66e-08	+13.89	6.65e-40
	OptMerge* ($\lambda = 1.0$)	+4.67	1.20e-07	-47.69	0.00e+00	+6.23	4.68e-14	+15.59	6.01e-47
	TA* ($\lambda = 0.2$)	+4.74	1.16e-06	-8.13	2.98e-24	+2.18	1.44e-02	+9.43	3.83e-17
	TIES ($\lambda = 0.2$)	+1.70	9.89e-02	-6.36	4.20e-15	+4.43	3.71e-07	+5.80	6.91e-06
	ISO ($\lambda = 0.2$)	+1.33	2.07e-01	+5.91	5.03e-14	+3.80	8.24e-06	+11.64	2.07e-20
	TSV* ($\lambda = 0.2$)	+3.26	1.17e-03	+2.08	7.87e-03	+4.65	1.02e-07	+7.54	6.81e-10
	WUDI ($\lambda = 0.2$)	+4.11	3.67e-05	+7.15	6.29e-21	+3.55	7.58e-05	+11.45	9.42e-22
	OptMerge* ($\lambda = 0.2$)	+2.41	1.91e-02	+6.02	8.59e-15	+5.48	5.69e-10	+9.25	8.34e-14
	DCAT (Ours)	+6.25	3.85e-09	+8.60	7.24e-23	+10.18	6.03e-30	+18.36	5.51e-57

The results in Table A3 for Audio+Vision and Table A4 for Video+Vision show that our method yields statistically significant improvements across all reported settings. In the tables, p-values below 0.05 are highlighted in yellow, and those below 0.01 are highlighted in red. While baseline merging methods sometimes improve performance, their gains are not consistently significant and vary depending on the benchmark and backbone. In contrast, our method achieves positive improvements with statistically significant p-values in all cases, demonstrating that the performance gains reported in the main text are not only numerically favorable but also consistently supported at the sample level.

F.2 Additional quantitative results

For some of the results presented in Tab.2, we reported only the overall or average scores rather than detailed subcategory results due to space constraints. In this section, we provide these detailed subcategory results, as well as additional benchmark results on EgoSchema[47] and LongVideoBench [78]. The results are reported in Tab.A5 and Tab.A6. These results show that DCAT consistently improves performance across diverse subcategories and additional benchmarks, while maintaining a clear performance advantage over other merging methods.

Table A5: **Detailed results for MVBench benchmark.** We report the specific performance measured from subcategories on MVBench.

		Action Antonym	Action Count	Action Localization	Action Prediction	Action Sequence	MVBench Character Order	Counterfactual Inference	Egocentric Navigation	Episodic Reasoning	Fine Grained Action
Vicuna-based	Original	51.50	38.00	27.50	29.00	35.00	47.00	34.00	27.00	32.00	40.00
	TA	62.00	32.50	36.00	46.00	48.50	45.00	34.00	34.00	40.50	38.00
	TIES	55.50	16.50	26.00	24.50	32.00	28.50	31.50	25.00	22.50	30.50
	ISO	54.50	38.00	28.00	49.00	46.00	46.50	41.00	34.00	37.50	41.00
	TSV	60.50	36.50	31.00	48.00	50.00	46.00	38.50	34.00	41.00	40.50
	WUDI	63.00	32.00	34.00	46.50	50.00	46.50	35.00	32.00	38.50	39.50
	OptMerge	61.50	36.50	28.50	49.00	52.50	44.00	43.00	31.50	43.50	38.50
	DCAT (Ours)	65.00	37.00	37.00	49.00	53.00	49.00	46.00	34.00	45.50	42.00
LLaMA 3.1-based	Original	61.50	18.00	30.00	47.00	45.00	50.00	39.50	21.00	33.50	36.50
	TA* ($\lambda = 1.0$)	71.00	45.50	26.00	44.00	45.50	49.00	31.00	25.00	44.50	36.50
	TIES ($\lambda = 1.0$)	75.00	39.50	27.50	46.00	50.00	55.50	32.50	33.50	47.00	40.50
	ISO ($\lambda = 1.0$)	64.50	44.50	27.50	47.00	50.00	52.00	36.00	37.00	46.00	40.00
	TSV* ($\lambda = 1.0$)	76.50	41.00	28.50	45.50	50.50	55.00	36.00	26.00	45.00	39.50
	WUDI ($\lambda = 1.0$)	69.00	42.00	30.50	45.00	53.00	55.50	30.00	29.50	46.00	38.00
	OptMerge* ($\lambda = 1.0$)	71.00	36.50	32.50	45.00	53.00	57.00	33.00	34.50	46.00	41.50
	TA* ($\lambda = 0.2$)	63.50	28.00	24.00	45.00	54.00	44.50	35.50	1.50	45.00	38.50
	TIES ($\lambda = 0.2$)	64.50	27.50	21.00	44.00	57.00	49.50	32.50	36.50	49.50	36.50
	ISO ($\lambda = 0.2$)	58.00	27.50	25.50	40.50	56.50	48.00	40.00	36.50	51.00	35.50
	TSV* ($\lambda = 0.2$)	62.00	28.00	24.50	40.50	59.00	50.00	39.50	38.00	50.50	36.50
	WUDI ($\lambda = 0.2$)	61.50	23.00	27.50	39.00	55.00	49.50	35.50	37.50	49.50	38.50
	OptMerge* ($\lambda = 0.2$)	64.00	28.50	26.50	42.50	58.50	49.50	40.00	37.50	50.50	34.50
	DCAT (Ours)	75.50	48.50	33.50	48.50	60.00	56.00	39.50	39.50	51.50	42.00

		Fine Grained Pose	Moving Attribute	Moving Count	Moving Direction	Object Existence	MVBench Object Interaction	Object Shuffle	Scene Transition	State Change	Unexpected Action	Average
Vicuna-based	Original	28.00	43.50	22.00	27.50	47.00	52.50	38.50	65.50	39.50	44.50	38.48
	TA	33.00	47.00	30.00	30.00	54.50	54.00	34.50	82.00	45.00	47.00	43.68
	TIES	31.00	42.00	20.50	19.50	39.00	47.00	30.00	54.00	39.00	41.00	32.78
	ISO	29.50	47.50	34.50	27.00	53.00	51.50	35.00	78.50	39.00	49.00	43.00
	TSV	32.00	49.50	29.00	31.50	52.50	55.00	37.00	82.50	46.50	50.00	44.58
	WUDI	30.50	47.00	27.50	28.50	55.00	52.00	34.00	82.00	45.00	47.50	43.30
	OptMerge	33.50	49.00	28.50	31.50	47.00	56.50	37.00	83.00	46.50	55.50	44.83
	DCAT (Ours)	35.00	51.00	36.50	33.00	54.50	58.50	41.00	83.00	48.50	54.50	47.65
LLaMA 3.1-based	Original	34.50	45.50	31.00	25.00	51.00	58.50	31.50	79.00	38.50	43.50	41.00
	TA* ($\lambda = 1.0$)	35.50	48.00	38.00	22.50	53.00	54.00	29.50	82.00	44.50	52.50	43.88
	TIES ($\lambda = 1.0$)	34.00	47.00	32.00	30.50	53.50	67.00	35.50	87.00	45.50	58.50	46.88
	ISO ($\lambda = 1.0$)	30.50	46.00	36.50	31.00	48.00	55.50	32.00	88.00	43.50	65.00	46.03
	TSV* ($\lambda = 1.0$)	36.50	48.00	34.50	33.00	53.00	64.00	34.00	87.00	46.00	60.50	47.00
	WUDI ($\lambda = 1.0$)	35.00	44.50	36.00	27.00	53.50	59.50	30.00	83.50	41.50	58.00	45.35
	OptMerge* ($\lambda = 1.0$)	36.00	45.00	35.00	32.50	53.00	64.50	35.50	87.00	45.00	61.00	47.23
	TA* ($\lambda = 0.2$)	36.00	46.00	38.00	34.00	48.50	65.00	30.00	76.00	45.50	65.00	43.18
	TIES ($\lambda = 0.2$)	37.00	46.50	40.00	35.00	48.00	63.50	31.50	78.00	43.50	67.00	45.43
	ISO ($\lambda = 0.2$)	35.00	45.50	35.50	22.50	49.00	59.50	33.50	89.00	40.00	67.50	44.80
	TSV* ($\lambda = 0.2$)	37.00	45.50	37.00	26.00	48.00	63.50	32.00	82.50	45.00	68.00	45.65
	WUDI ($\lambda = 0.2$)	35.50	44.00	38.50	28.00	44.50	63.00	28.50	87.00	40.00	65.50	44.55
	OptMerge* ($\lambda = 0.2$)	36.50	44.00	39.50	34.00	45.50	64.50	33.00	88.00	42.50	70.00	46.48
	DCAT (Ours)	39.00	49.00	41.00	35.00	53.00	68.00	36.50	90.00	47.00	70.50	51.18

F.3 Results on vision benchmarks

We use vision MLLMs as donor models to improve the performance of audio and video MLLMs. A natural question is how the vision performance of the resulting transferred models changes after DCAT. We investigate this question using Vicuna-based models, and report the results in Tab.A7. We evaluate on a diverse set of vision benchmarks, including GQA[25], MMBench [43], MMStar [8], POPE [38], and Text-VQA [63]. When the audio MLLM is used as the recipient, the vision performance is largely preserved or slightly improved, indicating only minor perturbations. In contrast, when the video MLLM is used as the recipient, we observe some degree of degradation in vision performance. This behavior is expected because DCAT transfers alignment rather than modality-specific capabilities. We emphasize that our goal is not to obtain a single model that performs well on both modalities, but rather to transfer the capability of a strongly aligned MLLM from one modality to another in order to improve performance on the target modality.

F.4 Results on audio-video DCAT

Beyond using a vision MLLM as the donor in DCAT, we also conducted experiments with Vicuna-based models in two additional settings: Video-to-Audio, where a video MLLM is used as the donor and an audio MLLM as the recipient, and Audio-to-Video, where an audio MLLM is used as the donor and a video MLLM as the recipient. The results are reported in Tab.A8. In the Video-to-Audio setting, DCAT improves the performance of the audio MLLM, whereas the improvement in the Audio-to-Video setting is limited. We hypothesize that this discrepancy arises from differences in the

Table A6: **Detailed results for MLVU, TempCompass benchmark and other benchmarks.** We report the specific performance measured from subcategories on MLVU, TempCompass(Multi Choice, Caption Matching, Yes/No), and other two benchmarks(EgoSchema, LongVideoBench).

		MLVU								TempCompass - Multi Choice					
		Count	Topic Reasoning	Ego	Order	PlotQA	Needle	Anomaly Reco.	Average	Action	Attr.	Dir.	Order	Speed	Overall
Vicuna-based	Original	27.18	57.79	28.41	23.17	30.80	33.80	27.00	32.59	70.41	35.07	33.13	37.75	32.81	42.28
	TA	34.47	65.40	34.66	23.17	50.09	45.07	61.00	44.84	82.84	45.14	37.01	38.74	35.33	48.29
	TIES	25.73	58.56	27.27	23.17	32.47	31.83	33.00	33.15	63.61	33.33	32.54	35.76	30.60	39.56
	ISO	30.10	67.30	42.33	22.39	49.35	44.23	61.50	45.31	81.66	46.18	35.52	39.40	35.65	48.10
	TSV	33.98	70.72	42.90	23.94	48.98	42.25	53.50	45.18	86.69	47.57	36.42	42.38	35.02	50.06
	WUDI	33.01	65.78	35.51	23.17	50.28	43.38	59.50	44.38	82.25	45.49	36.12	40.07	36.28	48.48
	OptMerge	30.58	71.10	36.08	25.10	43.60	40.56	56.00	43.29	79.29	44.10	35.22	37.75	38.17	47.34
	DCAT (Ours)	35.74	72.37	44.24	25.73	51.73	46.04	62.33	46.34	86.27	48.16	38.06	43.54	38.45	51.62
LLaMA 3.1-based	Original	3.40	50.19	34.66	16.22	43.04	35.49	39.50	31.79	87.57	46.88	36.12	37.09	35.96	49.24
	TA* ($\lambda = 1.0$)	37.86	70.72	43.75	23.94	47.50	44.79	48.00	45.22	86.98	45.14	33.73	37.09	36.91	48.48
	TIES ($\lambda = 1.0$)	38.35	73.00	46.02	23.17	51.21	47.61	50.50	47.12	91.12	49.31	34.03	35.10	42.27	50.89
	ISO ($\lambda = 1.0$)	24.27	67.30	42.05	25.10	49.17	49.58	59.50	45.28	67.46	38.19	26.27	18.21	17.03	33.86
	TSV* ($\lambda = 1.0$)	38.35	74.52	47.73	23.55	51.02	47.89	53.50	48.08	90.53	48.61	33.13	34.77	43.85	50.70
	WUDI ($\lambda = 1.0$)	38.35	70.72	46.31	23.55	50.09	49.01	54.50	47.51	17.16	8.68	6.57	4.64	6.62	8.86
	OptMerge* ($\lambda = 1.0$)	39.32	74.14	48.01	26.25	51.95	49.01	56.00	49.24	6.51	5.21	4.18	0.99	2.21	3.86
	TA* ($\lambda = 0.2$)	13.11	70.72	47.16	14.67	50.46	48.45	42.00	40.94	93.20	48.96	34.93	36.76	47.00	52.72
	TIES ($\lambda = 0.2$)	21.36	54.37	51.14	11.97	41.19	42.54	47.50	38.58	89.65	43.75	28.96	31.79	37.85	46.96
	ISO ($\lambda = 0.2$)	21.84	73.38	49.15	23.55	47.68	46.76	49.00	44.48	90.53	48.61	34.03	37.09	42.27	51.01
	TSV* ($\lambda = 0.2$)	17.48	68.82	49.15	13.51	43.04	41.97	49.00	40.42	92.60	48.96	33.73	34.11	41.64	50.76
	WUDI ($\lambda = 0.2$)	13.59	75.29	48.30	21.62	48.24	50.14	49.50	43.81	92.31	49.65	36.42	38.08	43.53	52.53
	OptMerge* ($\lambda = 0.2$)	17.96	69.20	48.58	18.92	45.83	45.35	47.00	41.83	91.72	47.57	31.94	35.43	42.27	50.32
	DCAT (Ours)	37.38	76.81	51.70	27.03	52.69	50.99	54.50	50.15	93.58	50.50	37.75	39.25	47.90	53.15

		TempCompass - Caption Matching						TempCompass - Yes/No						Others	
		Action	Attr.	Dir.	Order	Speed	Overall	Action	Attr.	Dir.	Order	Speed	Overall	EgoSchema	LongVideo Bench
Vicuna-based	Original	85.86	59.38	55.96	49.67	59.11	61.88	79.39	62.72	54.05	53.66	46.37	60.01	48.60	39.04
	TA	67.00	54.17	54.43	51.67	47.08	54.89	81.08	57.59	48.58	55.24	53.82	60.25	48.80	37.62
	TIES	86.87	58.33	49.54	48.67	52.92	59.08	78.72	57.14	53.85	56.28	57.36	61.60	38.20	35.90
	ISO	80.47	57.29	52.60	47.00	55.33	58.42	77.53	56.47	49.80	51.57	50.84	58.21	47.20	38.07
	TSV	82.49	62.15	56.57	51.00	56.01	61.54	80.57	56.92	49.60	53.40	50.65	59.23	51.20	40.02
	WUDI	64.65	55.21	52.91	49.67	45.70	53.63	80.07	58.04	48.38	54.71	53.45	59.89	49.60	38.30
	OptMerge	81.48	57.99	53.21	46.33	57.39	59.15	81.08	61.38	51.82	55.50	49.53	60.70	47.60	40.46
	DCAT (Ours)	86.08	60.67	57.23	51.28	59.77	62.67	81.24	63.56	55.61	56.23	59.02	63.11	52.05	40.54
LLaMA 3.1-based	Original	78.11	32.99	48.32	41.00	32.30	46.71	76.35	50.45	49.60	47.38	47.49	55.40	37.20	40.17
	TA* ($\lambda = 1.0$)	67.00	44.79	42.51	47.67	42.27	48.77	78.04	55.36	50.00	50.79	54.00	58.74	41.80	40.31
	TIES ($\lambda = 1.0$)	58.59	41.67	41.28	37.67	38.83	43.58	73.48	43.53	47.17	39.01	26.63	47.09	48.40	44.35
	ISO ($\lambda = 1.0$)	54.55	29.17	29.05	32.00	37.46	36.33	25.34	29.91	25.10	10.47	12.66	21.04	41.20	45.25
	TSV* ($\lambda = 1.0$)	82.49	54.86	57.19	56.33	54.30	61.01	79.56	54.24	48.79	50.00	51.96	58.09	46.60	44.35
	WUDI ($\lambda = 1.0$)	2.36	1.74	0.61	0.33	0.69	1.13	29.39	20.54	20.24	6.02	6.89	17.37	34.80	42.78
	OptMerge* ($\lambda = 1.0$)	0.67	0.35	1.53	0.00	0.69	0.67	7.94	5.58	6.28	4.19	1.68	5.22	37.20	45.10
	TA* ($\lambda = 0.2$)	49.49	37.85	36.09	45.00	46.05	42.78	53.55	31.25	37.45	28.01	30.54	37.22	49.20	48.09
	TIES ($\lambda = 0.2$)	53.20	31.25	32.72	32.33	42.27	38.26	57.94	38.17	45.75	42.67	49.72	47.70	44.60	46.67
	ISO ($\lambda = 0.2$)	87.21	56.94	52.60	56.00	62.54	62.87	77.03	51.79	52.63	47.38	53.26	57.68	42.80	46.30
	TSV* ($\lambda = 0.2$)	75.76	50.00	44.34	55.00	59.11	56.62	72.97	45.09	49.39	42.93	48.23	53.04	46.80	48.17
	WUDI ($\lambda = 0.2$)	93.94	64.58	56.88	62.00	63.23	67.93	77.87	48.88	49.39	46.86	52.33	56.42	47.40	48.54
	OptMerge* ($\lambda = 0.2$)	93.60	63.54	56.88	61.00	67.35	68.26	74.66	47.10	50.00	45.29	51.77	55.08	47.00	47.64
	DCAT (Ours)	94.25	65.36	58.10	63.35	68.15	69.14	78.30	52.70	54.05	48.50	54.25	58.54	49.80	49.06

intrinsic alignment quality of the donor model, which is consistent with the conjecture discussed in Sec.6 of the main manuscript.

F.5 Calibration-matched baselines

We compare DCAT with LoRA fine-tuning (LoRA rank=256, batch size=8, 128 steps), RegMean [30], and RegMean++ [27] under a matched calibration budget of 256 AVQA samples. Tab. A9 report results on Vicuna-based models. DCAT outperforms the strongest calibration-aware baseline on both MMAU and Video-MME (by 1.9 and 1.3 points overall, respectively) across all categories in each benchmark, suggesting that DCAT's gains arise from its alignment-transfer objective, which leverages the donor model to jointly improve subspace overlap, spectral diversity, and text-compatible directions.

F.6 Calibration sample robustness

To study the robustness of DCAT with respect to the calibration sample, we randomly sampled 256 AVQA training examples using 10 different random seeds and applied DCAT to both Audio+Vision and Video+Vision settings. The results in Tab. A10 show low overall standard deviations of 0.33 on MMAU and 0.14 on VideoMME across calibration subsets, demonstrating that DCAT is robust to calibration sampling and random seed within AVQA.

Table A7: **Results on vision benchmarks (Vicuna-based).** The performance of Vicuna based audio-vision(A+V) and video-vision(Vd+V) models are measured from five different benchmarks: GQA, MMBench-Dev(EN), MMStar, POPE, TextVQA-Val.

	GQA		MMBench		MMStar		POPE		Text-VQA		Average	
Original	61.91		63.83		35.44		84.31		47.76		58.65	
Merging Method	A+V	Vd+V	A+V	Vd+V	A+V	Vd+V	A+V	Vd+V	A+V	Vd+V	A+V	Vd+V
TA	61.76	58.08	62.97	60.40	35.62	34.69	85.34	84.83	49.10	45.39	58.96	56.68
TIES	43.89	53.02	57.99	59.36	33.98	33.61	79.56	85.06	42.99	46.26	51.68	55.46
ISO	58.75	57.62	61.43	60.74	34.07	32.48	85.31	82.51	48.15	46.89	57.54	56.05
TSV	60.92	60.75	63.83	63.66	34.26	33.58	86.02	85.58	49.86	47.87	58.98	58.29
WUDI	61.64	58.28	63.06	60.82	35.39	33.97	85.32	84.36	48.88	45.71	58.86	56.63
OptMerge	58.08	59.34	62.71	62.20	34.06	33.82	86.08	86.20	49.35	48.71	58.06	58.05
DCAT (Ours)	61.43	40.62	63.75	51.46	33.53	35.06	85.13	61.08	49.30	24.34	58.63	42.51

Table A8: **Audio-video cross-modal transfer results.** We report DCAT performance when transferring alignment between audio and video models. The left table evaluates video-to-audio transfer on MMAU [69], while the right table evaluates audio-to-video transfer on Video-MME [16].

MMAU (Video-to-Audio)

Model	Music	Sound	Speech	Overall
Original	54.79	60.66	31.23	48.90
DCAT	57.49	66.07	37.24	53.60

Video-MME (Audio-to-Video)

Model	Short	Medium	Long	Overall
Original	39.89	34.22	30.89	35.00
DCAT	37.78	33.78	29.33	33.63

F.7 Language ability on alignment-transferred models

We evaluate general language ability after DCAT on 13 text benchmarks [5, 13, 14, 21, 32, 49, 55, 57, 58, 66, 76, 87, 90] for Vicuna-based Audio+Vision and Video+Vision models in Tab. A11. McNemar's test shows no significant change for Audio+Vision ($p = 0.506$) and a significant improvement for Video+Vision ($p \ll 0.01$), indicating no degradation.

F.8 Additional ablation study

In this section, in addition to the component-level ablation studies conducted in Tab. 3, we present ablation studies on key hyperparameters. Specifically, we study the balancing coefficient β in $\mathcal{L}_{\text{cross}}(\tau)$, the balancing coefficient η in Eq. 10, the subspace dimension k used to construct P_k, and the size of the calibration dataset. All experiments are conducted with Vicuna-based models. For each hyperparameter, we select the sweep range of the parameters to the values where marked performance degradation is observed on target-modality benchmarks such as MMAU and Video-MME. The results of these additional ablation studies are reported in Tab. A12. We observe that appropriate choices of β and η are important for achieving strong performance. In contrast, DCAT maintains robust performance with respect to the subspace dimension k and the number of calibration samples. These experiments provide empirical rationale for our hyperparameter choices.

Table A13: **Summary of benchmarks used in our evaluation.** Each benchmark's question type, example prompt, number of questions, and evaluation method are listed. For benchmarks's prompt where we modified the default lmms-eval prompt, both the default and custom prompts are provided.

Modality	**Benchmark**	**Question Type**	**Example Prompt**	**# Q**	**Evaluation Method**
Audio	MMAU	MCQ	"Based on the given audio, identify the source of the speaking voice. A. Man B. Woman C. Child D. Robot Answer with the option's letter from the given choices directly."	1000	rule and based parsing and matching

continued on next page

Table A13 – continued from previous page

Modality	Benchmark	Question Type	Example Prompt	# Q	Evaluation Method
	AIR-Bench-Foundation	MCQ	"In terms of mood, what does this music suggest? A. party B. emotional C. soundscape D. trailer Answer with the option's letter from the given choices directly."	24683	rule and based parsing and matching
	MuChoMusic	MCQ	"Which instrument initiates the piece? (A) Synthesizer (B) Rueful tune (C) Acoustic guitar (D) Vocals Answer with the option's letter from the given choices directly: "	1187	rule-based parsing and matching
	Clotho-AQA (filtered binary subset)	binary QA	"Are people talking? Answer the question using a single word only. "	1442	exact match
	VocalSound	6-class classification	*We use the custom prompt for both Vicuna and LLaMA evaluation. Default: "Classify the human vocal sound to VocalSound in English: " Custom: "What type of vocal sound is in the audio? Answer with exactly one of: Laughter, Sniff, Throat clearing, Cough, Sigh, Sneeze."	3591	rule-based parsing and matching
Video	VideoMME	MCQ	*We use the custom prompt for Vicuna evaluation. Default: "Select the best answer to the following multiple-choice question based on the video and the subtitles. Respond with only the letter (A, B, C, or D) of the correct option. How many national flags appear in the video? A. 3. B. 4. C. 2. D. 5. Answer with the option's letter from the given choices directly." Custom: "Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option. How many national flags appear in the video? A. 3. B. 4. C. 2. D. 5. The best answer is:"	2700	rule-based parsing and matching

continued on next page

Table A13 – continued from previous page

Modality	Benchmark	Question Type	Example Prompt	# Q	Evaluation Method
	TempCompass (multi-choice)	MCQ	*We use the custom prompt for Vicuna evaluation. Default: "What is the man doing in the video? A. dunking a basketball B. dribbling a basketball C. passing a basketball Please directly give the best option:" Custom: "What is the man doing in the video? A. dunking a basketball B. dribbling a basketball C. passing a basketball The best answer is:"	1580	rule-based parsing and matching / calls gpt-3.5-turbo-1106 [6] for judging on rule failure
	TempCompass (caption-matching)	MCQ	*We use the custom prompt for Vicuna evaluation. Default: "Which description is a more suitable match for the video? Option 1: The man is dribbling a basketball. Option 2: A man is dunking a basketball. Please directly give the best option:" Custom: "Which description is a more suitable match for the video? Option 1: The man is dribbling a basketball. Option 2: A man is dunking a basketball. The best answer is:"	1503	rule-based parsing and matching / calls gpt-3.5-turbo-1106 for judging on rule failure
	TempCompass (yes/no)	binary QA	*We use the custom prompt for Vicuna evaluation. Default: "Is the man dunking? Please answer yes or no:" Custom: "Is the man dunking? Please answer yes or no:"	2453	rule-based parsing and matching / calls gpt-3.5-turbo-1106 for judging on rule failure
	MVBench	MCQ	*We use the custom prompt for Vicuna evaluation. Default: "Question:What is the action performed by the person in the video? Option: (A) Not sure (B) Scattering something down (C) Piling something up Only give the best option. " Custom: "Question:What is the action performed by the person in the video? Option: (A) Not sure (B) Scattering something down (C) Piling something up The best answer is:"	4000	rule and based parsing and matching

continued on next page

Table A13 – continued from previous page

Modality	Benchmark	Question Type	Example Prompt	# Q	Evaluation Method
	MLVU	MCQ	"What color is the water on the tropical beach in the video? (A) Dark blue (B) Clear (C) Green (D) Azure blue Only give the best option. Best option: ("	2174	rule-based parsing and matching * uses per-task-type accuracy, not overall accuracy
	EgoSchema (subset)	MCQ	"Taking into account all the actions performed by c, what can you deduce about the primary objective and focus within the video content? A. C is cooking. B. C is doing laundry. C. C is cleaning the kitchen. D. C is cleaning dishes. E. C is cleaning the bathroom. Answer with the option's letter from the given choices directly."	500	rule-based parsing and matching
	LongVideo-Bench (Val)	MCQ	"In front of an ancient building, a woman walks by holding something. When the subtitle says 'temples scattered throughout the', what color is the woman's clothing at that moment? A. yellow B. green C. blue D. red Answer with the option's letter from the given choices directly. "	1337	rule-based parsing and matching
Vision	GQA	Free-form short-answer	"Who is wearing the dress? Answer the question using a single word or phrase."	12578	exact match
	MMBench (Dev_EN)	MCQ	"Identify the emotion expressed in this image. There are several options: A. happiness B. sadness C. anger D. loneliness Answer with the option's letter from the given choices directly."	1164	rule-based parsing and matching / calls gpt-4o-2024-11-20 [26] for parsing if rule-based parsing fails *adapts circular evaluation
	MMStar	MCQ	"What seems to be the theme of the image? Options: A: Hanging Posters, B: Music, C: Home decoration, D: Playing guitar Answer with the option's letter from the given choices directly"	1500	rule-based parsing and matching * uses per-task-type accuracy, not overall accuracy
	POPE	binary QA	"Is there a snowboard in the image? Answer the question using a single word or phrase."	9000	exact match

continued on next page

Table A13 – continued from previous page

Modality	Benchmark	Question Type	Example Prompt	# Q	Evaluation Method
	TextVQA (Val)	Free-form short-answer	"What is the brand of this camera? Answer the question using a single word or phrase."	5000	VQA soft accuracy

F.9 Efficiency study

All efficiency numbers are measured on a single NVIDIA L40S 48GB GPU with the same numerical precision as the original checkpoints. For each method and recipient setting, we time only the core merging algorithm, excluding per-checkpoint adapter loading and saving so that the reported cost reflects merging rather than I/O. All methods use the merging hyperparameters from their respective references (300 Adam iterations for WUDI and 300 SGD iterations for OptMerge) from the source of OptMerge [75].

Tab. A14 shows that DCAT remains practical despite using calibration activations. For the LLaMA 3.1-based settings, DCAT has a merging time comparable to stronger baselines such as TSV and OptMerge: 28.45/34.57 minutes for LLaMA-Audio/Video, compared with 25.61/33.02 for TSV and 29.96/35.55 for OptMerge. Its peak memory is also comparable, requiring 16.72GB mainly to store layer-wise outputs for a small number of calibration samples, while TSV and OptMerge use 12.12GB and 18.66GB. For Vicuna, DCAT is more expensive than lightweight baselines because it merges the full weights after merging LoRA adapters, making its cost closer to the full-weight LLaMA setting. Even so, DCAT finishes within about 24–25 minutes and uses 14.56GB on a single L40S GPU, indicating that its one-time merging cost remains highly feasible.

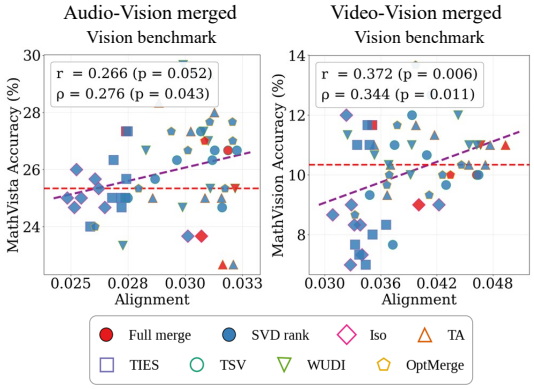

Figure A8: **Failure case on alignment–performance correlation.** Although our experiments across hundreds of model weights and diverse benchmarks demonstrate a strong correlation between the alignment metric and performance, we observe a small number of benchmarks for which this correlation is not pronounced. These cases suggest that further study may be needed to explain the underlying factors driving performance improvements.

G Limitation

First, although our analyses across hundreds of model checkpoints and diverse benchmarks reveal a strong and statistically meaningful correlation between the proposed Alignment score and downstream performance, this relationship is not universal across all evaluation settings. As shown in Fig. A8, a small number of benchmarks exhibit weak or negligible correlation, suggesting that factors beyond modality–text alignment may dominate performance in certain tasks. For example, some benchmarks may depend more heavily on modality-specific perception, reasoning format, dataset bias, or evaluation protocol than on the alignment properties captured by our score. Therefore, our Alignment score should be interpreted as a useful and broadly predictive indicator for cross-modal transfer, rather than as a complete explanation of all performance variations. Understanding when alignment is insufficient, and identifying complementary factors that better explain such exceptional cases, is an important direction for future work. Second, DCAT assumes that the models to be

Table A9: **Calibration-matched comparison.** We compare DCAT with LoRA fine-tuning, RegMean [30], and RegMean++ [27] under a matched calibration budget of 256 AVQA samples. MMAU is evaluated with Audio+Vision Vicuna-based models, and Video-MME with Video+Vision Vicuna-based models. LoRA is trained with rank 256, batch size 8, and 128 steps. The best results are in **bold**, and the second-best are underlined.

Method	MMAU				Video-MME			
	Music	Sound	Speech	Overall	Short	Medium	Long	Overall
Original	54.8	60.7	31.2	48.9	39.9	34.2	30.9	35.0
LoRA fine-tuned	53.6	62.2	33.0	49.6	44.2	35.7	31.3	37.1
RegMean [30]	58.5	66.5	37.9	54.3	45.8	38.4	34.8	39.7
RegMean++ [27]	58.8	66.5	35.5	53.6	46.9	39.3	34.6	40.3
DCAT (ours)	**60.2**	**69.1**	**41.7**	**56.2**	**48.0**	**40.5**	**36.2**	**41.6**

Table A10: **Robustness to calibration sampling.** We randomly sample 256 AVQA training examples with 10 different random seeds and apply DCAT to the Audio+Vision (MMAU) and Video+Vision (Video-MME) settings. We report per-seed results, together with the mean and standard deviation over the 10 calibration subsets.

Seed	MMAU				Video-MME			
	Music	Sound	Speech	Overall	Short	Medium	Long	Overall
0	59.38	68.57	40.04	56.00	48.64	40.53	35.53	41.57
1	59.08	68.27	39.74	55.70	48.76	39.98	35.53	41.42
2	60.88	67.67	39.74	56.10	48.09	40.31	35.76	41.39
3	59.68	66.47	40.34	55.50	47.64	41.09	35.31	41.35
4	57.89	67.67	40.34	55.30	48.76	40.98	35.20	41.64
5	57.89	68.87	41.54	56.10	47.87	41.09	35.09	41.35
6	59.68	68.27	40.34	56.10	47.53	41.09	35.09	41.24
7	59.68	67.67	40.94	56.10	47.98	40.64	35.64	41.42
8	58.19	67.97	40.34	55.50	48.20	40.20	35.20	41.20
9	60.28	67.67	40.64	56.20	48.31	40.31	35.09	41.24
Mean	**59.26**	**67.91**	**40.40**	**55.86**	**48.18**	**40.62**	**35.34**	**41.38**
Std.	**1.01**	**0.66**	**0.54**	**0.33**	**0.44**	**0.42**	**0.25**	**0.14**

merged have compatible parameter shapes, as it operates directly in weight space. This assumption is shared by many training-free merging methods, but it limits applicability to in-the-wild MLLMs whose language backbones differ in tokenizer size, embedding matrix size, LM head size, or other architectural components. Such layers are particularly important in LLMs, since they directly determine token representation and generation behavior; naïve resizing, padding, or partial copying can severely degrade model performance and would also introduce confounding factors beyond the merging method itself. For this reason, we restrict our experiments to models with shape-compatible language-model parameters and do not apply additional retraining or ad-hoc architectural surgery to force compatibility. Extending directional cross-modal alignment transfer to models with different vocabularies, embedding spaces, or heterogeneous backbone shapes would substantially broaden the coverage of DCAT and represents a promising future research direction.

H Broader Impacts

This work proposes an efficient model-merging framework that transfers alignment across MLLMs trained for different modalities, improving target-modality performance without additional large-scale training. This may have positive societal impact by reducing the data and compute costs required to develop capable multimodal systems, especially for data-scarce, high-volume, or domain-specific modalities such as audio, video, medical, or scientific data. As a result, the approach could improve accessibility for researchers and organizations with limited computational resources. However, stronger audio and video understanding capabilities may also be misused for surveillance, sensitive attribute inference, large-scale profiling, or downstream systems that make unfair or unsafe decisions. In addition, cross-modality alignment transfer may propagate or amplify biases, privacy risks, or

Table A11: **Language ability before and after DCAT.** We evaluate general language ability on 13 text benchmarks for the Vicuna-based Audio+Vision and Video+Vision models before and after applying DCAT. Δ denotes After $-$ Before. McNemar's test shows no significant change for Audio+Vision ($p = 0.506$) and a significant improvement for Video+Vision ($p \ll 0.01$).

Task	Audio+Vision			Video+Vision		
	Before	After	Δ	Before	After	Δ
MMLU	0.492	0.491	-0.001	0.501	0.510	$+0.009$
ARC	0.550	0.514	-0.036	0.506	0.521	$+0.015$
HellaSwag	0.763	0.745	-0.018	0.777	0.775	-0.002
WinoGrande	0.731	0.710	-0.021	0.722	0.722	0.000
PIQA	0.787	0.784	-0.003	0.778	0.786	$+0.008$
BoolQ	0.821	0.841	$+0.020$	0.837	0.844	$+0.007$
SciQ	0.890	0.900	$+0.010$	0.886	0.902	$+0.016$
OpenBookQA	0.446	0.454	$+0.008$	0.450	0.462	$+0.012$
CommonsenseQA	0.567	0.699	$+0.132$	0.651	0.708	$+0.057$
Social IQa	0.506	0.500	-0.006	0.479	0.491	$+0.012$
RACE	0.426	0.435	$+0.009$	0.426	0.434	$+0.008$
CoPA	0.870	0.860	-0.010	0.870	0.860	-0.010
IFEval	0.314	0.279	-0.035	0.307	0.311	$+0.004$
Mean	**0.628**	**0.632**	**+0.004**	**0.630**	**0.641**	**+0.011**

Table A12: **Hyperparameter study on DCAT.** We conduct a hyperparameter study for DCAT. We evaluate a range of options for β, η, k, and the calibration sample size, and provide the rationale behind our hyperparameter choices.

Study subject	MMAU				Video-MME			
	Music	Sound	Speech	Overall	Short	Medium	Long	Overall
Original	54.79	60.66	31.23	48.90	38.33	33.67	30.67	34.22
Base (DCAT)	60.22	69.06	41.74	56.20	46.33	39.67	34.33	40.11
$\beta = 0.1$ in Eq. 6	50.90	60.66	34.53	48.70	37.33	29.00	26.33	30.89
$\beta = 0.2$ in Eq. 6	61.08	66.97	39.64	55.10	49.33	36.67	34.67	40.22
$\beta = 0.4$ in Eq. 6	60.18	67.27	39.34	54.80	49.33	37.00	37.67	41.33
$\beta = 0.5$ in Eq. 6	58.98	66.67	39.64	54.30	49.67	37.00	38.33	41.67
$\beta = 0.6$ in Eq. 6	58.68	65.47	35.44	52.40	48.33	37.67	35.00	40.33
$\beta = 1.0$ in Eq. 6	51.50	58.56	33.63	47.90	33.00	26.33	27.33	28.89
$\eta = 10^3$ in Eq. 10	37.72	44.44	25.23	35.80	1.33	1.67	1.67	1.56
$\eta = 10^4$ in Eq. 10	56.29	66.67	37.54	53.50	46.67	34.33	34.67	38.56
$\eta = 5 \times 10^4$ in Eq. 10	58.38	66.37	39.04	54.60	49.67	38.00	36.00	41.22
$\eta = 5 \times 10^5$ in Eq. 10	57.49	66.67	38.74	54.30	50.67	37.33	34.33	40.78
$\eta = 10^6$ in Eq. 10	58.89	66.67	38.14	53.90	50.00	37.00	35.33	40.78
$\eta = 10^7$ in Eq. 10	56.89	66.67	33.33	52.30	45.67	37.33	34.67	39.22
$k = 32$	56.89	66.37	33.33	52.20	41.00	34.67	30.00	35.22
$k = 64$	58.68	66.67	39.34	54.90	49.33	37.67	35.67	40.89
$k = 256$	58.08	66.37	37.24	53.90	49.67	37.00	37.33	41.33
$k = 512$	58.68	66.97	33.63	53.10	44.33	37.67	34.67	38.89
#calib. samples $= 64$	55.99	65.17	35.44	52.20	40.33	38.00	32.00	36.78
#calib. samples $= 1024$	56.89	67.57	34.83	53.10	44.33	40.00	32.67	39.00

failure modes inherited from the donor model, and incorrect multimodal understanding could cause harm when used in high-stakes applications. To mitigate these risks, practical deployments should include careful evaluation across demographic groups and domains, privacy-preserving calibration data, safety testing before and after merging, restrictions on sensitive use cases, and human oversight in high-risk settings. Although this work is primarily foundational, its direct path to improving multimodal capability warrants consideration of these potential impacts.

Table A14: **Efficiency study of model merging methods.** We compare the computational cost of DCAT with existing merging baselines across four recipient settings: Vicuna-Audio, Vicuna-Video, LLaMA-Audio, and LLaMA-Video. We report the wall-clock merging time and peak GPU memory measured during the merging procedure. All results are measured on the same hardware and with the same precision.

Method	**Merging Time (min.)** ↓				**Peak GPU Memory (GB)** ↓			
	Vicuna-Audio	Vicuna-Video	LLaMA-Audio	LLaMA-Video	Vicuna-Audio	Vicuna-Video	LLaMA-Audio	LLaMA-Video
TA	0.14	0.45	0.74	0.93	0.00	0.00	2.94	2.94
TIES	1.18	0.51	1.81	1.64	0.00	0.00	22.51	22.51
ISO	0.58	0.53	8.96	12.03	0.04	0.04	8.09	8.09
TSV	0.68	0.69	25.61	33.02	0.05	0.05	12.12	12.12
WUDI	5.76	5.75	23.20	23.20	2.29	2.29	4.89	4.89
OptMerge	5.86	5.79	29.96	35.55	2.33	2.33	18.66	18.66
DCAT	23.64	25.39	28.45	34.57	14.56	14.56	16.72	16.72

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 identify a phenomenon in which merging MLLMs trained on different modalities using existing model-merging methods can substantially improve performance on specific target modalities. We theoretically characterize this phenomenon and leverage it to propose a method for improving the performance of modality-specific models in settings where both models and data are scarce. These contributions are clearly presented in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We've included limitations in Sec. G.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: We present Theorem 1 and provide detailed proofs of the theorem and its accompanying propositions in Appendix Sec. B. We also explicitly state the Gaussian assumption required for the theorem.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We describe the experimental settings and implementation details in the main manuscript Sec. 5.1. To ensure fair comparison, we report the hyperparameter values used for the baselines in Sec. 5.2. Additional details are provided in Appendix Sec. D.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: All datasets used for our experiments and evaluation are open-source and publicly accessible, and we will release our code upon acceptance. In addition, reviewers are provided with instructions to reproduce the main experimental results in the `README.md` file included with the supplementary code.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide detailed implementation and experimental details, including hyperparameters, the models used, and the analysis procedures, in the main manuscript Sec. 5.1, Appendix Sec. C, and Sec. D.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: We report Pearson and Spearman correlation coefficients with their corresponding p-values in Fig. 2. We also include standard deviations in Fig. 3 and Fig. 4. In Appendix Sec. F.1, we conduct statistical significance tests for performance changes across all results presented in the main manuscript and report the complete results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We report the compute resources used in our experiments in Appendix Sec. C.1. The computational costs of our method and the baselines are provided in Appendix Sec. F.9.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We provide broader impacts in Appendix Sec. H.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our experiments are conducted using existing pretrained models and do not aim to introduce new capabilities. Therefore, safeguard design is outside the scope of this work, and we rely on the safeguards already incorporated in the original models.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We appropriately cite the papers from which the models were obtained, provide citations for all baselines, and credit in-the-wild models by including direct links to their sources.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: We provide details of the provided code in supplementary `README.md`.

Guidelines:

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

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

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

Answer: [N/A]

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

Guidelines:

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

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

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

Answer: [N/A]

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

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: Our experiments include some benchmarks that use an LLM-as-a-judge and parsing.

Guidelines:

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