CHORD: Cross-modal Harmonization with Optimized Regularized Distributions for Mitigating Language Degradation

Anonymous ACL submission

Abstract

The integration of visual perception into Large Language Models (LLMs) often introduces a trade-off between cross-modal alignment and the preservation of original language capabilities. We argue that the standard cross-entropy loss used in Supervised Fine-Tuning (SFT) focuses on token-level prediction but ignores the global geometric alignment of different modality distributions. We propose Cross-modal Harmonization with Optimized Regularized Distributions (CHORD), a distribution-level constraint based on Maximum Mean Discrepancy (MMD) that regularizes the SFT process of Multimodal Large Language Models (MLLMs). We demonstrate that the MMD loss acts as a critical stabilizer that mitigates distributional drift of the language manifold during the injection of visual information. Extensive experiments show that CHORD consistently outperforms standard LLaVA-style tuning, particularly on language-heavy reasoning tasks. Specifically, CHORD achieves improvements of 4.9% and 4.1% on MMLU for LLaVA-1.5-7B and LLaVA-1.5-13B, respectively, while remaining robust on vision-language benchmarks. We further analyze the embedding distribution and find that MMD-induced alignment creates a more coherent joint-modality space, offering a scalable strategy for training more balanced MLLMs without additional data or architectural changes.

1 Introduction

The rapid evolution of Large Language Models (LLMs) has paved the way for Multimodal Large Language Models (MLLMs), which extend textual reasoning capabilities to the visual domain. The predominant paradigm, exemplified by models like LLaVA, aligns a pre-trained visual encoder with a frozen LLM via visual instruction tuning. While effective, this process often incurs a "modal alignment tax": the aggressive injection of visual signals frequently degrades the model's original linguistic capabilities, particularly in complex reasoning tasks. This phenomenon suggests a fundamental tension between accommodating new modalities and preserving the pre-trained knowledge manifold.

We argue that the standard optimization objective—token-level Cross-Entropy (CE) loss—is insufficient for resolving this tension. While CE effectively aligns individual visual tokens with their textual descriptions, it operates locally and is agnostic to the global geometric structure of the feature space. Consequently, naively minimizing prediction error can distort the semantic distribution of the LLM, causing the language manifold to drift significantly as it adapts to visual inputs. This structural degradation explains why MLLMs often suffer from catastrophic forgetting in pure-text benchmarks despite converging on visual tasks.

To address this, we introduce CHORD (Cross-modal Harmonization with Optimized Regularized Distributions), which incorporates Maximum Mean Discrepancy (MMD) into the instruction tuning process. Unlike token-level constraints, MMD operates at the distribution level, measuring and minimizing the discrepancy between visual and textual representations in a reproducing kernel Hilbert space. By adding MMD as a regularization term, we enforce a global geometric alignment that acts as a stabilizer. This encourages the model to integrate visual information into the existing language space without excessively distorting its original distributional structure, thereby maintaining the LLM's reasoning priors.

Empirically, our method demonstrates that strict distributional alignment leads to superior generalist capabilities. Without requiring additional data or complex architectural modifications, our approach significantly boosts performance on language-heavy benchmarks—achieving a 4.9% and 4.1%

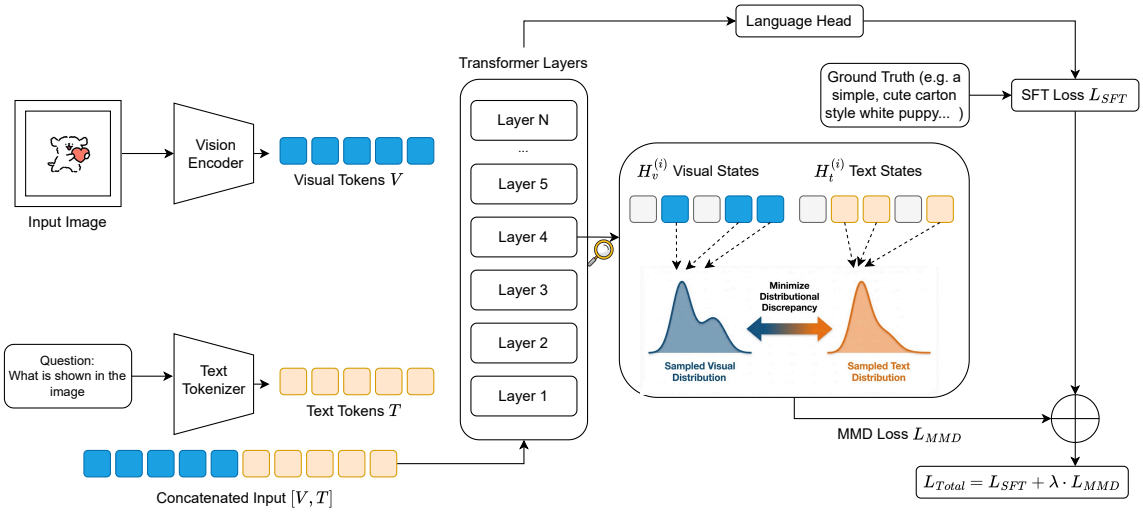


Figure 1: Illustration of our CHORD. After jointly processing concatenated visual tokens V and text tokens T through a shared Transformer backbone, CHORD explicitly minimizes the distributional discrepancy between intermediate visual and text hidden states ($H_v^{(i)}$ and $H_t^{(i)}$) at layer i using an MMD-based loss. The total loss $L_{Total} = L_{SFT} + \lambda \cdot L_{MMD}$ encourages the model to learn a unified representation space while maintaining language generation performance.

improvement on MMLU for LLaVA-1.5-7B and 13B, respectively—while maintaining performance on standard vision-language tasks. Furthermore, our analysis of the embedding space reveals that MMD regularization fosters a more coherent joint-modality space, effectively mitigating the distributional drift observed in standard fine-tuning.

Our contributions are as follows:

- We identify that standard cross-entropy loss during visual instruction tuning causes distributional drift in the language manifold, explaining the degradation of reasoning capabilities in MLLMs.
- We propose CHORD, a distribution-level regularization method using Maximum Mean Discrepancy (MMD) with a stop-gradient mechanism that preserves linguistic priors while enabling cross-modal alignment.
- We conduct extensive experiments on LLaVA-1.5 (7B/13B) and Qwen2.5-7B, demonstrating that CHORD achieves up to 4.9% improvement on MMLU while maintaining competitive vision-language performance across six benchmarks.

2 Related Work

Modality Interference in MLLMs. The integration of visual encoders with Large Language Models (LLMs) (Touvron et al., 2023a,b) has empowered models to perceive and reason across modalities (Liu et al., 2023; Radford et al., 2021). However, recent studies identify a significant trade-off known as the "alignment tax" or "modality interference," where visual instruction tuning degrades the model's inherent linguistic capabilities, particularly in complex reasoning and instruction following (Zhai et al., 2024). Recent investigations into the optimization dynamics of MLLMs further corroborate the severity of this conflict; for instance, Wu et al. (2025) demonstrate that standard gradient updates induce catastrophic forgetting of pre-trained knowledge due to conflicting optimization directions between modalities. While data-centric approaches, such as replaying high-quality text corpora during multimodal training, can mitigate this forgetting, they inevitably increase computational overhead and are often constrained by the availability of the original pre-training data. Consequently, research focus has shifted toward algorithmic regularization techniques that can preserve the language manifold without relying on extensive data replay.

Parameter and Output-level Regularization. Existing regularization strategies primarily operate at the parameter or output level. Parameter-level methods, such as Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) or LoRA-based isolation (Hu et al., 2022), attempt to restrict up-

dates to the model's weights. However, Euclidean constraints in parameter space do not guarantee semantic preservation in the functional space and may overly restrict the model's plasticity required for learning visual features. Alternatively, output-level regularization, such as Knowledge Distillation via KL divergence (Hinton et al., 2015), aligns the final probability distributions of the MLLM with its base model. While effective for surface-level alignment, KL divergence overlooks the internal reasoning process. It fails to prevent the degradation of intermediate representations, which are crucial for maintaining the deep cognitive structures required for tasks like Chain-of-Thought (CoT) reasoning.

Distributional Alignment via Feature-level Regularization. To mitigate the degradation of language capabilities in Multimodal LLMs (MLLMs), recent research has shifted focus from superficial output alignment to feature-level regularization. Notable approaches such as **VIRAL** (Yoon et al., 2025) introduce auxiliary objectives to mitigate visual feature drift, employing cosine similarity to align the LLM's internal visual representations with the original visual encoder. Similarly, **SEA** (Yin et al., 2025) proposes Supervised Embedding Alignment, which utilizes contrastive learning to force projected visual tokens to spatially overlap with semantic text prototypes, effectively "disguising" visual inputs as text embeddings.

However, these methods largely rely on rigid constraints. Whether minimizing the $\mathcal{L}_2$ distance or optimizing cosine similarity, they impose a strict *point-wise* or *direction-wise* correspondence, forcing the student model's features to mimic the teacher's exact spatial coordinates. This rigidity can severely restrict the model's plasticity, hindering the necessary feature shifts required to accommodate complex cross-modal reasoning.

In contrast to these point-wise methods, our work introduces **Maximum Mean Discrepancy (MMD)** (Gretton et al., 2012) as a flexible distributional regularization term. While MMD has been widely used in domain adaptation (Long et al., 2015), we innovatively apply it to sample visual and text tokens at specific transformer layers during SFT. Unlike VIRAL or SEA, which enforce strict feature mimicry, MMD imposes a *distribution-level* constraint via the Reproducing Kernel Hilbert Space (RKHS). This allows the model's latent representations to drift within the valid language manifold to assimilate visual semantics, provided that the overall statistical topology—captured by high-order moments—remains consistent with the pre-trained language prior.

3 CHORD

In this section, we first revisit the standard training paradigm of Multimodal Large Language Models (MLLMs). We then present our empirical observations regarding the modality gap in the latent space, which motivates our proposed approach. Finally, we introduce our method, which incorporates a Maximum Mean Discrepancy (MMD) regularization term with a stop-gradient mechanism to enhance cross-modal alignment during instruction tuning.

3.1 Preliminaries: Standard MLLM Training

Current state-of-the-art MLLMs typically adopt a two-stage training strategy to align a pre-trained visual encoder with a Large Language Model (LLM).

Stage 1: Feature Alignment. In the first stage, the goal is to align the visual features with the text embedding space. Given an input image $\mathbf{X}_v$ and its associated caption $\mathbf{X}_t$, the visual encoder (e.g., CLIP-ViT (Radford et al., 2021; Dosovitskiy et al., 2021)) extracts visual features $\mathbf{H}_v$. A learnable projector ϕ (e.g., an MLP) maps these features into the word embedding space of the LLM, denoted as $\mathbf{H}'_v = \phi(\mathbf{H}_v)$. The LLM backbone θ and the visual encoder are typically frozen, and only the projector ϕ is optimized using the auto-regressive language modeling loss.

Stage 2: Visual Instruction Tuning. In the second stage, the model is fine-tuned on visual instruction data to improve its instruction-following capabilities. The training objective is the standard Supervised Fine-Tuning (SFT) loss, maximizing the likelihood of the target response $\mathbf{Y}$ given the instruction $\mathbf{X}_{\text{instruct}}$ and the image $\mathbf{X}_v$:

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{L} \log P(y_i \mid \mathbf{X}_v, \mathbf{X}_{\text{instruct}}, y_{<i}; \theta, \phi) \tag{1}$$

where visual tokens and text tokens are treated as a unified sequence fed into the LLM. In this stage, the projector ϕ and the LLM θ (fully or via parameter-efficient tuning) are updated.

3.2 Motivation: The Persistent Modality Gap

While standard SFT successfully equips MLLMs with instruction-following abilities, we investigate

whether it achieves sufficient deep alignment between visual and textual representations in the feature space. We employ Maximum Mean Discrepancy (MMD) to measure the distributional distance between visual tokens and text tokens within the transformer layers.

We define the *Modality Alignment Ratio* (MAR, r) as the ratio of the cross-modal MMD between image and text tokens to the average intra-modal MMD:

$$r = \frac{\mathrm{MMD}(\mathrm{Image}, \mathrm{Text})}{\frac{1}{2}\left(\mathrm{MMD}_{\mathrm{text}} + \mathrm{MMD}_{\mathrm{vision}}\right)} \tag{2}$$

where $\mathrm{MMD}_{\mathrm{text}}$ and $\mathrm{MMD}_{\mathrm{vision}}$ measure the intra-modal discrepancy within text tokens and visual tokens, respectively. A larger r indicates that visual and textual representations occupy more distinct regions of the feature space relative to their own internal spread.

Table 1: Modality Alignment Ratio across different MLLMs. Higher r indicates a larger cross-modal gap relative to intra-modal compactness.

Model	$\mathrm{MMD}_{\mathrm{cross}} \downarrow$	$\mathrm{MMD}_{\mathrm{text}}$	$\mathrm{MMD}_{\mathrm{vision}}$	**MAR**$\downarrow$
LLaVA-1.5-7B	0.6285	0.0147	0.0068	58.47
Qwen2.5-VL-7B	0.3880	0.0128	0.0117	31.67

Observation. As shown in Table 1, there is a strong correlation between model capability and modality alignment. The stronger model, Qwen2.5-VL-7B, exhibits significantly lower cross-modal discrepancy ($\mathrm{MMD}_{\mathrm{cross}} = 0.39$) and a reduced Modality Alignment Ratio (MAR $= 31.67$) compared to the baseline LLaVA-1.5-7B ($\mathrm{MMD}_{\mathrm{cross}} = 0.63$, MAR $= 58.47$). This observation suggests that superior multimodal reasoning is underpinned by a tighter coupling between visual and textual manifolds. While advanced architectures may implicitly achieve better alignment, standard SFT on weaker baselines fails to close this distributional gap, leaving visual tokens distinct from the text manifold. Motivated by this, we propose to explicitly regularize the cross-modal MMD during training. By mathematically enforcing the distributional compactness characteristic of stronger models, we aim to alleviate the misalignment bottleneck and enhance the model's language preservation.

Because cross-family comparisons can conflate alignment with architecture, data, and vision encoder differences, we also examine MAR within each model family. Table 2 shows the same trend in a less confounded setting: larger variants have lower MAR but still exhibit a non-trivial cross-modal gap. This motivates an explicit regularizer even for stronger backbones, while also predicting smaller gains when the initial gap is already lower.

Table 2: Within-family MAR diagnostics measured on a separate evaluation split. Lower MAR indicates a smaller cross-modal gap relative to intra-modal compactness.

Family	**Variant**	**MAR**$\downarrow$
LLaVA-1.5	7B	132.97
LLaVA-1.5	13B	72.77
Qwen2.5-VL	7B	81.20
Qwen2.5-VL	32B	76.82

3.3 MMD Regularization with Multi-Kernel Support

To mitigate the modality gap, we introduce a regularization term based on the Maximum Mean Discrepancy (MMD). MMD is a statistical test that measures the distance between two probability distributions P and Q by mapping them into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$.

Formal Definition. Let $\mathcal{H}_k$ be an RKHS associated with a characteristic kernel $k(\cdot,\cdot)$. The squared MMD between the visual representation distribution $\mathcal{D}_v$ and the textual representation distribution $\mathcal{D}_t$ is defined as:

$$\begin{aligned}
&\mathrm{MMD}_k^2(\mathcal{D}_v, \mathcal{D}_t) \\
&\quad = \big\| \mathbb{E}_{x\sim\mathcal{D}_v}[\varphi(x)] \\
&\qquad - \mathbb{E}_{y\sim\mathcal{D}_t}[\varphi(y)] \big\|_{\mathcal{H}_k}^2 \\
&\quad = \mathbb{E}_{x,x'}[k(x,x')] \\
&\qquad + \mathbb{E}_{y,y'}[k(y,y')] \\
&\qquad - 2\,\mathbb{E}_{x,y}[k(x,y)]
\end{aligned} \tag{3}$$

where $\varphi(\cdot)$ is the feature map such that $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}_k}$.

Empirical Estimation. In practice, we compute an unbiased empirical estimate of the MMD squared using the sampled visual tokens $\tilde{\mathbf{H}}_v^{(l)} = \{v_1, \ldots, v_m\}$ and text tokens $\tilde{\mathbf{H}}_t^{(l)} = \{t_1, \ldots, t_n\}$ at layer l. Incorporating the stop-gradient operator (sg) on the text anchors as proposed, the regularized loss component is formulated as:

$$\mathcal{L}_{\text{MMD}}^{(l)} = \frac{1}{m^2}\sum_{i,j} k(v_i, v_j) + \frac{1}{n^2}\sum_{i,j} k(t'_i, t'_j) - \frac{2}{mn}\sum_{i,j} k(v_i, t'_j) \quad (4)$$

where $t' = sg(t)$ denotes the text tokens with gradients detached.

Kernel Selection. The kernel function $k(\cdot, \cdot)$ determines the geometry of the alignment objective. We support Gaussian RBF, multi-scale RBF, and cosine kernels, and use the cosine kernel in main experiments because angular similarity is often more robust in high-dimensional LLM representations (Yoon et al., 2025). Full kernel definitions are provided in Appendix C.

4 Experiments

In this section, we present a comprehensive evaluation of our proposed method, **CHORD**, across multiple multimodal and language benchmarks. We aim to answer two primary research questions: (1) Can optimizing the Maximum Mean Discrepancy (MMD) between visual and textual representations effectively improve modality alignment? (2) Does the proposed stop-gradient strategy on text tokens help in preserving the intrinsic language capabilities of the LLM during multimodal instruction tuning?

4.1 Experimental Setup

Architectures and Baselines. We conduct our experiments primarily on the LLaVA-1.5 (Liu et al., 2023) architecture, a standard and robust baseline in the MLLM community. We evaluate our method on two model scales: **LLaVA-1.5-7B** and **LLaVA-1.5-13B**, both utilizing Vicuna-v1.5 (Zheng et al., 2023) as the language backbone and CLIP-ViT-L-336px (Radford et al., 2021) as the vision encoder. To demonstrate the generalization capability of our approach, we also extend our experiments to **Qwen2.5-7B** (Yang et al., 2024) as the language backbone and SigLIP-SO400M-Patch14-384 as the vision encoder. We compare CHORD against two standard training paradigms:

- **LoRA Fine-tuning:** Standard Parameter-Efficient Fine-Tuning where only low-rank adaptation matrices are updated.
- **Full Fine-tuning (FFT):** Updating all parameters of the LLM backbone.

We additionally include a simple MSE ($\mathcal{L}_2$) regularization baseline based on *first-moment matching* (centroid alignment). Since the number of visual tokens (m) and text tokens (n) generally differs, a point-to-point $\mathcal{L}_2$ alignment is not feasible. Instead, we compute the mean (centroid) of visual tokens and text tokens at the chosen layer and minimize the squared distance between the two centroids:

$$\mathcal{L}_2 = \left\| \frac{1}{m}\sum_{i=1}^{m} v_i - \frac{1}{n}\sum_{j=1}^{n} t_j \right\|_2^2 . \quad (5)$$

To compare against data-level mitigation, we further report a text-replay baseline that mixes pure-text data into multimodal SFT, as well as CHORD combined with the same replay protocol. All baselines are trained using the official LLaVA codebase to ensure fair comparison.

Datasets and Metrics. Following the standard LLaVA-1.5 protocol, we utilize the LLaVA-v1.5-mix665k dataset for Supervised Fine-Tuning (SFT). Evaluation is performed on a diverse set of multimodal benchmarks, including POPE (Li et al., 2023) (hallucination), MMStar, MMBench (Liu et al., 2024), AI2D (Kembhavi et al., 2016) (diagram understanding), OCRBench, and SciQA (Lu et al., 2022). Crucially, to evaluate the preservation of language reasoning capabilities—a key motivation of our work—we report performance on **MMLU** (Hendrycks et al., 2021) (Massive Multitask Language Understanding), and further evaluate ARC-Easy and HellaSwag in the expanded analysis. All evaluation experiments are conducted using lmms-eval (Zhang et al., 2025).

4.2 Implementation Details

Our method introduces a regularization term to the standard SFT objective. During training, we extract hidden states from specified layers of the LLM. We apply a sampling strategy with a fixed *sampling_ratio* to select a subset of visual tokens and text tokens. We then calculate the Maximum Mean Discrepancy (MMD) between these two sets of tokens. Formally, the total loss is defined as $\mathcal{L}_{total} = \mathcal{L}_{SFT} + \lambda\mathcal{L}_{MMD}$. To prevent the language representation from drifting towards the visual modality, we strictly apply a **stop-gradient** operation to the text tokens within the MMD calculation. This ensures that the optimization pressure is directed solely towards aligning visual features

Table 3: **Main Results on Multi-modal and Language Benchmarks.** We compare our proposed method (**CHORD**) against standard LoRA fine-tuning and other regularization baselines. Our method achieves significant gains on the language-heavy benchmark (MMLU) while maintaining competitive performance on vision-language tasks. Metrics are reported in percentage (0-100).

Language Backbone	Method	POPE	MMStar	MMBench	AI2D	OCRBench	SciQA	Avg (VL)	MMLU
				LLaVA-v1.5-7B					
LLaVA-1.5-7B	LoRA	87.2	35.2	64.9	55.5	**32.3**	70.3	57.6	33.4
	$\mathcal{L}_2$ (LoRA)	87.5	33.9	64.2	55.7	32.2	71.3	57.4	37.0
	Chord (LoRA)	**87.6**	**35.2**	**66.2**	**56.6**	31.8	**70.9**	**58.0**	**38.3**
	FFT	86.8	**35.3**	**66.7**	**55.8**	32.0	**70.5**	**57.8**	40.6
	$\mathcal{L}_2$ (FFT)	86.8	35.6	65.8	52.2	32.2	69.6	57.0	41.9
	Chord (FFT)	**86.9**	35.0	65.3	55.4	**32.7**	70.0	57.6	**44.7**
				LLaVA-v1.5-13B					
LLaVA-1.5-13B	LoRA	87.3	37.8	**68.4**	58.2	34.1	**73.9**	59.9	41.5
	Chord (LoRA)	**87.8**	**38.0**	67.5	**59.1**	**34.2**	73.4	**60.0**	**45.6**
				Qwen2.5-7B					
Qwen2.5-7B	LoRA	87.5	50.3	**76.8**	**73.7**	47.0	**82.8**	69.7	65.1
	Chord (LoRA)	**87.8**	**50.6**	75.8	72.7	**49.4**	82.5	**69.8**	**66.7**

to the language manifold, not vice versa. Hyperparameters such as the sampling ratio and the coefficient λ are determined via grid search on a held-out validation set.

4.3 Main Results

The main comparative results are summarized in Table 3. We observe consistent improvements driven by CHORD across different settings.

Mitigating Language Forgetting. The most significant observation is the performance on the pure language benchmark, MMLU. Standard multimodal SFT (both LoRA and FFT) often suffers from catastrophic forgetting, degrading the LLM's original reasoning abilities.

- On **LLaVA-1.5-7B** (LoRA), the standard baseline reaches 33.4% MMLU, while a simple $\mathcal{L}_2$ regularization baseline improves it to 37.0%. CHORD further boosts MMLU to **38.3%**, suggesting that matching the full distribution is more effective than only matching the mean.
- In the **Full Fine-Tuning** setting on LLaVA-1.5-7B, CHORD achieves **44.7%** on MMLU, outperforming both the standard FFT baseline (40.6%) and the $\mathcal{L}_2$-regularized FFT baseline (41.9%).
- Similar trends are observed on the 13B model and Qwen2.5-7B, where CHORD consistently improves MMLU while maintaining competitive multimodal performance.

Table 4: **Comparison with data replay on LLaVA-1.5-7B LoRA.** CHORD outperforms first-moment regularization and text replay, and remains complementary to replay when both are combined.

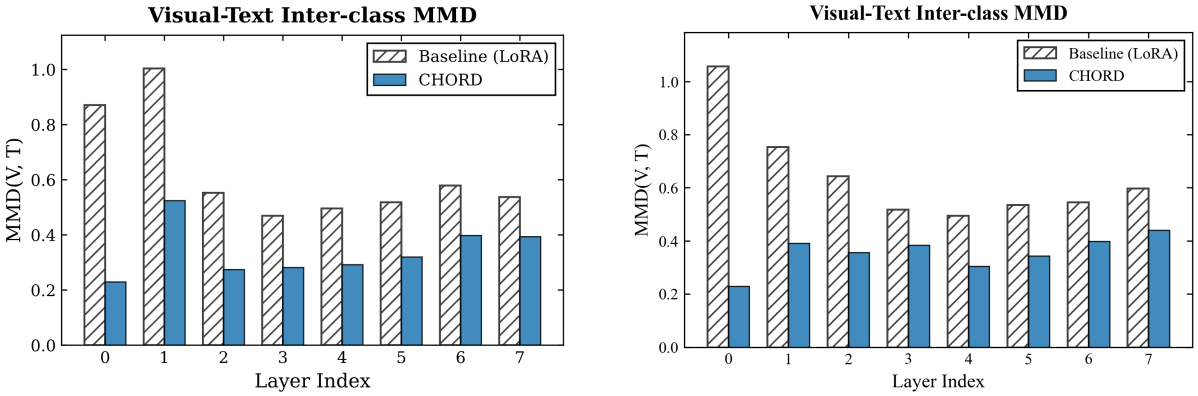
Method	Avg VL	MMLU	Extra Data
Vanilla LoRA	57.6	33.4	No
$\mathcal{L}_2$ regularization	57.4	37.0	No
Text replay	57.9	37.8	Yes
CHORD	**58.0**	38.3	No
CHORD + replay	**58.0**	**38.9**	Yes

These results strongly support the efficacy of CHORD, which effectively preserves the text representation space while forcing visual tokens to align with it.

Table 4 shows that CHORD is not merely a proxy for data replay. It improves both Avg VL and MMLU without additional data, and combining it with replay further improves MMLU from 37.8 to 38.9. This suggests that intermediate distributional regularization captures a failure mode that is orthogonal to data composition.

To test whether language preservation is specific to MMLU, we further evaluate ARC-Easy and HellaSwag. Table 5 shows that CHORD improves both tasks at 7B and 13B scales, indicating that the language gains extend to science and commonsense evaluations.

Together, the replay and additional-language results strengthen the interpretation that CHORD improves a general preservation property rather than overfitting to one evaluation suite. Replay



(a) LLaVA-v1.5-7B

(b) LLaVA-v1.5-13B

Figure 2: Layer-wise Visual-Text Inter-class MMD. We compare the Maximum Mean Discrepancy (MMD) between visual and textual token distributions across the first 8 layers of LLaVA-v1.5 models. The baseline (hatched bars) exhibits high discrepancy, particularly peaking at early layers. In contrast, CHORD (solid blue bars) consistently reduces MMD across both the 7B and 13B models, indicating superior distributional alignment regardless of model scale.

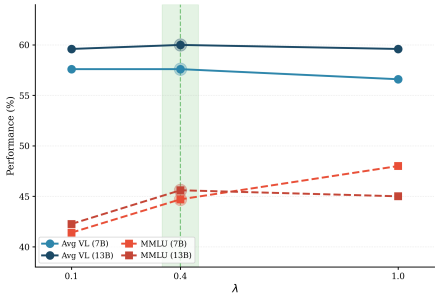


Figure 3: λ sensitivity. CHORD uses $\lambda = 0.4$ as the default configuration.

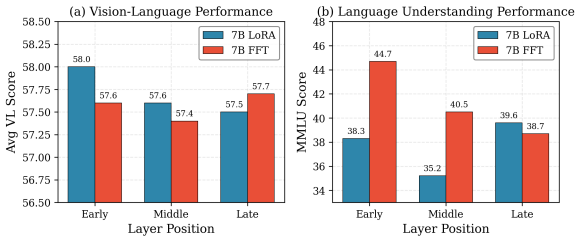


Figure 4: Target-layer ablation. CHORD uses early-layer regularization as the default configuration.

changes the training data distribution, whereas CHORD changes the geometry of intermediate representations; their complementarity suggests that distribution-level alignment and data-level rehearsal address different sources of forgetting.

Competitive Vision-Language Performance. While significantly boosting language capabilities, CHORD maintains or improves performance on vision-language tasks. On LLaVA-1.5-7B (LoRA), our method achieves the best results on MMStar (35.2), MMBench (66.2), and AI2D (56.6), leading to the highest average VL score of **58.0**. In the FFT setting, although there are minor fluctuations in specific metrics like MMBench, the overall average VL performance remains highly competitive (57.6 vs. 57.8) while delivering substantially better language understanding. The results on Qwen2.5-7B further validate robustness, where CHORD achieves the highest average VL score (69.8) alongside the best MMLU performance (66.7).

In summary, CHORD successfully regularizes the training process, preventing overfitting to visual instruction data at the expense of general language logic, yielding a more balanced and robust multimodal learner.

Table 5: **Additional text-only language benchmarks.** CHORD improves beyond MMLU on ARC-Easy and HellaSwag.

Model / Setup	Benchmark	Baseline	CHORD	Δ
LLaVA-1.5-7B LoRA	ARC-Easy	58.33	59.21	+0.88
LLaVA-1.5-7B LoRA	HellaSwag	66.22	67.83	+1.61
LLaVA-1.5-13B LoRA	ARC-Easy	66.13	67.19	+1.06
LLaVA-1.5-13B LoRA	HellaSwag	74.38	75.62	+1.24

4.4 Discussion

Layer-wise center similarity analysis To understand why standard SFT degrades language capabilities, we analyze the geometric relationship between visual and textual representations. On MMBench-Dev, CHORD raises the center cosine similarity between visual and textual token representations above 0.97 from the earliest layers, whereas the LoRA baseline starts around 0.75 and

only gradually approaches the text manifold. This centroid-level diagnostic is complementary to the inter-class MMD analysis: center similarity captures first-order geometric drift, while MMD measures distributional mismatch beyond the mean. Their consistent trend suggests that CHORD does not merely optimize a single statistic, but broadly anchors visual representations to the language manifold; the full layer-wise plot is provided in Appendix A.

Layer-wise Visual-Text Inter-class MMD On the dev split of MMBench_en (Figure 2), the baseline exhibits a sharp MMD spike at Layer 1 (> 1.0), indicating that standard SFT struggles to project visual tokens onto the text manifold immediately after the initial projection. CHORD reduces this peak by roughly 50% and keeps the value below 0.4 across the remaining profiled layers, confirming that the regularizer closes the modality gap at the distributional level.

4.5 Ablation Summary

We ablate the key CHORD design choices and use $\lambda = 0.4$, the cosine kernel, stop-gradient on text tokens, and early-layer regularization (layers 1–8) as the default setting. These choices provide the best trade-off between language preservation and vision-language performance: cosine-kernel MMD reaches 44.7 MMLU without reducing Avg VL, while removing stop-gradient raises MMLU in isolation but drops Avg VL from 57.6 to 56.7. Figures 3 and 4 further show that increasing λ from 0.1 to 0.4 improves language capability, but $\lambda = 1.0$ begins to trade off multimodal performance; they also confirm that early-layer regularization outperforms middle- or late-layer targets. Full ablation details are provided in Appendix B.

These choices align with the structure of the problem: stop-gradient prevents the text manifold from drifting toward visual features, early-layer regularization targets where the modality gap is largest before substantial cross-modal mixing occurs, and the cosine kernel matches the anisotropic geometry of LLM hidden states, where directional similarity better reflects semantic compatibility than vector magnitude.

5 Conclusion

In this paper, we identify a critical limitation of standard visual instruction tuning for Multimodal Large Language Models (MLLMs): optimizing token-level cross-entropy can distort the pre-trained language manifold, degrading reasoning ability. To address this, we propose **CHORD** (Cross-modal Harmonization with Optimized Regularized Distributions), a simple regularization framework that enforces a distribution-level constraint in the latent feature space.

CHORD minimizes the Maximum Mean Discrepancy (MMD) between sampled visual and text representations, while applying stop-gradient on text tokens to anchor visual features to the robust pre-trained language geometry. In this way, the model absorbs visual information while mitigating drift in its linguistic structure.

Extensive experiments on LLaVA-1.5 (7B/13B) and Qwen2.5-7B show that CHORD mitigates catastrophic forgetting, yielding up to a 4.9% MMLU gain over standard baselines while maintaining or improving vision-language performance. Layer-wise analyses further confirm that CHORD reduces the modality gap in early layers, producing a more cohesive joint representation space.

Overall, CHORD is a simple plug-and-play regularizer for the alignment–stability dilemma in MLLMs. Future work will extend CHORD to larger-scale training with more data and stronger backbones.

Table 6: **Kernel and stop-gradient ablation.** Avg VL denotes the average score across multimodal benchmarks, and MMLU measures language capability.

Method	Avg VL	MMLU
Baseline (FFT)	57.6	40.6
FFT with MMD Regularization		
+ Multi-scale Kernel	**57.6**	41.4
+ Gaussian RBF Kernel	57.3	42.5
+ Cosine Kernel	**57.6**	**44.7**
+ Cosine Kernel (w/ text gradient)	56.7	47.7

Limitations

Our method requires caching intermediate hidden states from selected transformer layers to compute the MMD regularizer. In our LLaVA-1.5-7B setting, peak GPU memory increases from 39 GB to 46 GB, an 18% overhead (+7 GB), while the MMD computation itself has negligible throughput impact under token sampling. This limitation can be alleviated by reducing the number of regularized layers or lowering the visual/text token sampling ratio.

Ethical Considerations

This paper presents CHORD, a method for improving the training of multimodal large language models by preserving language capabilities during visual instruction tuning. Our work contributes to making MLLMs more balanced and capable across both visual and linguistic tasks.

Potential Benefits. By mitigating catastrophic forgetting of language abilities, our method may lead to more reliable MLLMs that maintain strong reasoning capabilities. This could improve applications in education, accessibility (e.g., image description for visually impaired users), and scientific research where both visual understanding and complex reasoning are required.

Environmental Considerations. Our method adds moderate memory overhead during training but does not require additional parameters, architecture changes, or replay data in its default form, making it a relatively lightweight alternative to data-level mitigation approaches.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251. Springer.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, and 1 others. 2025. Mitigating visual knowledge forgetting in mllm instruction-tuning

via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yuanyang Yin, Yaqi Zhao, Yajie Zhang, Yuanxing Zhang, Ke Lin, Jiahao Wang, Xin Tao, Pengfei Wan, Wentao Zhang, and Feng Zhao. 2025. Sea: Supervised embedding alignment for token-level visual-textual integration in mllms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1058–1070.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, and 1 others. 2025. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2025. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Additional Representation Analysis

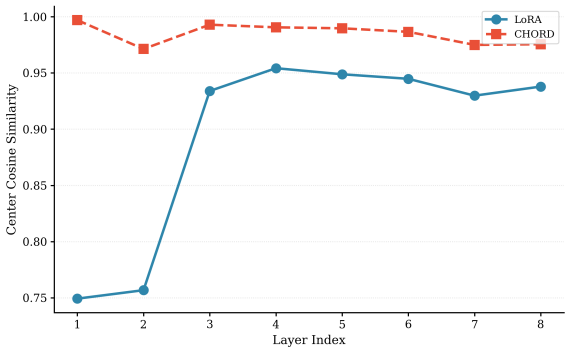


Figure 5: Layer-wise Center Cosine Similarity. We visualize the cosine similarity between the centroids of visual and textual token representations across the first 8 transformer layers on MMBench-Dev. The baseline (LoRA) exhibits a significant modality gap in early layers, whereas our method (CHORD) maintains consistent high-level alignment.

B Ablation Details

B.1 Ablation on λ

We investigate the impact of the MMD regularization weight λ on model performance, varying $\lambda \in \{0.1, 0.4, 1.0\}$. As illustrated in Figure 3, increasing λ from 0.1 to 0.4 yields a consistent improvement in language capabilities (MMLU) for both 7B and 13B models. However, a further increase to $\lambda = 1.0$ leads to diminishing returns or slight regressions in language tasks while also reducing multimodal performance (Avg VL). We therefore adopt $\lambda = 0.4$ as the default setting.

B.2 Impact of Kernel Choice and Stop Gradient

We compare Gaussian RBF, multi-scale, and cosine kernels under the full fine-tuning setting on LLaVA-v1.5-7B. As shown in Table 6, all MMD variants improve language capability over the FFT baseline. The cosine kernel achieves the best trade-off, yielding 44.7 MMLU without reducing Avg VL. Allowing gradients to propagate through text tokens raises MMLU to 47.7 but reduces Avg VL to 56.7, supporting our default use of stop-gradient as a fixed text-manifold anchor.

B.3 Ablation on Target Layers

We study how the choice of transformer layers for computing $\mathcal{L}_{\text{MMD}}$ affects the trade-off between language preservation and vision-language performance. On LLaVA-v1.5-7B, we ablate early layers (1–8), middle layers (9–16), and late layers (16–24), keeping all other hyperparameters fixed. Figure 4 shows that applying MMD on early layers achieves the best overall performance, suggesting that early supervision anchors visual tokens to the pre-trained text manifold before substantial cross-modal mixing occurs.

C Kernel Definitions

The choice of the kernel function $k(\cdot, \cdot)$ determines the feature geometry in which distributional alignment is measured. We consider three kernel variants.

Gaussian RBF Kernel. The standard Radial Basis Function (RBF) kernel captures local structures based on Euclidean distance:

$$k_{\text{RBF}}(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \tag{6}$$

where σ is the bandwidth parameter.

Multi-Scale Kernel. Single-scale Gaussian kernels are sensitive to the choice of bandwidth. We therefore also use a multi-scale kernel, defined as a linear combination of RBF kernels with varying bandwidths σ_u:

$$k_{\text{MK}}(x, y) = \sum_{u=1}^{U} \beta_u \exp\left(-\frac{\|x - y\|^2}{2\sigma_u^2}\right). \tag{7}$$

This captures both fine-grained and coarse-grained discrepancies between modalities.

Cosine Kernel. Because LLM latent spaces are often anisotropic and semantic similarity is better captured by angular distance than Euclidean magnitude, we also use a cosine similarity-based kernel:

$$k_{\text{Cos}}(x, y) = \exp\left(\alpha \cdot \frac{x^\top y}{\|x\|\|y\|}\right), \tag{8}$$

where α is a scaling factor. This kernel aligns directional components while ignoring vector norms.