DYNAMIC MODALITY SCALING FOR VISION-LANGUAGE MODEL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in vision-language models have achieved remarkable results in making language models understand visual inputs. However, a unified approach to aligning these models remains a challenge. In LLM-centric VLMs, language dominance is particularly plausible, because the decoder holds a strong prior over answers and explanations. Yet visual dominance is also possible in OCR or scene-heavy settings, where noisy patches overwhelm a simple question. The correct intervention must therefore be conditional, rather than globally image-boosting or text-suppressing. In this paper we propose dynamic modality scaling (VL-DyMS), a framework built on two frozen unimodal teachers and a multimodal student. Each teacher is obtained by masking one input: the text teacher sees only the question, the vision teacher only the image. The student is supervised by an answer loss and by a modality-aware distillation loss against each teacher. A dynamic controller adjusts the contribution of textual and visual knowledge at every step, so supervision follows the modality the student is currently underusing rather than a fixed schedule. We evaluate VL-DyMS on five datasets using Qwen2.5-VL and LLaVA-1.5. It shows an improved performance over the pretrained baseline and standard LoRA adaptation.

1 INTRODUCTION

Vision-language models (VLMs) have moved quickly from research prototypes to general-purpose systems. Models such as LLaVA Liu et al. (2024), Qwen-VL Bai et al. (2023), Qwen2.5-VL Bai et al. (2025), and InternVL Chen et al. (2024) combine pretrained vision and language components to answer questions, describe scenes, reason over visual content, and read text in images. Despite their strong performance, learning from the two modalities remains unbalanced. In particular, the language component starts from a strong pretrained model with an established prior over answers and explanations, while the visual modality must carry dense information that is not always recoverable from language priors.

A typical LLM-centric VLM connects a pretrained vision encoder to a pretrained language model through a multimodal projector. The image is first encoded into visual representations and then projected into a representation space that can be processed together with textual tokens. The text input, in contrast, is processed directly by the pretrained language model. This creates a structural asymmetry between the two modalities, language enters multimodal learning with a strong pretrained representation, while visual information must first be encoded and mapped into a form that the language model can use.

One consequence of this asymmetry is that VLMs can rely heavily on text even when visual information is available. A model asked what colour a bus is may answer *yellow* because it inspected the image, or because buses are commonly associated with yellow. Both answers receive the same accuracy score. Goyal et al. (2017) established this issue in visual question answering, and it motivated the construction of more balanced datasets. However, the underlying behaviour did not disappear with them. Frank et al. (2021) and Parcalabescu & Frank (2023) later showed that vision-language models can remain strongly dependent on text even when their final answers are correct.

Work on multimodal training has also shown that the two modalities do not necessarily learn at the same rate. Wang et al. (2020) attribute multimodal imbalance to individual modalities overfitting and generalizing at different rates within the same jointly trained network. Peng et al. (2022) show

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

that a dominant modality can suppress the optimization of the other modality and propose gradient modulation to compensate. Wu et al. (2022) describe a similar behaviour as a greedy tendency to rely on whichever modality reduces the training loss faster. These findings suggest that combining vision and language in the same model does not guarantee balanced learning between them.

At the same time, the visual modality has different learning and scaling requirements from language. An image can carry information about objects, attributes, spatial relations, scene text, and other visual details that are difficult to represent through a short textual description. Recent work in Tong et al. (2026) provides further evidence that vision and language do not scale equally. Its IsoFLOP analysis shows that language follows a near Chinchilla-like scaling behaviour, where the compute-optimal amount of language data grows approximately as $C^{0.53}$, while vision is more data-hungry and grows approximately as $C^{0.63}$. Therefore, as compute increases, the optimal amount of visual data grows faster than the optimal amount of language data, and a fixed data mixture cannot remain optimal for both modalities.

Although this observation concerns data allocation during multimodal pretraining, it motivates a broader question for multimodal learning: if vision and language learn and scale differently, why should their contribution remain fixed throughout training? A fixed weighting assumes that the relative learning needs of the two modalities remain unchanged. In practice, one modality may learn faster or contribute more strongly than the other at a particular point in training.

This motivates a dynamic rather than fixed approach to modality scaling. The objective is not to force vision and language to contribute equally, but to adjust the learning pressure according to the current state of the two modalities.

Based on this motivation, we propose Dynamic Modality Scaling (VL-DyMS), a framework that dynamically adjusts modality-specific weights according to the current contribution to the multimodal learning. The contributions of this paper are as follows:

- Proposed Dynamic Modality Scaling (VL-DyMS), a framework that dynamically adjusts each modality's contribution during training.
- Evaluated the proposed VL-DyMS across five vision-language benchmarks and two multimodal backbones.

2 Related Work

2.1 Vision-Language Alignment

Early vision-language models focused on aligning image and text representations in a shared space. CLIP Radford et al. (2021) and ALIGN Jia et al. (2021) use large-scale contrastive learning, while BLIP-2 Li et al. (2023a) introduces a Q-Former to connect frozen vision and language models. Generative VLMs such as LLaVA Liu et al. (2023; 2024), Qwen-VL Bai et al. (2023), Qwen2.5-VL Bai et al. (2025), and InternVL Chen et al. (2024) instead integrate visual representations into pretrained language models. In these LLM-centric architectures, strong language priors may remain influential during multimodal adaptation.

Recent studies examine this alignment more explicitly. MIR Huang et al. (2025) measures multimodal integration during VLM pretraining and links stronger integration to downstream performance, while Schrodi et al. (2025) identify information imbalance between images and captions as a major source of modality gap and object bias. These findings suggest that alignment depends not only on representational proximity, but also on how effectively each modality contributes information.

2.2 Modality Imbalance and Scaling

Joint multimodal training does not necessarily optimize all modalities equally.Wang et al. (2020) relate this imbalance to different generalization rates, Peng et al. (2022) show that dominant modalities can suppress weaker ones, and Wu et al. (2022) describe a tendency to rely on modalities that reduce training loss faster. Dynamic Modality-Gap Learning Yang et al. (2024) addresses differences in convergence rates, while Wei & Hu (2024) targets gradient conflicts between multimodal and unimodal objectives.

Importantly, equal modality dependence is not necessarily optimal. Wei et al. (2025) show that optimal dependence varies with modality-specific prediction variance. Similarly, controlled multimodal scaling experiments Tong et al. (2026) show different compute-optimal scaling laws for language and vision. Together, these results motivate adaptive rather than uniformly equal modality weighting.

2.3 Modality Rebalancing and Knowledge Distillation

Several methods dynamically regulate modality learning. Wang et al. (2020) adjusts losses based on generalization behaviour, OGM by Peng et al. (2022) modulates dominant-modality gradients, and prototype-based modulation Fan et al. (2023) strengthens slower modalities. Other approaches adapt the objective itself: Dynamic Modality-Gap Learning Yang et al. (2024) combines supervised and contrastive learning, MMPareto by Wei & Hu (2024) resolves gradient conflicts, and SCOPE by Yang et al. (2026) preserves complementary modality-specific information while aligning shared semantics.

Knowledge distillation provides an alternative source of modality-specific supervision. Du et al. (2021) use unimodal teachers with fixed contributions, whereas CKD in Zhang et al. (2024) estimates modality-specific causal effects to dynamically weight distillation in multimedia recommendation. CMoB in Wang et al. (2025) further estimates modality contribution at the sample level and dynamically strengthens insufficiently learned modalities.

VL-DyMS builds on these ideas but targets LLM-centric VLM adaptation. Unlike methods developed mainly for classification or recommendation, it combines separate visual and textual distillation signals with dynamic scaling while retaining the original pretrained multimodal model as a behavioral reference.

2.4 Measuring Modality Contribution

Final task accuracy alone cannot determine whether both modalities contribute meaningfully. Input-ablation studies Frank et al. (2021) reveal strong textual dependence in multimodal transformers, while MM-SHAP study Parcalabescu & Frank (2023) uses Shapley values to quantify modality contribution. MIR by Huang et al. (2025) measures representation-level multimodal integration, whereas POPE introduced by Li et al. (2023b) evaluates object hallucination when language priors override visual evidence. These diagnostics complement task accuracy by measuring how multimodal information is used.

2.5 Comparison of Dynamic Modality Rebalancing Methods

Table 1 compares representative modality rebalancing approaches. Existing methods dynamically regulate modality learning through loss weighting, gradient modulation, causal valuation, or modality-specific supervision. However, they do not jointly combine dynamic modality-specific distillation with the behavior of the original pretrained multimodal model in an LLM-centric VLM setting. VL-DyMS addresses this gap by dynamically scaling visual and textual distillation signals according to the learning state while retaining the frozen pretrained model as a behavioral reference.

3 Methodology

Dynamic Modality Scaling (VL-DyMS) consists of a frozen pretrained multimodal baseline, two frozen masked teachers, and a trainable multimodal student. The student is initialized from the same vision-language backbone and adapted using LoRA Hu et al. (2022), while the backbone remains frozen. For each training example, the student and pretrained baseline predictions are evaluated against the target and routed to one of three branches: (i) preserve the student when it is correct, with no optimizer update; (ii) recover pretrained behavior when the student is wrong but the baseline is correct; or (iii) apply dynamically weighted text-vision distillation when both are wrong. Dynamic modality scaling is therefore activated only in the third branch. Figure 1 illustrates the framework.

Table 1: Comparison of representative modality rebalancing methods.

Method	Base Model	Dynamic Rebalancing	Modality-Specific Signal	Learning-State Adaptive	Distillation-Based	Pretrained Baseline Reference
Gradient Blending Wang et al. (2020)	Multimodal classification	✓	✓	✓	×	×
OGM Peng et al. (2022)	Multimodal classification	✓	✓	✓	×	×
Dynamic Modality-Gap Learning Yang et al. (2024)	Multimodal classification	✓	✓	✓	×	×
MMPareto Wei & Hu (2024)	Multimodal classification	✓	✓	✓	×	×
CKD Zhang et al. (2024)	Multimedia recommendation	✓	✓	×	✓	×
CMoB Wang et al. (2025)	Multimodal classification	✓	✓	✓	×	×
SCOPE Yang et al. (2026)	Multimodal learning	×	✓	×	×	×
VL-DyMS (ours)	**LLM-centric VLM**	✓	✓	✓	✓	✓

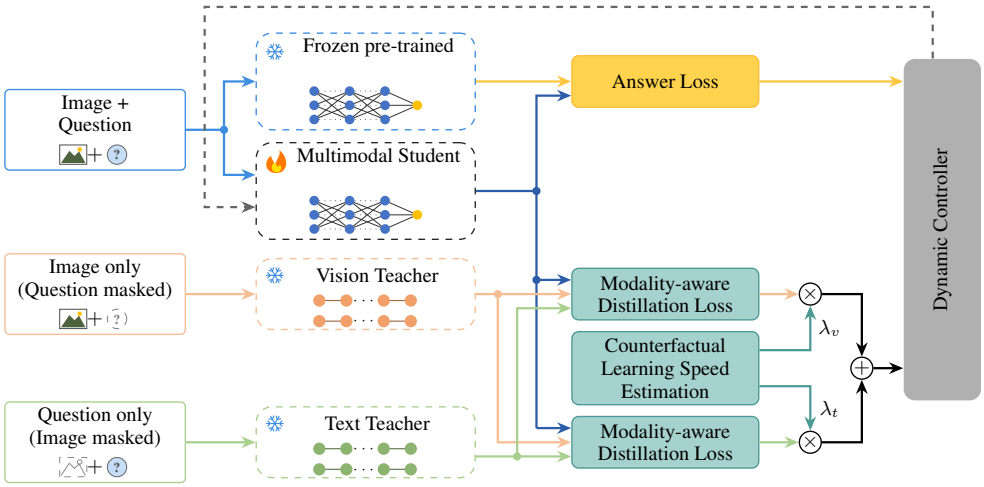


Figure 1: The proposed VL-DyMS framework.

3.1 Problem Formulation

Let $\mathcal{D} = \{(I_i, q_i, a_i)\}_{i=1}^{N}$ denote a visual question answering dataset, where I_i, q_i, and a_i are the image, question, and target answer. We denote the frozen pretrained model by f_{θ_0} and the trainable multimodal student by f_θ. Only the student's LoRA parameters and the VL-DyMS controller are optimized.

For each example, three views are constructed: the complete image-question input (I_i, q_i), a text-only view with the image removed, and a vision-only view with the original question removed. The complete view is processed by both the frozen baseline and the student, while the masked views are processed by frozen copies of the same pretrained backbone and serve as modality-specific teachers. Dynamic modality supervision is used only when both the student and pretrained baseline fail; otherwise, the student is either preserved or guided to recover pretrained behavior.

3.2 Baseline and Uni-modal Teachers

The frozen pretrained model evaluated on the complete input acts as a fixed behavioral reference. The text teacher observes the original question without the image, while the vision teacher observes the image without the original question. Both are frozen copies of the same pretrained backbone, eliminating the need for separately trained unimodal teachers. Throughout training, the baseline and

both teachers remain fixed, while only the student's adaptation parameters and the dynamic controller are optimized.

3.3 Modality-Aware Distillation

The student is trained with the standard autoregressive answer loss $\mathcal{L}_{\text{ans}}$ on the complete image-question input. Modality-specific supervision is provided through separate textual and visual distillation losses:

$$\mathcal{L}_{\text{KD}}^{t} = \beta_{\text{logit}}\mathcal{L}_{\text{KL}}\left(f_{\theta}^{\text{full}}, f_{\theta_0}^{t}\right) + \beta_{\text{rep}}\mathcal{L}_{\text{rep}}\left(z_{\theta}^{t}, z_{\theta_0}^{t}\right), \tag{1}$$

$$\mathcal{L}_{\text{KD}}^{v} = \beta_{\text{logit}}\mathcal{L}_{\text{KL}}\left(f_{\theta}^{\text{full}}, f_{\theta_0}^{v}\right) + \beta_{\text{rep}}\mathcal{L}_{\text{rep}}\left(z_{\theta}^{v}, z_{\theta_0}^{v}\right), \tag{2}$$

where f_{θ}^{full} is the multimodal student output, $f_{\theta_0}^{t}$ and $f_{\theta_0}^{v}$ are the frozen text-only and vision-only teacher outputs, and z^t and z^v are the corresponding modality representations. $\mathcal{L}_{\text{KL}}$ is the temperature-scaled distillation loss Hinton et al. (2015), while $\mathcal{L}_{\text{rep}}$ is the representation-matching loss. β_{logit} and β_{rep} weight the output- and representation-level terms. The two modality losses remain separate so that textual and visual supervision can be regulated independently.

3.4 Dynamic Modality Scaling

VL-DyMS estimates the current learning state from modality-specific distillation behavior, the evolution of the task objective, and the relationship between the pretrained baseline and the current student. A dynamic controller produces the textual and visual scaling coefficients at training step k:

$$\left(\lambda_t^{(k)}, \lambda_v^{(k)}\right) = \pi_{\phi}\left(\mathbf{s}_k\right), \tag{3}$$

where $\mathbf{s}_k$ denotes the current multimodal learning state and π_{ϕ} is the VL-DyMS controller. The objective is not to enforce equal modality contribution, but to adjust their relative training influence according to the current learning condition. Thus, visual or textual supervision can be strengthened when the corresponding modality requires greater correction. The learning state is defined as:

$$\mathbf{s}_k = [g_k, c_t^{(k)}, c_v^{(k)}, |c_t^{(k)} - c_v^{(k)}|, \xi_k, \Delta_k] \tag{4}$$

capturing the current text-vision representation gap, reliability-weighted teacher confidences, teacher disagreement, and task-loss change. The controller adjusts the relative textual and visual supervision according to this learning state rather than enforcing equal modality contribution. The three-branch objective performs no update when the student is correct, applies pretrained-behavior recovery through KL and representation matching when only the baseline is correct, and otherwise optimizes the answer loss with dynamically weighted textual and visual distillation.

3.5 VL-DyMS Training Objective

The final objective combines the supervised answer loss with dynamically weighted modality-specific distillation:

$$\mathcal{L}_{\text{VL-DyMS}} = \mathcal{L}_{\text{ans}} + \lambda_{\text{KD}}\left[\lambda_t^{(k)}\mathcal{L}_{\text{KD}}^{t} + \lambda_v^{(k)}\mathcal{L}_{\text{KD}}^{v}\right], \tag{5}$$

where λ_{KD} controls the overall distillation strength, while $\lambda_t^{(k)}$ and $\lambda_v^{(k)}$ determine the relative textual and visual contributions. The frozen pretrained model serves only as a behavioral reference for branch selection and learning-state estimation and does not introduce an additional prediction loss in Equation 5.

4 Experiments

4.1 Datasets

We evaluate VL-DyMS on five vision-language benchmarks selected to represent different dependencies on textual and visual information: TextVQA Singh et al. (2019), GQA Hudson & Manning

Table 2: Summary of the datasets used to evaluate VL-DyMS.

Dataset	Main Capability	Full Dataset Size
TextVQA Singh et al. (2019)	Scene-text understanding and reasoning	45,336 questions & 28,408 images
GQA Hudson & Manning (2019)	Compositional visual reasoning	22.67M questions & 113,018 images
TallyQA Acharya et al. (2019)	Visual counting and counting-based reasoning	287K questions & 165K images
POPE Li et al. (2023b)	Object hallucination and visual grounding	27,000 question-answer pairs
POPEv2 Li et al. (2026)	Counterfactual object hallucination	1,000 questions & 1,000 images

Table 3: Sample question-answer pairs from TallyQA Acharya et al. (2019)

Image	Question	Answer
	How many chairs are pictured?	2
	How many chairs are at the table?	2
	How many things are purple?	1
	How many curtains are there?	2
	How many chairs?	2
	How many pillows are on the couch?	1
	How many chairs are near the table?	2
	How many stuffed animals are visible?	1
	How many chairs are visible?	2
	How many table legs are on the rug?	1
	How many pillows are on the futon?	1

(2019), TallyQA Acharya et al. (2019), POPE Li et al. (2023b), and POPEv2 Li et al. (2026). Together, these benchmarks cover scene-text understanding, compositional visual reasoning, counting, and object hallucination. Table 2 summarizes the role of each benchmark.

Table 3 shows an example from TallyQA Acharya et al. (2019), illustrating image, questions and answers pair.

4.2 Evaluation Protocol

For each backbone, we evaluate three multimodal configurations.
Frozen pretrained model: The original pretrained vision-language model is evaluated without task-specific adaptation.
LoRA + answer loss: The pretrained backbone is adapted using LoRA Hu et al. (2022) and optimized only with the supervised answer loss.
VL-DyMS: The same backbone and LoRA adaptation configuration are used, VL-DyMS training strategy described in Section 3.

The text-only and vision-only teachers are also evaluated to characterize the predictive information available from each isolated modality. These masked-input models are used as diagnostic views and are not treated as competing multimodal methods.

Final evaluation is performed on the complete standard evaluation split used for each benchmark. No arbitrary evaluation subset is used for the reported results.

All trainable variants use LoRA with rank $r = 8$, $\alpha = 16$, and dropout 0.05, and are trained for one epoch with seed 42. Experiments use a single NVIDIA A100 with 4-bit NF4 quantization and mixed-precision computation.

Table 4: Multimodal performance across five datasets using Qwen2.5-VL and LLaVA-1.5.

Model \ Dataset	TextVQA	GQA	TallyQA	POPE	POPEv2
Qwen2.5-VL 3B					
Qwen2.5 Baseline	80.00	60.00	70.27	91.8	88.70
LoRA	75.57	61.21	72.50	92.1	80.2
VL-DyMS	**81.67**	**63.75**	**72.50**	**93.9**	**89.10**
LLaVA-1.5 7B					
LLava-1.5 Baseline	50.00	60.56	65.20	85.66	78.20
LoRA	50.41	53.18	66.5	84.10	79.6
VL-DyMS	**52.08**	**66.87**	**68.70**	**89.44**	**80.50**

5 Results and Discussion

5.1 Overall Performance

Tables 4 and 5 report the performance of VL-DyMS using Qwen2.5-VL-3B and LLaVA-1.5-7B, together with the corresponding text-only and vision-only teachers. For each backbone, we compare the frozen pretrained baseline, LoRA adaptation, and VL-DyMS using the standard metric of each benchmark.

Across both backbones, VL-DyMS achieves a higher point estimate than the pretrained baseline on all five datasets. With Qwen2.5-VL-3B, the performance improvement gain is observed on GQA (3.75 points), followed by TallyQA (2.23 points).

With LLaVA-1.5-7B, VL-DyMS the largest improvement is again observed on GQA (6.29 points), with gains of 3.50 and 3.78 points on TallyQA and POPE, respectively. Improvements across all ten backbone-benchmark combinations indicate that VL-DyMS is not restricted to a single pretrained model or task type.

Table 5: Uni-modal teachers performance across five datasets using Qwen2.5-VL and LLaVA-1.5.

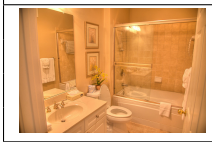
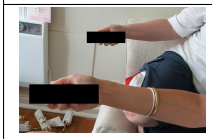
Model \ Dataset	TextVQA	GQA	TallyQA	POPE	POPEv2
Qwen2.5-VL 3B					
Text-Teacher	1.25	27.50	12.50	40.17	50.00
Vision-Teacher	6.67	6.25	22.50	88.50	92.10
LLaVA-1.5 7B					
Text-Teacher	7.50	37.84	36.60	50.00	50.00
Vision-Teacher	10.41	12.64	49.00	84.55	79.00

5.2 Cross-Models Analysis

The comparison between the pretrained baseline, standard LoRA adaptation, and VL-DyMS shows that dynamic modality scaling yields higher performance gains over the pretrained baseline than standard fixed adaptation across the same benchmarks. While LoRA applies a fixed answer-supervised adaptation strategy, VL-DyMS uses modality-specific distillation signals to dynamically regulate the relative influence of textual and visual supervision while preserving pretrained behavior.

The pattern across both backbones suggests that explicitly adapting modality-specific learning pressure to the current learning state provides additional benefit beyond standard parameter-efficient adaptation.

Table 6: Qualitative examples comparing the text-only teacher, vision-only teacher, pretrained baseline, and VL-DyMS outputs for Qwen2.5-VL and LLaVA-1.5 across the five evaluated datasets.

TextVQA				
Question: What is the brand of this camera? **Answer:** Dakota Digital	**Qwen2.5-VL 3B**		**LLaVA-1.5 7B**	
	**Text-Teacher**	I'm sorry, but I can't identify the brand of	**Text-Teacher**	Canon
	Vision-Teacher	Dakota Digital Single-Use Camera	**Vision-Teacher**	Digital
	Baseline	dakota	**Baseline**	Dakota digital
	VL-DyMS	dakota digital	**VL-DyMS**	dakota digital
GQA				
Question: What is in the basket? **Answer:** broccoli	**Qwen2.5-VL 3B**		**LLaVA-1.5 7B**	
	**Text-Teacher**	apples	**Text-Teacher**	Nothing
	Vision-Teacher	What is the price of broccoli?	**Vision-Teacher**	$2.25
	Baseline	broccoli	**Baseline**	broccoli
	VL-DyMS	broccoli	**VL-DyMS**	broccoli
TallyQA				
Question: How many players are in the picture? **Answer:** 1	**Qwen2.5-VL 3B**		**LLaVA-1.5 7B**	
	**Text-Teacher**	1	**Text-Teacher**	1
	Vision-Teacher	1	**Vision-Teacher**	1
	Baseline	1	**Baseline**	1
	VL-DyMS	1	**VL-DyMS**	1
POPE				
Question: Is there a potted plant in the image? **Answer:** Yes	**Qwen2.5-VL 3B**		**LLaVA-1.5 7B**	
	**Text-Teacher**	Yes	**Text-Teacher**	Yes
	Vision-Teacher	Yes	**Vision-Teacher**	Yes
	Baseline	Yes	**Baseline**	Yes
	VL-DyMS	Yes	**VL-DyMS**	Yes
POPEv2				
Question: Is there a remote in the image? **Answer:** No	**Qwen2.5-VL 3B**		**LLaVA-1.5 7B**	
	**Text-Teacher**	Yes	**Text-Teacher**	Yes
	Vision-Teacher	Yes	**Vision-Teacher**	Yes
	Baseline	No	**Baseline**	Yes
	VL-DyMS	No	**VL-DyMS**	No

5.3 Modality Analysis

The masked teachers indicate how much task-relevant information remains when one modality is removed. Table 6 further illustrates their behavior relative to the pretrained baseline and VL-DyMS.

On TextVQA, both masked views perform substantially below the multimodal baseline. For Qwen2.5-VL, the text- and vision-only teachers obtain 1.25 and 6.67 compared with 80.00 for the full model; for LLaVA-1.5, they obtain 7.50 and 10.41 compared with 50.00. This reflects the complementary nature of TextVQA, where the question identifies the required information and the image provides the scene text.

The modality pattern differs across tasks and backbones. On GQA, the text-only teacher outperforms the vision-only teacher for both models, including 37.84 versus 12.64 for LLaVA-1.5. On TallyQA, however, the LLaVA-1.5 vision-only teacher reaches 49.00 compared with 36.60 for the text-only teacher. On the hallucination benchmarks, vision-only teachers are relatively strong, particularly for Qwen2.5-VL, reaching 88.50 on POPE and 92.10 on POPEv2.

These masked-view scores do not directly measure causal modality contribution in the full multimodal model. A modality that performs poorly alone may still provide essential complementary information. VL-DyMS therefore does not select the stronger teacher or enforce equal modality contribution; instead, it dynamically adjusts separate modality-specific supervision while preserving the complete multimodal input.

5.4 Discussion

The results support three observations. First, VL-DyMS consistently improves the pretrained multimodal baseline across vision-text understanding, compositional reasoning, counting, and object-hallucination benchmarks using two VLM backbones. Second, the magnitude of improvement varies across tasks and models, supporting the motivation for adapting modality-specific supervision to the current learning state rather than relying on standard fixed adaptation. Third, the teacher results distinguish *modality availability* from *modality usefulness*: isolated predictive strength does not directly determine a modality's importance during multimodal inference. VL-DyMS therefore regulates modality-specific learning pressure dynamically while preserving full multimodal inference.

6 Conclusion

This paper addressed modality imbalance in LLM-centric vision-language models, where the relative contribution of text and vision may vary across tasks and throughout adaptation. Rather than assuming a fixed direction of imbalance, we argue that modality-specific supervision should respond dynamically to the current learning state of the model.

We proposed VL-DyMS, a dynamic modality scaling framework that uses frozen textual and visual teachers together with a pretrained multimodal baseline to regulate modality-specific supervision during adaptation. Across TextVQA, GQA, TallyQA, POPE, and POPEv2, VL-DyMS achieves higher point estimates than the pretrained baseline while introducing no additional inference-time component. These results provide initial evidence that dynamically controlling the relative learning influence of vision and language can improve downstream adaptation across tasks with different multimodal requirements.

Limitations and Future Work. The current VL-DyMS focuses on modality-specific contribution through modality distillation and task loss. While this methodology demonstrates it's effectiveness, other signals that may characterize cross-modal relationships have not yet been explored. Future work will investigate concept-aware dynamic scaling by incorporating semantic concepts into the controller state. We will also extend the evaluation to additional VLM families and multiple training seeds to determine the robustness and generality of the proposed approach.

AI use statement

AI were used only to improve the clarity, readability, and wording of the manuscript. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of the manuscript.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 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 (CVPR)*, 2024.

Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Weiming Zhang, and Nenghai Yu. Deciphering cross-modal alignment in large vision-language models via modality integration rate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 218–227, October 2025.

Drew A. Hudson and Christopher D. Manning. 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 (CVPR)*, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023a.

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

Yifan Li, Kun Zhou, Xin Zhao, Lei Fang, and Jirong Wen. Analyzing and mitigating object hallucination: A training bias perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6636–6643, 2026. doi: 10.1609/aaai.v40i8.37594.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Letitia Parcalabescu and Anette Frank. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models and tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8238–8247, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uAFHCZRmXk.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, Naila Murray, Marjan Ghazvininejad, Mike Lewis, Nicolas Ballas, Amir Bar, Michael Rabbat, Jakob Verbeek, Luke Zettlemoyer, Koustuv Sinha, Yann LeCun, and Saining Xie. Beyond language modeling: An exploration of multimodal pretraining. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Jun Wang, Fuyuan Cao, Zhixin Xue, Xingwang Zhao, and Jiye Liang. CMoB: Modality valuation via causal effect for balanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 91756–91778, 2025. doi: 10.52202/085713-2763.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, 2020.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2250–2259, October 2025. doi: 10.1109/ICCV51701.2025.00217.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52559–52572. PMLR, 2024.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Jingfan Yang, Rui Zhang, Liang Hong, and Ke Yuan. SCOPE: Preserving modality-specific cues to mitigate modality laziness in multimodal learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29066–29078, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1453. URL https://aclanthology.org/2026.findings-acl.1453/.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62108–62122, 2024.

Jinghao Zhang, Guofan Liu, Qiang Liu, Shu Wu, and Liang Wang. Modality-balanced learning for multimedia recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.