

Dive into the implicit biases of low-rank vision-language alignment

Mingjia Shi¹ Shuo Wang^{1†} Xiaobo Wang^{2‡} Sifan Zhou¹ Kai Wang³
Tianyu Fu¹ Chenxu Zhao^{1†} Anyang Su¹ Ping Jiang¹ Minghui Wu¹

¹ Mininglamp

² Shenzhen University of Advanced Technology

³ National University of Singapore

{wangshuo.e,zhaochenxu}@mininglamp.com, 3101ihs@gmail.com,
wangxiaobo@suat-sz.edu.cn

Abstract. Vision-language alignment—the stage that bridges pretrained vision encoders and large language models—is widely treated as a form of pretraining requiring full-parameter updates. We challenge this view and investigate what happens when low-rank adaptation is applied to the LLM during this stage instead. We find that low-rank alignment not only reduces computational costs but also outperforms full-parameter alignment on most benchmarks. To understand this phenomenon, we systematically characterize the implicit biases introduced by low-rank adaptation during alignment. Empirically, we find that low-rank alignment shifts model behavior from hallucinatory to conservative and preserves per-token linear separability of visual features that full-parameter alignment disrupts—a phenomenon we term *LS-curse*. Geometrically, low-rank aligned models exhibit more homogeneous and structurally stable visual representations, maintaining modality-specific knowledge rather than prematurely fusing entity-level semantics. Theoretically, we establish two theorems showing that low-rank alignment induces preferences for parameter subspaces with flat gradients and feature subspaces robust to perturbations, providing a principled explanation for the observed structure-preserving behavior. Extensive experiments cover ablation over 100 alignment configurations, three families of low-rank operators, and various rank, encoder, and other settings.

1 Introduction

The construction of vision-language models (VLMs) hinges on a critical stage: *vision-language alignment*, where pretrained vision encoders are bridged with large language models (LLMs) through adapter modules and joint training [5, 8, 25]. As illustrated in Figure 1, this stage precedes instruction tuning and establishes the cross-modal feature correspondence upon which all downstream capabilities depend. Standard practice updates the full LLM parameters during alignment, treating it as a form of pretraining. However, this view constitutes a

³ [†] Corresponding author. [‡] Project lead.

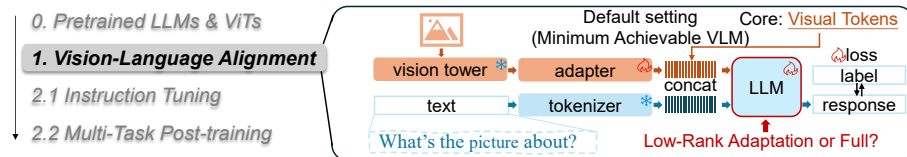


Fig. 1: Vision-language alignment framework (default): a minimum achievable VLM with a cross-modal adapter (MLP). How does *low-rank alignment* shape *visual tokens*?

misconception: unlike LLM pretraining, which builds linguistic representations from scratch, vision-language alignment operates atop pretrained knowledge and reasoning priors—it is, in fact, *supervised fine-tuning*, precisely the regime for which low-rank adaptation methods were designed [10, 12, 14, 55]. Yet, applying low-rank methods during this stage remains largely unexplored.

When low-rank adaptation is applied to the LLM during alignment (*i.e.*, *low-rank alignment*), we observe that not only does the training cost drop substantially, but downstream performance *empirically improves* across all involved model scales (from 1.4B to 14B). This improvement holds across diverse benchmarks spanning perception, knowledge, reasoning, and hallucination, and generalizes across three families of low-rank operators (LoRA, LoHa, LoKr). This finding raises a natural question: **how do the low-rank methods reshape the visual features to yield superior representations, and technically, what *implicit biases* does the *low-rank alignment* introduce?**

Empirical observations. We investigate this question first at the behavior and feature levels. At the behavior level, hypothesis testing on visual perception benchmarks reveals that low-rank alignment shifts model decision-making from hallucinatory to conservative, reducing overconfident predictions while preserving cautious uncertainty (§2.2). At the feature level, token-wise linear probes uncover that full-parameter alignment disrupts the linear separability of certain visual tokens—a phenomenon we term *LS-curse*—whereas low-rank alignment preserves the per-token knowledge structure inherited from the pretrained vision tower (§2.3). Together, these observations suggest that the performance gain stems from low-rank alignment’s stronger preservation of modality-specific knowledge structures, rather than the premature entity-level mixing induced by full-parameter updates.

Geometric characterization. We further characterize these biases through geometric analysis of the visual feature space. Fine-grained and coarse-grained linear probes reveal that fully aligned models exhibit greater feature diversity and broader angular coverage across tokens, reflecting aggressive entity-level fusion. In contrast, low-rank aligned models maintain homogeneous and structurally conservative representations. Manifold visualizations corroborate this finding: per-token feature distributions under low-rank alignment are markedly more consistent than those under full alignment (§2.4).

Theoretical analysis. We establish two theorems characterizing the mathematical mechanisms underlying these biases. Theorem 1 reveals that the low-rank gradient flow exhibits a noise-weighted smoothing effect, preferentially reinforcing updates along feature directions robust to stochastic variations—encoding

an implicit bias toward flatter regions of the loss landscape. Theorem 2 demonstrates that the steady-state distribution of low-rank parameters concentrates on subspaces with flat gradients and noise-robust features. These two results jointly explain why low-rank alignment preserves general knowledge structures instead of prematurely fusing entity-specific characteristics (§3).

We conduct extensive experiments across over 55 vision-language alignment settings—unique combinations of model, vision encoder, low-rank operator, rank, vision-tower depth, and learning-rate schedule—spanning model architectures from MobileLLaMA-1.4B to Qwen3-14B and three families of low-rank operators, which together yield over 100 configurations when evaluated across benchmark categories. We summarize our contributions as follows:

- We demonstrate that low-rank adaptation during vision-language alignment not only improves efficiency but also consistently enhances downstream performance across all evaluated model scales (1.4B–14B), challenging the prevailing assumption that alignment requires full-parameter updates.
- We provide a systematic empirical and geometric characterization of the implicit biases introduced by low-rank alignment, identifying the LS-curse in full alignment and revealing how low-rank methods preserve modality-specific knowledge structures—collectively supporting a *progressive fusion* view in which general visual representations are retained during alignment and entity-specific fusion is deferred to instruction tuning.
- We establish two theorems identifying the theoretical mechanisms: noise-robust feature preferences during optimization (Theorem 1), and flat-gradient subspace concentration at steady state (Theorem 2)—providing a principled explanation for the structure-preserving behavior observed empirically.
- We conduct comprehensive ablations covering rank, operators, model scales, architectures, encoder generalization (CLIP, DINOv2, SigLIPv2) and other details, providing practical guidance for efficient multimodal training.

2 Empirical Study: Low-Rank Alignment’s Preferences

Training protocol. As illustrated in Figure 1, during the vision-language alignment stage, we employ perceptual question answering (QA) tasks with general prompting templates. Trainable parameters comprise the LLM backbone and the vision-tower adapter; the vision encoder is frozen by default. For low-rank alignment, we apply adaptation exclusively to all linear modules in the LLM, since the adapters are already lightweight—imposing low-rank constraints on them would force all vision features into a

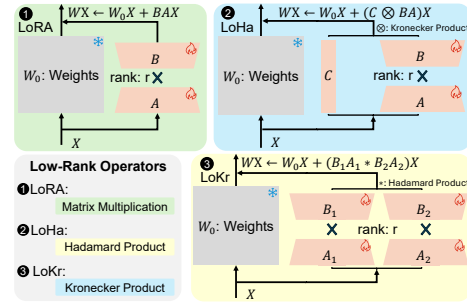


Fig. 2: Operators for low-rank alignment: LoRA/LoHa/LoKr (Matrix, Hadamard and Kronecker Products).

low-rank subspace. Both full-parameter and low-rank baselines are tuned under matched hyperparameter budgets to ensure fair comparison (best out of 9 times). For behavioral evaluation, instruction tuning is necessary to ensure proper instruction following. To assess zero-shot capabilities, we divide instruction tuning into two phases: initial tuning with a base visual question answering (VQA) instruction dataset, followed by post-training with an extended instruction dataset covering task-specific and knowledge-intensive data. All benchmarks except GQA and TextVQA are treated as unseen tasks. Extended ablations on encoder choice, frozen *vs.* unfrozen settings, low-rank placement, learning rate schedules, and overfitting controls are provided in §4.

Low-rank operators. We focus on linear transformations, specifically LoRA [14], LoHa [17], and LoKr [51], which encompass the three principal product-based constructions for low-rank modules: matrix multiplication, Hadamard product, and Kronecker product. The formulations and operator details are in Figure 2.

Linear probe. The procedure extracts features from a classification dataset and trains a linear classifier on them to measure accuracy. This approach enables us to assess both the robustness of the corresponding knowledge and the degree of feature entanglement.

Resources. We conduct experiments across a range of language models spanning compact to commonly used scales, including MobileLLaMA-1.4/2.7B [8], Vicuna-7/13B [7], and Qwen3-8/14B⁴ [49]. The vision tower is CLIP-Vision [38]. For vision-language alignment, we utilize datasets including LAION [42], CC [43], SBU [37], SAM [39], and MS-COCO [24]. Instruction tuning leverages default instruction datasets, including LLaVA [25], ShareGPT4V-PT [5], GQA [15], MS-COCO-train-2017, OCRVQA [34], TextVQA [44], and VG [20], along with extended instruction datasets: SBU, ScienceQA [30], MS-COCO, GQA, IconQA [31], SAM, ShareGPT4V-TextVQA [5], Web-Celebrity [29], Web-Landmark [8], and WikiArt [16]. Evaluation is performed on a diverse set of benchmarks and tasks, including GQA, MMBench [28], MME [52], MMMU-Pro-Vision [54], MMVet-gpt [53], POPE [22], ScienceQA, and TextVQA.

Default settings. Since this work focuses on the alignment stage, we minimize the VLM pipeline to its essential components, deliberately removing unnecessary architectural tricks to facilitate clean ablation and isolate alignment-stage effects. Unless otherwise specified, instruction fine-tuning uses GQA, TextVQA, and OCRVQA as task-specific datasets, and LoRA with rank 128 by default. For clarity in reporting, we categorize the main benchmarks into four groups: *perception*, *knowledge*, *reasoning*, and *hallucination*, corresponding to MME-perception, GQA, MME-reasoning, and POPE, respectively. For simplicity, we refer to models by their size rather than their full names. Default visualizations and fine-grained analyses are performed using the 1.4B model to reduce the

⁴ Detailed experiments regarding Qwen3 and other settings are in the Appendix.

Table 1: Performance of low-rank vs. full-parameter alignment across model scales (1.4B–13B). Low-rank results report the best configuration across LoRA, LoHa, and LoKr after three rounds of rank tuning. Metrics cover perception (MME-P), knowledge (GQA), reasoning (MME-R), and hallucination (POPE). Training time on 8×A100.

size	perception $\uparrow$	$\Delta \uparrow$	knwldg. $\uparrow$	$\Delta \uparrow$	reason $\uparrow$	$\Delta \uparrow$	hallu. $\uparrow$	$\Delta \uparrow$	time (h)	$\downarrow -\Delta \uparrow$
1.4B	208.1[453.5]		245.418.9[28.0]		9.122.9[766.6]		743.7 2.5[72.8]		70.3 4.53[2.52]	2.01
2.7B	15.7[669.0]		653.311.6[42.4]		30.840.4[201.4]		161.0 0.5[76.4]		75.9 8.61[4.29]	4.32
7B	521.0[965.3]		444.323.4[38.5]		15.186.8[332.5]		245.735.0[80.4]		45.4 19.75[7.68]	12.07
13B	236.6[1079.7]		843.129.5[43.0]		13.552.5[219.6]		167.166.9[82.9]		16.035.32[13.23]	21.09

computational and storage overhead associated with complex processing (*e.g.*, feature extraction and dimensionality reduction).

2.1 Pre-Experiments: Performance of Low-Rank Alignment

We evaluate models aligned via low-rank methods under the default instruction-tuning setting described above. Table 1 demonstrates that low-rank vision-language alignment substantially enhances final model performance across all scales, even when implemented solely through low-rank branches prior to instruction tuning. It raises the question: *why and how does low-rank alignment boost performance?* We then address this question through both experimental and theoretical analyzes of the mechanism of low-rank alignment.

2.2 Behavioral Preferences: From Hallucinatory to Conservative

We first examine whether low-rank alignment alters model decision-making behavior. To this end, we evaluate aligned models on a yes-or-no visual QA dataset and measure Type I and Type II error rates after instruction tuning. In this context, a Type I error occurs when the model incorrectly answers “yes” to a negative-ground-truth question, indicating an overconfident or hallucinated response. A Type II error occurs when the model incorrectly answers “no” to a positive-ground-truth question, reflecting cautious behavior that defaults to rejection under uncertainty. At the sentence level, such cautiousness manifests as increased perplexity in the generated answers.

Figure 3 demonstrates that low-rank adaptation shifts model behavior from hallucinatory to conservative. At the 1.4B and 7B scales, LoHa substantially reduces the Type I error rate compared to full fine-tuning, effectively reversing the model’s behavioral tendency from overconfident affirmation toward cautious

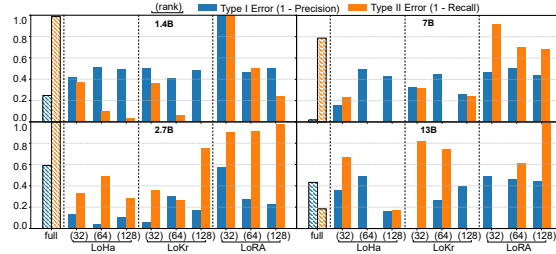


Fig. 3: Type -I and -II error rate: models and low-rank operators. The results are on POPE, a visual perception hallucination benchmark.

rejection. Through grid search over rank values, we find that certain low-rank operators, such as LoHa, can simultaneously reduce both Type I and Type II error rates without inducing a mere trade-off between the two failure modes. We use *conservative* to denote a reduced Type I (hallucination) rate and *decisive* to denote that this reduction is not obtained at the cost of a higher Type II rate, *i.e.*, both error modes decrease jointly rather than trading off. We note that this shift is context-dependent: at the smallest (1.4B) scale the gap narrows, since an under-capacity model already tends toward conservative default decisions, and conservatism is not universally beneficial—settings that require affirmative identification may favor less cautious behavior.

Obs.1 Low-rank alignment is associated with conservative/decisive behaviors.

2.3 Feature Preferences: Linear Separability Preservation

We next investigate how the visual features—*i.e.*, the LLM’s visual inputs—are shaped by alignment before multimodal fusion takes place. We employ ImageNet [41] as the test dataset to evaluate general vision capabilities.

We measure *per-token linear separability* (LS), which assesses whether each visual token independently retains a linearly separable knowledge structure. Concretely, for each token position t in the visual sequence, we train a separate linear classifier on the corresponding C -dimensional feature vectors across all samples and report classification accuracy. A high per-token LS indicates that the knowledge encoded at that position remains well-structured and discriminative; a low value suggests that the token’s representation has been entangled with information from other tokens or modalities.

The original vision tower is well-trained and exhibits near-perfect token-wise linear separability. Figure 4 reveals that in a low-level perception setting (*i.e.*, a 1.4B model with full or low-rank alignment using a 2-layer vision-language connector), the vision tokens of aligned models before instruction tuning no longer maintain full linear separability on ImageNet. For the low-rank aligned model (LoRA with rank 128), individual tokens remain largely linearly separable, whereas full-parameter alignment disrupts this property in a subset of tokens—a phenomenon we term *LS-curse*: certain tokens lose their discriminative structure entirely, with accuracy dropping from near-perfect to as low as 83.6%. As shown in the following table, while some fully aligned tokens retain the knowledge structure from generic image datasets, others exhibit a mixing of representations that destroys linear separability.

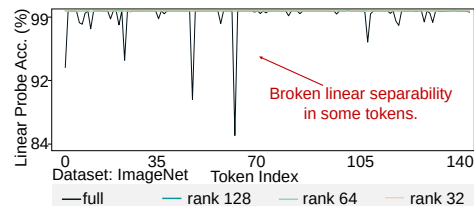


Fig. 4: Broken per-token linear separability of visual tokens, marked in Figure 1, tested by the default settings on ImageNet.

	origin	full	low-rank
avg. acc. $\pm$ std. (%) [min]	100 $\pm$ 0.00	99.52 $\pm$ 1.89 [83.61]	99.99 $\pm$ 0.02

This mixing of partial representations is not necessarily detrimental; it can introduce beneficial diversity across token representations. For instance, adversarial samples with similar representations may align semantically, thus appearing non-adversarial in certain tasks or contexts. However, it reflects a fundamentally different alignment strategy: full-parameter alignment aggressively restructures the visual feature space, whereas low-rank alignment operates conservatively within a constrained subspace.

Obs.2 Low-rank alignment preserves per-token linear separability.

Interpretation. The combination of Obs. 1 and 2 reveals a coherent picture of the implicit biases introduced by low-rank alignment. At the behavior level, low-rank aligned models exhibit greater conservatism and reduced susceptibility to hallucination. At the feature level, this conservatism is grounded in the preservation of modality-specific knowledge structures: low-rank alignment maintains the per-token linear separability inherited from the pretrained vision tower, whereas full alignment disrupts this structure through aggressive entity-level mixing. The subspace-constrained nature of low-rank updates limits token-wise diversity, preserving commonalities within each modality and preventing premature fusion of entity-level characteristics, while adjusting the subspace that most strongly affects the majority of features to maximize alignment loss reduction.

	full	low-rank
(behavior) decisiveness	hallucinatory	conservative
(feature) preference	entity characteristics	modal commonality

2.4 Geometric Characterization

To further characterize the structural differences between full and low-rank alignment, we conduct two token-level geometric analyses: (1) testing each token’s linear separability on diverse, unseen fine-grained classification datasets (*e.g.*, birds and dogs) to disentangle general from entity-specific knowledge; and (2) examining the angular coverage and manifold structure of token features on general coarse-grained datasets (*e.g.*, ImageNet).

General vs. specific knowledge. If low-rank alignment preserves general knowledge structures without prematurely fusing entity-specific characteristics, the two alignment types should yield comparable per-token LS on fine-grained entity categories while differing substantially on coarse-grained, general-purpose datasets. Table 2 confirms this prediction. On both STF-dogs [19] and CUB [47]—fine-grained datasets targeting dogs and birds, respectively—the fully aligned model exhibits high peak accuracy (*e.g.*, 95.7% on STF-dogs, 94.0% on CUB) but also pronounced inter-token variability, with the worst tokens degrading significantly. The low-rank aligned model, by contrast, maintains narrower and more uniform LS distributions: it trades peak fine-grained separability for consistent preservation of coarse, foundational representations across all tokens.

Table 2: Linear probe results across datasets of varying knowledge granularity. The fully aligned (full) vision tower’s features are more diverse (wider range of linear probe acc. on fine-grained datasets). ImageNet is general classification dataset. STF-dogs and CUB are fine-grained classification dataset of dogs and birds. Ranks of LoRA are 32/64/128. Range is reported by the maximum and minimum.

dataset	fine-grained	full	rank=128	rank=64	rank=32
ImageNet	×	83.6~100.0	99.9~100.0	99.8~100.0	99.8~100.0
STF-dogs	✓	32.8~95.7	39.2~88.0	36.3~85.9	36.7~86.8
CUB	✓	49.9~94.0	61.2~91.7	60.8~91.5	61.6~91.5

This asymmetry directly supports the progressive fusion hypothesis—low-rank alignment retains the general knowledge backbone while deferring entity-specific fusion to subsequent tuning stages.

Representation geometry. We quantify how alignment reshapes the angular structure of the visual feature space. For each entity sampled from ImageNet, we compute the minimum pairwise cosine angle among its corresponding visual tokens—a smaller angle indicates that tokens have spread farther apart on the hyperspherical surface. As shown in Figure 5, the horizontal axis tracks alignment progress and the vertical axis reports this angular coverage. The fully aligned model covers a progressively wider angular span, at times exceeding half of the hyperspherical surface, indicating aggressive spread and fusion of token representations. The low-rank aligned model, by contrast, maintains a compact angular footprint throughout, preserving the original structural layout of the visual features.

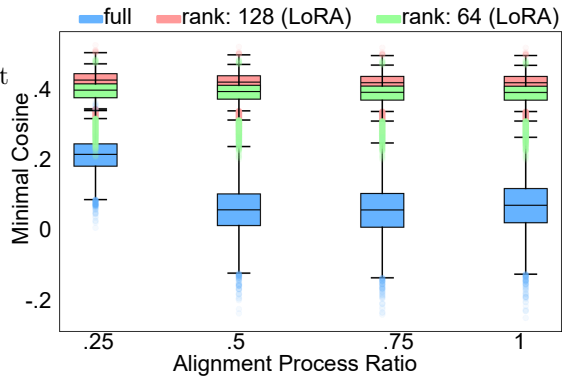


Fig. 5: Full aligned vision tokens have larger coverage of each given input images in ImageNet. Each image’s feature angle coverage (*i.e.*, minimal cosine $\downarrow := \min_{i,j} \cos(i, j)$, $i, j \in \{\text{visualtokens}|\text{images}\}$).

Feature manifold. Figure 6 visualizes the token-wise feature manifolds (where colors represent different tokens), corroborating the angular analysis above. Under full alignment, per-token manifolds are geometrically diverse and spatially dispersed across the feature space, reflecting aggressive restructuring of visual representations. Under low-rank alignment, token clusters are markedly more homogeneous in shape and collectively occupy a more compact region, confirming that the low-rank constraint induces a structural bias toward conservative, stable representations. This geometric regularity is consistent with updates confined to a dominant low-rank subspace, preserving the spatial organization inherited from the pre-trained encoder rather than aggressively overwriting it.

Empirical summary. Low-rank alignment introduces systematic implicit biases that manifest at multiple levels. At the behavior level, it produces conservative and decisive perception. At the feature level, it preserves per-token knowledge structure and prevents the LS-curse. Geometrically, it maintains homogeneous and structurally stable visual representations while preventing the premature spread and overgeneralization of entity-level features observed in full alignment. These converging observations motivate the theoretical analysis in Section 3, where we identify the mathematical mechanisms underlying these biases.

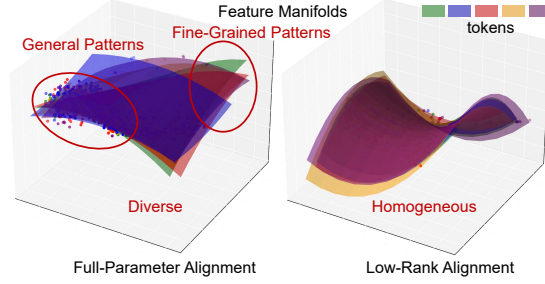


Fig. 6: The feature manifolds of each visual token, as depicted in Figure 1, are diverse under full alignment and homogeneous under low-rank alignment (LoRA, rank 128). The surface is fitted using a polynomial, retaining first- and second-order information. Each colored manifold represents the token feature distribution on MS-COCO.

3 Theoretical Study: Preferences on Parameter and Representation Spaces, Flatness and Robustness

The empirical study in Section 2 reveals that low-rank alignment exhibits systematic implicit biases: conservative behavior, preservation of per-token knowledge structure, and geometrically stable visual representations. Theoretical analysis is provided to identify the sufficient conditions under which such patterns arise and derive their mathematical characterizations. *Due to the presence of higher-order tensors beyond second order in our derivations, we adopt the Einstein summation convention by default, wherein duplicated indices imply summation.*

Analysis model: unconstrained features. We adopt the Unconstrained Feature Model [35] (UFM) as our analytical framework. UFM assumes that an over-parameterized neural network can fit any feature, thereby treating features H as free variables rather than fixed outputs of a specific architecture. This enables us to study global properties of the learned feature structure, such as which directions are preferentially reinforced. In practice, this assumption is well-supported by the over-parameterized nature of modern LLMs used in the alignment stage.

Low-rank modeling. Consider a low-rank adaptation parameterized by θ with forward mapping $O(\theta, H)$ and total loss

$$F(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathcal{L}(O(\theta, H(x)), y)],$$

$$O^i = W_0^{i,j} H^j + b^i + l^i(H; \theta),$$

where $O^i \in \mathbb{R}^p$ and $H^j \in \mathbb{R}^q$ denote the output and input features indexed by i and j ; $W_0^{i,j} \in \mathbb{R}^{p \times q}$ and $b^i \in \mathbb{R}^p$ are the frozen base weight and bias; $l^i(H; \theta)$

is the low-rank adaptation term; $\mathcal{D}$ is the training dataset; and $\mathcal{L}$ is the loss function. For LoRA, $l^i(H; \theta) = B_k^i A_j^k H^j$ with low-rank factors $A_j^k \in \mathbb{R}^{r \times q}$, $B_k^i \in \mathbb{R}^{p \times r}$, and $\theta = \{B, A\}$. Analogous parameterizations for LoHa and LoKr, involving Hadamard factor pairs $B_1, B_2 \in \mathbb{R}^{p \times r}$, $A_1, A_2 \in \mathbb{R}^{r \times q}$ and Kronecker factor $C^s_t \in \mathbb{R}^{s \times t}$, are detailed in the Appendix.

3.1 Process Preference: Robustness to Feature Noise

Theorem 1 (Feature preferences: noise smoothing.). *Under the UFM assumption, the feature space with additive noise exhibits a weighted smoothing effect. Expanding features with zero-mean noise ξ , we have*

$$\mathbb{E}_\xi \left[\frac{\partial F}{\partial W_{0j}^i} \right] \simeq \frac{\partial F}{\partial W_{0j}^i}(\bar{H}) + \frac{1}{2} \frac{\partial^3 F}{\partial W_{0j}^i \partial H^k \partial H^l} \Sigma^{kl},$$

where $\Sigma^{kl} = \mathbb{E}[\xi^k \xi^l]$ is the feature noise covariance, F is the alignment loss function, W_0 is the original weight, $\bar{H}$ the mean of $H := \bar{H} + \xi$.

Corollary 1 (LoRA’s⁵ noise robustness preference .). *Substituting this into the LoRA gradient flow yields*

$$\begin{aligned} \dot{A}^p_j &\simeq -\eta B^i_p \left(\frac{\partial F}{\partial W_{0j}^i}(\bar{H}) + \frac{1}{2} \frac{\partial^3 F}{\partial W_{0j}^i \partial H^k \partial H^l} \Sigma^{kl} \right), \\ \dot{B}^i_p &\simeq -\eta \left(\frac{\partial F}{\partial W_{0j}^i}(\bar{H}) + \frac{1}{2} \frac{\partial^3 F}{\partial W_{0j}^i \partial H^k \partial H^l} \Sigma^{kl} \right) A^p_j, \end{aligned}$$

demonstrating an implicit smoothing effect along directions with feature noise. This shows that gradient flow preferentially aligns updates along subspaces that are robust to stochastic variations in the features, encoding an implicit bias towards flatter directions in the loss landscape.

Intuition and explanation. The core insight lies in the coupled gradient flow: $\dot{A} \sim B \cdot (\text{grad} + \text{noise})$ and $\dot{B} \sim (\text{grad} + \text{noise}) \cdot A$, where $\dot{A}$ and $\dot{B}$ denote the gradient flows of A and B , respectively, and $\cdot$ and $\otimes$ denote the Hadamard and Kronecker products where applicable. Since updates to A are scaled by B and vice versa, whichever directions are already dominant receive proportionally larger updates—a self-reinforcing dynamic. The noise-induced third-order term further amplifies this effect along directions with large feature-noise covariance, creating a Matthew effect: strong directions grow stronger while weak directions are suppressed. Consequently, LoRA naturally concentrates updates on noise-robust principal directions in the feature space.

⁵ Corollaries about LoHa and LoKr can be obtained. See Appendix.

Comparison: full vs. low-rank alignment. In full alignment, the weight matrix W is updated directly and all directions receive gradient updates on an equal footing. In low-rank alignment, A and B act as mutual subspace projectors, confining the effect of third-order smoothing terms to the principal subspace. This confinement favors conservative feature integration along dominant directions while suppressing changes in weaker subspaces, helping to explain the homogeneous manifold structure and preserved linear separability observed in Section 2.3. Here δ_i^j denotes the Kronecker delta arising in the projection identities. Analogous conclusions hold for LoHa and LoKr as detailed in the Appendix.

3.2 Steady-State Preferences: Flatness and Robustness

Theorem 1 characterizes which directions are reinforced *during* optimization. We now ask a complementary question: *where* do the low-rank parameters converge at the end of training?

Theorem 2 (Low-rank steady-state distribution). *Let the gradient flow of θ be $\dot{\theta}^\alpha = -\eta \frac{\partial F}{\partial \theta^\alpha}$. Define the effective low-rank subspace mapping $MH = l(H)$ (for example, $M = BA$ in LoRA, $M = B_1 A_1 * B_2 A_2$ in LoHa, $M = C \otimes BA$ in LoKr) and the corresponding effective loss:*

$$F_{\text{eff}}(M) = F(M) + \frac{\eta \log \det \tilde{D}(M)}{4} + \frac{\partial^2 F}{2\partial O^i \partial O^m} \Sigma^{jk} (W_{0j}^i + M_{\cdot j}^i)(W_0^m{}_{\cdot k} + M^m{}_{\cdot k}),$$

where $F(M)$ is the original loss expressed in the low-rank subspace M ; $\tilde{D}(M)$ is the covariance of gradient noise projected onto M , encoding flatness preference; Σ^{jk} is the covariance of feature noise. Then, under small-step stochastic gradient descent or noisy gradient flow, the steady-state M distributes approximately as:

$$p_{\text{stat}}(M) \propto \exp\left(-\frac{2}{\eta} F_{\text{eff}}(M)\right).$$

Remark 1. The effective loss F_{eff} augments F with two penalty terms. The log-determinant term penalizes subspaces where gradient noise is small or isotropic, biasing convergence toward flat regions of the loss landscape. The feature-noise term penalizes subspaces sensitive to input perturbations, biasing convergence toward noise-robust directions. Together, they ensure that steady-state parameters concentrate on subspaces that are simultaneously flat and robust.

Intuition and takeaways. The two penalty terms reveal a dual implicit bias: low-rank parameters are drawn toward configurations that are flat with respect to the loss and robust with respect to input perturbations. The net effect is that the learned subspace naturally avoids directions sensitive to small variations—the system settles into configurations where outputs remain stable under feature noise, inherently reducing susceptibility to adversarial perturbations and aligning with the conservative behavior observed empirically in Section 2.

Table 3: Comparison of between low-rank operators on Vicuna-13B. Full-parameter alignment serves as the baseline. The selection of low rank operators is highly correlated with rank. A suitable range of rank can be empirically found with better perception.

	full	LoRA			LoKr			LoHa		
metric	–	32	64	128	32	64	128	32	64	128
MME-P	236.6	375.9	288.0	–	79.4	78.5	449.3	380.6	621.4	1079.7
POPE	42.2	50.0	19.7	–	9.0	12.7	46.8	21.6	51.5	82.8

Table 4: Ablation on low-rank operators across model scales. Results are reported on 7B / 13B MobileVLM-V2 models. Best results in **bold**.

operators on 7/13B	<i>perception</i> ↑	<i>knowledge</i> ↑	<i>reasoning</i> ↑	<i>hallucination</i> ↑
full parameter	521.0/236.6	23.4/29.5	86.8/52.5	35.0/66.9
LoRA: mat. mul.	551.8/375.9	27.0/36.3	251.4/147.5	40.9/66.7
LoHa: Hadamard	965.3/1079.7	38.5/43.0	332.5/219.6	80.4/82.9
LoKr: Kronecker	844.4/449.3	37.4/40.7	312.5/62.5	74.8/70.2

4 Ablation Study: Low-rank Vision-Language Alignment

Overview. The ablation studies investigate the effects of rank, model scale, alignment depth, low-rank operators, vision encoder choice, frozen *vs.* unfrozen encoder settings, low-rank placement, learning rate schedules, and overfitting controls. Unless otherwise specified, experiments follow the default settings with targeted modifications. Extended results are provided in the Appendix.

Ranks. In perception-intensive benchmarks, increasing the rank of the low-rank module yields significant performance gains, particularly for larger models. This is evident from the improvements on MME-perception and POPE using the 13B model In Table 3, in contrast, knowledge-intensive or task-specific benchmarks (*e.g.*, GQA, ScienceQA, TextVQA) show no comparable gains from rank increases, suggesting that the performance bottleneck in these tasks lies elsewhere. We explore this further in the extended-data ablation in the Appendix.

Operators. Ablation of operators is conducted using Vicuna-7B and -13B under default settings with the best-tuned rank for each operator. The results reveal operator-specific strengths across benchmarks, indicating that the choice of low-rank operator is a promising avenue for future optimization.

Model scales and architectures. As shown in Tables 5 and 6, low-rank alignment consistently matches or outperforms full fine-tuning across scales from 1.4B to 14B, and this advantage generalizes to architecturally distinct models including Qwen3-8B/14B under a one-shot protocol. Two exceptions—POPE and MMMU-Pro—appear consistently across both model families, where higher-rank updates partially overwrite pre-trained knowledge structures. The rank dependence of

Table 5: Performance across model scales and alignment methods. Full-parameter alignment serves as the baseline. MobileVLM-V2 results use rank 64 for all low-rank methods.

model	metric	Full	LoRA	LoKr	LoHa
1.4B	MME-P	208.1	568.7	766.6	450.8
	POPE	2.5	51.5	72.8	63.1
2.7B	MME-P	15.8	52.0	432.6	318.3
	POPE	0.5	15.0	71.3	66.1
7B	MME-P	521.0	551.8	825.1	608.3
	POPE	35.0	37.2	70.7	67.3
13B	MME-P	236.6	288.0	78.5	621.4
	POPE	66.9	44.8	37.6	67.4

Table 6: Performance of alignment methods on Qwen3. Low-rank alignment works on most of the benchmarks, with notable exceptions on POPE and MMMU-Pro where rank-constrained updates prove insufficient.

metric	full 8B	LoRA ($r=128$, 8B)	Δ	full 14B	LoRA ($r=32$, 14B)	Δ
GQA	28.4	37.2	+8.8	6.6	20.2	+13.6
MMBench	18.1	48.7	+30.6	12.0	45.5	+33.5
MME-P	536.0	752.3	+216.3	526.3	988.9	+462.6
MME-R	212.1	242.1	+30.0	234.3	308.6	+74.3
MMMU-Pro	12.1	5.7	-6.4	11.8	11.8	0
POPE	92.7	25.1	-67.6	—	62.3	—
ScienceQA	63.7	71.3	+7.6	59.3	71.9	+12.6
TextVQA	5.9	9.7	+3.8	4.2	10.2	+6.0

these anomalies suggests that the low-rank constraint acts as a regularizer limiting deviation from the pre-trained weight manifold, a property that becomes increasingly important as rank grows.

Other ablation on default settings. These ablations jointly validate the default configuration of our alignment pipeline across four dimensions. (a) Visual encoder. Both DINOv2 [36] and SigLIPv2 [46] backbones benefit from low-rank alignment over full fine-tuning, with consistent gains on GQA and MMBench. The improvement is more pronounced under DINOv2, suggesting that the low-rank constraint interacts more favorably with dense spatial features than with the global token representations of SigLIPv2. (b) Encoder freezing. Freezing the visual encoder during alignment consistently outperforms unfreezing across nearly all metrics, regardless of backbone or alignment method. Unfreezing introduces instability, particularly on TextVQA and MME-p, indicating that gradient updates to the encoder during alignment disrupt pre-trained visual representations rather than refining them. (c) Adaptation scope. Applying low-rank adaptation jointly to both attention and MLP modules yields the best overall

Table 7: Ablation studies on alignment design choices. All configurations use full fine-tuning in the instruction tuning stage with other default settings unchanged. Row colors denote ablation groups: visual encoder , frozen vs. unfrozen encoder , low-rank adaptation scope , and overfitting control .

bench.	GQA	MME-p	POPE	TextVQA	MMB-dev	SQA
dinov2-full	61.56	1503.69	86.86	51.56	72.51	74.27
dinov2-LoRA	61.88	1481.45	86.52	55.19	74.66	76.15
siglipv2-full	62.19	1406.61	88.24	29.58	72.68	73.42
siglipv2-LoRA	62.73	1474.70	87.52	30.82	73.37	72.08
frz. clip-LoRA	62.47	1511.61	87.52	44.73	72.41	71.42
un-frz. clip-full	59.98	1444.91	85.79	40.18	71.74	72.53
un-frz. clip-LoRA	60.18	1373.54	86.49	42.97	71.22	72.53
un-frz. dinov2-full	61.78	1403.20	87.60	28.61	71.82	72.38
un-frz. dinov2-LoRA	63.07	1481.98	88.18	37.27	76.03	72.04
all (attn. + MLP)	62.47	1511.61	87.52	44.73	72.41	71.42
MLP only	59.99	1384.71	87.95	41.29	72.77	75.41
attn. only	62.26	1391.84	87.66	39.38	69.62	74.32
full + weight decay	61.30	1432.68	86.21	43.55	70.34	70.61
default (LoRA)	62.47	1511.61	87.52	44.73	72.41	71.42

performance, with the full configuration outperforming MLP-only and attention-only variants on MME-p and TextVQA. This suggests that the two module types capture complementary aspects of visual-language alignment and should not be decoupled. (d) Overfitting. Our default low-rank alignment matches or exceeds full fine-tuning with tuned weight decay across all six benchmarks, confirming that the performance advantage is not attributable to overfitting in the full fine-tuning baseline. The low-rank constraint is an effective overfitting regularizer.

5 Related Works

Low-rank Adaptation and Vision-Language Models. Low-rank adaptation [14] introduces trainable low-rank matrices into frozen linear layers for parameter-efficient fine-tuning, with extensions including QLoRA [9], LoHa [17], LoKr [51], and further variants [13, 23, 27, 33, 48, 56]. On the VLM side, Flamingo [1] pioneered visual-language integration, inspiring a broad family of models [6, 8, 21, 26, 50, 57] that differ in scale, architecture, and alignment strategy. In this context, VLoRA [32] suggests that cross-modal information transfer may be intrinsically low-rank, providing empirical motivation for applying low-rank adaptation in the alignment stage. However, whether such constraints introduce systematic implicit biases in visual token representations during alignment remains uncharacterized. Our work targets this gap: rather than proposing a new method, we analyze the alignment dynamics induced by LLM-side low-rank adaptation—a question orthogonal to connector design. Since many open-source VLMs release only trained weights without the full reproduction materials required to replicate their pretraining stage, we adopt a more general alignment pipeline that

abstracts away model-specific details while remaining representative of mainstream practice, as evidenced by its continued use in recent popular models such as Kimi2.5-K2 [45]. Our simplified pipeline is therefore deliberately distinct from architecture-specific designs such as Qwen3-VL-style [3] multi-layer visual injection (deepstack), which depends on proprietary training configurations beyond public reproduction, and we cannot reproduce the alignment stages of Qwen3-VL. Moreover, most current open-source models employ a ViT inherited from a previously aligned version rather than trained from unaligned ones, making it difficult to isolate for academic research.

Implicit Bias of Low-Rank Adaptation and Matrix Factorization. A classical line of work studies the implicit regularization of factorized parameterizations: gradient descent on matrix factorization is biased toward low-rank / minimal nuclear-norm solutions [11], an effect that deep factorization intensifies [2] and that need not be captured by any norm [40]; more recent analyses study LoRA from a parameter-efficiency or optimization-landscape standpoint [4, 18]. Our setting departs from this line in four respects—a non-linear LLM backbone, fine-tuning from a pretrained checkpoint, stochastic (noisy) dynamics, and a cross-entropy alignment objective—and consequently yields a *flat-gradient, noise-robust subspace* bias (Sec. 3) rather than the min-norm/low-rank bias of the classical setting; we expand on these four differences in the Appendix.

6 Conclusion

This work investigates the implicit biases introduced by low-rank adaptation during vision-language alignment—a stage widely treated as requiring full-parameter updates. We show that low-rank alignment not only reduces training costs but consistently outperforms full fine-tuning across model scales from 1.4B to 14B parameters. Through behavioral, feature-level, and geometric analyzes, we identify three systematic biases: a shift from hallucinatory to conservative perception, preservation of per-token linear separability (in contrast to full alignment), and structurally homogeneous visual representations. These observations reveal that low-rank alignment retains general modality-specific knowledge during the alignment stage while deferring entity-specific fusion to subsequent instruction tuning. Two theorems ground this behavior in optimization theory, showing that low-rank gradient flow preferentially reinforces noise-robust feature directions and that steady-state parameters concentrate on flat, perturbation-robust subspaces. Extensive experiments across over 100 alignment configurations—spanning model scales, operator families and vision encoders—validate these findings and offer practical guidance for vision-language alignment.

Acknowledgements

We thank the ECCV committee and reviewers for their careful reading and constructive feedback. We are also grateful to colleagues for helpful discussions and to the institutions that provided computational resources used in this study. The first author carried out and led this work as an independent researcher.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: A visual language model for few-shot learning. *NeurIPS* **35**, 23716–23736 (2022)
2. Arora, S., Cohen, N., Hu, W., Luo, Y.: Implicit regularization in deep matrix factorization. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2019)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631* (2025)
4. Biderman, D., Portes, J., Gonzalez Ortiz, J.J., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., Cunningham, J.P.: LoRA learns less and forgets less. *Transactions on Machine Learning Research (TMLR)* (2024)
5. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: *European Conference on Computer Vision*. pp. 370–387. Springer (2024)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 24185–24198 (2024)
7. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., 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)
8. Chu, X., Qiao, L., Zhang, X., Xu, S., Wei, F., Yang, Y., Sun, X., Hu, Y., Lin, X., Zhang, B., et al.: Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766* (2024)
9. Detrmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient fine-tuning of quantized llms. *Advances in Neural Information Processing Systems* **36** (2024)
10. Dong, M., Cai, H., Li, J., Zhou, S., Ren, B., Peng, K., Fu, Y.: Visnec: Measuring and leveraging visual necessity for multimodal instruction tuning. *arXiv preprint arXiv:2603.01195* (2026)
11. Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., Srebro, N.: Implicit regularization in matrix factorization. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2017)
12. Han, Z., Gao, C., Liu, J., Zhang, S.Q., et al.: Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608* (2024)
13. Hayou, S., Ghosh, N., Yu, B.: Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354* (2024)
14. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: *International Conference on Learning Representations*
15. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6700–6709 (2019)
16. HuggAN: WikiArt Dataset. <https://huggingface.co/datasets/huggan/wikiart> (2022), accessed: 2024-12-19

17. Hyeon-Woo, N., Ye-Bin, M., Oh, T.H.: Fedpara: Low-rank hadamard product for communication-efficient federated learning. In: The Twelfth International Conference on Learning Representations (2022)
18. Jang, U., Lee, J.D., Ryu, E.K.: LoRA training in the NTK regime has no spurious local minima. In: International Conference on Machine Learning (ICML) (2024)
19. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proc. CVPR workshop on fine-grained visual categorization (FGVC). vol. 2 (2011)
20. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **123**(1), 32–73 (2017)
21. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML* pp. 19730–19742 (2023)
22. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355* (2023)
23. Li, Y., Yu, Y., Liang, C., He, P., Karampatziakis, N., Chen, W., Zhao, T.: Loftq: Lora-fine-tuning-aware quantization for large language models. *arXiv preprint arXiv:2310.08659* (2023)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *NeurIPS* **36** (2024)
27. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353* (2024)
28. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
29. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
30. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* **35**, 2507–2521 (2022)
31. Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., Zhu, S.C.: Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214* (2021)
32. Ma, F., Xue, H., Zhou, Y., Wang, G., Rao, F., Yan, S., Zhang, Y., Wu, S., Shou, M.Z., Sun, X.: Visual perception by large language model’s weights. *Advances in Neural Information Processing Systems* (2024)
33. Meng, F., Wang, Z., Zhang, M.: PiSSA: Principal singular values and singular vectors adaptation of large language models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
34. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 international conference on document analysis and recognition (ICDAR). pp. 947–952. IEEE (2019)

35. Mixon, D.G., Parshall, H., Pi, J.: Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis* **20**(2), 11 (2022)
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
37. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems* **24** (2011)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *ICML*. pp. 8748–8763. PMLR (2021)
39. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024)
40. Razin, N., Cohen, N.: Implicit regularization in deep learning may not be explainable by norms. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2020)
41. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2015), <https://arxiv.org/abs/1409.0575>
42. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems* **35**, 25278–25294 (2022)
43. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: *Proceedings of ACL* (2018)
44. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards VQA models that can read. In: *CVPR*. pp. 8317–8326 (2019)
45. Team, K.: Kimi k2.5: Visual agentic intelligence. <https://www.kimi.com/blog/kimi-k2-5> (2025)
46. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025)
47. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
48. Xu, Y., Xie, L., Gu, X., Chen, X., Chang, H., Zhang, H., Chen, Z., Zhang, X., Tian, Q.: Qa-lora: Quantization-aware low-rank adaptation of large language models. *arXiv preprint arXiv:2309.14717* (2023)
49. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), <https://arxiv.org/abs/2505.09388>

50. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
51. Yeh, S.Y., Hsieh, Y.G., Gao, Z., Yang, B.B., Oh, G., Gong, Y.: Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In: The Twelfth International Conference on Learning Representations (2023)
52. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. *National Science Review* **11**(12), nwae403 (2024)
53. Yu, W., Yang, Z., Ren, L., Li, L., Wang, J., Lin, K., Lin, C.C., Liu, Z., Wang, L., Wang, X.: Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. arXiv preprint arXiv:2408.00765 (2024)
54. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al.: Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 15134–15186 (2025)
55. Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al.: Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792 (2023)
56. Zhou, S., Wang, S., Yuan, Z., Shi, M., Shang, Y., Yang, D.: Gsq-tuning: Group-shared exponents integer in fully quantized training for llms on-device fine-tuning. In: *Findings of the Association for Computational Linguistics: ACL 2025*. pp. 22971–22988 (2025)
57. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)