iGVLM: Dynamic Instruction-Guided Vision Encoding for Question-Aware Multimodal Understanding

Anonymous ACL submission

Abstract

Most existing Vision–Language Models rely on *static, instruction-agnostic* vision encoders whose representations are reused across arbitrarily different textual queries, limiting fine-grained reasoning where task-specific visual cues matter. We argue that effective instruction conditioning requires an *architectural* separation between pretrained and instruction-conditioned pathways, rather than relying on in-place injection alone. We propose **iGVLM**, an instruction-guided vision encoder built on this principle: a frozen static branch preserves pretrained visual priors, while a dynamic branch modulates visual features via Adaptive Layer Normalization under a CLIP-encoded instruction embedding; the two are fused through a zero-initialized projection. To evaluate instruction sensitivity directly, we further introduce **MM4**, a diagnostic probe of 180 images and 720 QA pairs that measures per-image consistency across four semantically distinct queries on the same image. Across three LLM backbones (Vicuna-7B/13B, Qwen2.5-3B) and six benchmarks under a controlled setting, iGVLM improves MMStar by +4.4/+3.6/+4.5 over LLaVA-1.5 and degrades more slowly than baselines on MM4 under strict consistency thresholds, while retaining ∼80% throughput.

1 Introduction

Vision–Language Models (VLMs) (Chen et al., 2024a; Lu et al., 2024; Chen et al., 2023; Jiang et al., 2024), built on rapid progress in computer vision (Zhang et al., 2024; Liu et al., 2021) and natural language processing (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020), have achieved strong performance on multimodal tasks such as image captioning, visual question answering, and grounded dialogue. Progress on these tasks has so far been driven largely by scaling data, model size, and training compute. In this work, we step back from this scaling trajectory and ask a *methodological* question that remains underexplored: *what is the right architectural principle for conditioning visual perception on linguistic instructions, without sacrificing the perceptual priors learned during large-scale pre-training?*

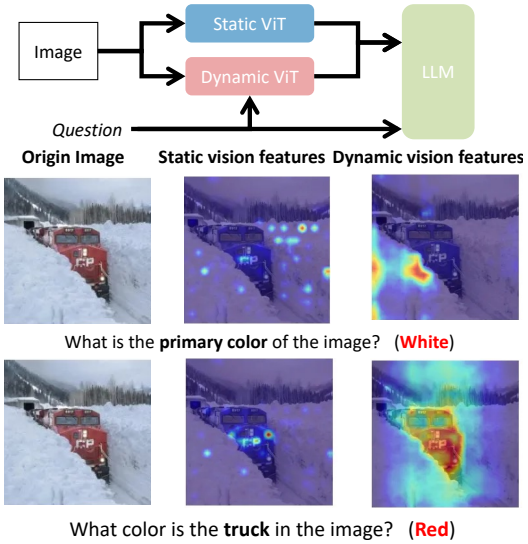


Figure 1: **Visualization of vision features.** We employ Grad-CAM to visualize the vision encoders of the two branches in iGVLM, highlighting the regions most relevant to the correct answer. The instruction-guided branch distinctly focuses on areas closely associated with the correct answer, whereas the static branch attends to a question-agnostic salient region.

Most existing VLMs rely on *static, instruction-agnostic* vision encoders such as CLIP-ViT (Radford et al., 2021), which extract visual representations independently of the textual query. The same visual features are therefore reused across arbitrarily different instructions, leaving the model unable to emphasize task-relevant cues for fine-grained reasoning. Figure 1 illustrates this concretely: static representations attend to the same salient region regardless of whether the question asks about the dominant color of the scene or the color of a specific object. The central difficulty, then, is not relearning visual perception, but *condi-*

tioning the utilization of pretrained visual features on linguistic instructions—and doing so without destabilizing the priors that make those features useful in the first place.

Recent work approaches this problem from two directions, each with a structural cost. QA-ViT (Ganz et al., 2024) injects textual representations into the upper layers of a single vision transformer; this is efficient but mixes instruction signals directly into the pretrained representation pathway, which we find yields only marginal gains on instruction-sensitive dimensions of MMStar (e.g., +1.6 on Instance Reasoning over LLaVA-1.5). DyFo (Li et al., 2025) instead formulates visual reasoning as Monte Carlo Tree Search over expert-model decisions, achieving stronger instruction adaptation at the cost of more than $20\times$ throughput degradation and reliance on external experts. These two regimes—weak in-place modulation versus expensive external search—expose a common gap: there is currently no lightweight mechanism that conditions visual features on instructions *while explicitly protecting the pretrained representation from being overwritten by the conditioning signal itself*.

We argue that effective instruction conditioning calls for an *architectural* separation between pretrained and instruction-conditioned pathways, rather than relying solely on training-time regularization or in-place injection. To this end, we propose **iGVLM**, an instruction-guided vision encoder whose central design principle is to *decouple representation preservation from instruction-conditioned adaptation* into two parallel pathways (shown in Figure 2). A frozen *static branch* retains the original CLIP-ViT features as a task-agnostic anchor. A parallel *dynamic branch* applies Adaptive Layer Normalization (AdaLN) (Perez et al., 2018) at every transformer layer, modulating attention and MLP activations under a CLIP-encoded instruction embedding. The two branches are fused through a zero-initialized projection, so training begins from the exact behavior of the frozen backbone and introduces instruction-conditioned adaptation only as it proves useful. This decoupling, rather than any single component in isolation, is what makes instruction-aware vision encoding stable: ablating the static branch alone causes MMStar to collapse from 36.4 to 27.3 and MM4 from 23 to 5 on the 13B model.

To evaluate not just average accuracy but *instruction sensitivity itself*, we further introduce **MM4**, a diagnostic benchmark in which each image is paired with four semantically distinct questions whose answers are uniformly distributed across options. Existing benchmarks score each query in isolation and therefore cannot distinguish a model that genuinely re-attends to the image per instruction from one that exploits a single dominant visual feature. MM4 is intended not as a leaderboard but as a controlled probe: a tool for verifying whether a proposed design principle actually produces the instruction-conditioned behavior it claims.

Our goal in this paper is not to pursue leaderboard performance through scale, but to isolate and validate a design principle for instruction-aware vision encoding. Under a controlled setting where all compared methods share the same training data, vision backbone, and language model, we ask whether *decoupling representation from modulation* consistently outperforms in-place conditioning—and find that it does, across three LLM backbones and six benchmarks. Our main contributions are:

- We identify *architectural decoupling* of representation preservation and instruction-conditioned modulation as a design principle for instruction-aware vision encoding, and instantiate it as **iGVLM**: a frozen static branch and a dynamic AdaLN-modulated branch fused through a zero-initialized projection.
- We introduce **MM4**, a diagnostic benchmark of 180 images and 720 manually verified QA pairs designed to probe multi-query instruction sensitivity under a shared visual input, complementing existing benchmarks that evaluate queries in isolation.
- Through controlled experiments across three LLM backbones (Vicuna-7B/13B, Qwen2.5-3B) and six benchmarks, we show that the proposed decoupling consistently improves instruction sensitivity over both in-place modulation (QA-ViT) and search-based methods (DyFo) under identical training data and backbones, improving MMStar by +4.4/+3.6/+4.5 over LLaVA-1.5 while retaining $\sim$80% throughput.

2 Related Work

Vision encoders in VLMs. A foundational line of work, exemplified by CLIP (Radford et al., 2021), aligns visual and textual representations via

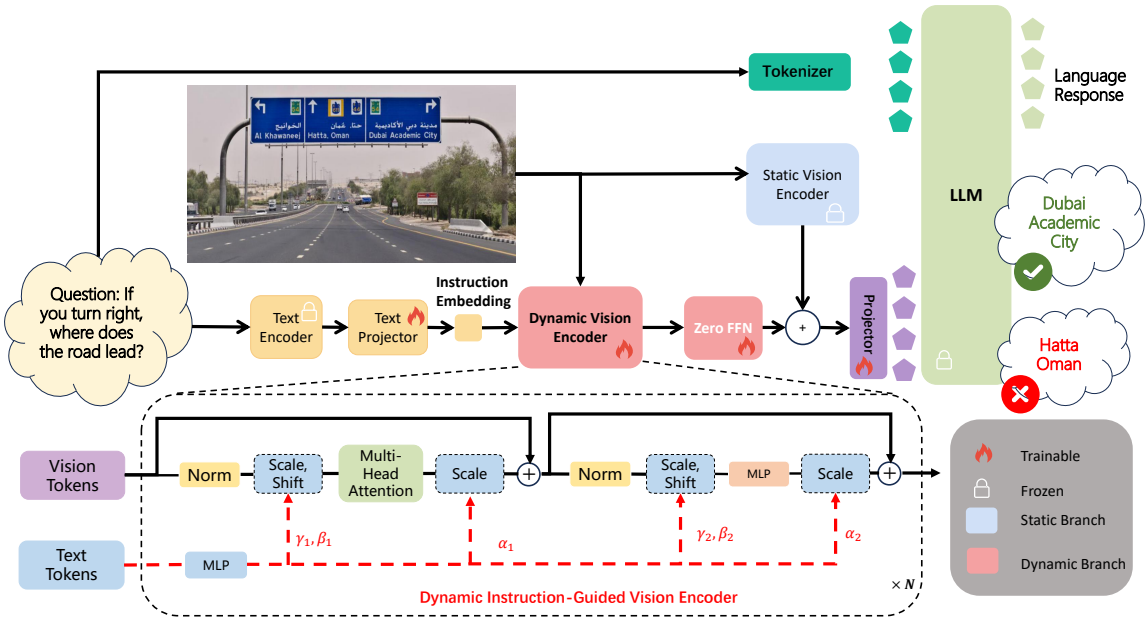


Figure 2: **(Top): The proposed iGVLM architecture.** The Text Encoder extracts features from the input instructions to guide the Vision Encoder, enabling dynamic modulation of visual representations. These instruction-conditioned features are then fused with static visual features. The fused representation is aligned with a Large Language Model (LLM) to generate responses. The illustrated example comes from a real-world VQA scenario rather than the MM4 benchmark. **(Bottom): AdaLN-Modified ViT.** We leverage textual information to modulate the multi-head attention and MLP modules within the ViT via the AdaLN adapter, enabling instruction-aware adjustment of visual attention.

large-scale contrastive learning, forming the basis of many modern VLMs (Chen et al., 2024a; Lu et al., 2024; Chen et al., 2023). To strengthen this static visual representation, prior work fuses multiple encoders (Ranzinger et al., 2024; Tong et al., 2024) or fine-tunes them for high resolution (Li et al., 2024), but the resulting encoders remain instruction-agnostic. A complementary line introduces query-awareness directly into the encoder: QA-ViT (Ganz et al., 2024) uses query-aware cross-attention within a single pathway, providing some instruction adaptation but with limited control over how pretrained representations are preserved. Our work explicitly decouples representation preservation from instruction-conditioned modulation via a dual-branch design.

VLM evaluation. Early benchmarks (Goyal et al., 2017; Sharma et al., 2018; Schwenk et al., 2022) provide task-specific assessments; recent efforts (Wu et al., 2024; Fu et al., 2023; Cheng et al., 2023; Chen et al., 2024b) offer broader and more challenging evaluations. However, most existing benchmarks evaluate queries in isolation and cannot directly measure whether a model adapts its visual perception across different instructions on the same image—a gap MM4 is designed to fill.

3 Method

3.1 Overall Architecture

As illustrated in Figure 2, given an image-text pair, iGVLM derives a global instruction embedding from a CLIP text encoder and uses it to condition a parallel dual-branch vision encoder. A frozen static branch retains task-agnostic visual priors; a dynamic branch applies AdaLN at every transformer layer under the instruction embedding. The two branches are fused via a zero-initialized projection, and the result is projected into the LLM input space together with the instruction tokens.

The outputs of these two branches are fused to obtain a balanced visual representation that combines general-purpose perceptual semantics with task-specific adaptation. The fused visual features are subsequently projected into the language embedding space and provided, together with the instruction tokens, to a large language model (LLM) for multimodal reasoning and response generation.

3.2 Instruction-Guided Visual Feature Modulation

To enable instruction-conditioned visual perception, we derive a global textual guidance signal from the instruction. We adopt the text encoder

from a pretrained CLIP model and truncate the input text to a maximum length of 77 tokens. The resulting [CLS] token embedding summarizes the semantic intent of the instruction and is mapped into the vision latent space through a lightweight linear projection:

$$c_t = \mathcal{F}_T(T_{\leq 77}), \quad \hat{c}_t = \mathcal{H}_t(\mathrm{Norm}(c_t)), \tag{1}$$

where $\mathcal{F}_T(\cdot)$ denotes the CLIP text encoder, and $\mathcal{H}_t(\cdot)$ aligns the text embedding with the vision feature space.

We incorporate Adaptive Layer Normalization (AdaLN) (Perez et al., 2018) into each transformer block of the CLIP vision encoder to inject textual conditioning in a stable and localized manner. The projected instruction embedding $\hat{c}_t$ is transformed into layer-wise modulation parameters that control feature scaling and shifting within both the self-attention and feedforward submodules. By integrating AdaLN across all transformer layers, iGVLM enables hierarchical instruction-conditioned modulation while preserving the pretrained weights of the vision backbone.

Formally, given an input image I and instruction embedding $\hat{c}_t$, the instruction-guided vision encoder produces:

$$y_{ct} = \mathcal{F}_{ct}(I, \hat{c}_t; \Theta_{\text{CLIP}}), \tag{2}$$

where $y_{ct} \in \mathbb{R}^{N_I \times D_I}$ denotes the instruction-conditioned visual features and Θ_{CLIP} represents the frozen pretrained parameters.

3.3 Dual-Branch Feature Fusion

While instruction-conditioned modulation enables task-specific adaptation, preserving the original perceptual semantics is essential for stable and generalizable visual understanding. To this end, iGVLM employs a dual-branch fusion mechanism that explicitly combines instruction-guided features with the original frozen visual representations.

Let $y_{ct} \in \mathbb{R}^{N_I \times D_I}$ denote the instruction-guided features from $\mathcal{F}_{ct}$, and let $y_0 = \mathcal{F}_I(I; \Theta_{\text{CLIP}})$ denote the corresponding frozen features from the original vision encoder. The fused visual representation is computed as:

$$y_I = \mathcal{Z}(\mathrm{Norm}(y_{ct})) + y_0, \tag{3}$$

where $\mathcal{Z}$ is a learnable linear projection initialized to zero. This initialization ensures that the fused representation initially matches the pretrained visual features, allowing instruction-conditioned adaptation to be introduced gradually and safely during training.

Following the LLaVA-1.5 framework, the fused visual features y_I are projected into the input embedding space of the LLM via a learnable linear transformation. Training proceeds in two stages: first, the instruction-guided vision encoder and projection layers are optimized while keeping the pretrained vision backbone and LLM frozen; second, all components are jointly optimized to enable coherent multimodal reasoning.

3.4 MM4: A Diagnostic Probe for Question-Aware Visual Perception

Motivation. Existing multimodal benchmarks (Chen et al., 2024b) evaluate each (image, question) pair in isolation, which cannot directly answer the question we care about: *does a model re-attend to the same image differently across different questions, or does it exploit a single dominant visual feature regardless of the query?* Such benchmarks also tend to pose questions about the most salient object in each image, so a model that always answers about the salient object correctly can score well while failing this notion of instruction sensitivity. We therefore design MM4 not as a leaderboard but as a controlled diagnostic probe, reusing the same image across multiple semantically distinct queries so that per-image consistency can be measured directly.

Construction protocol. MM4 contains 180 images and 720 multiple-choice QA pairs (four per image), manually written by three annotators with graduate-level training in computer science; we deliberately avoided using LLMs for annotation to prevent circular evaluation bias. Construction proceeded in two stages: we first collected 150 images and 600 QA pairs from publicly available web sources under permissive use, then classified each question under the six MMStar dimensions (CP, FP, IR, LR, ST, MA) and added 30 images / 120 QA pairs targeting ST and MA, which were initially under-represented (<16% of the pool). All images are used under permissive licenses; per-image URLs, attribution, and the question annotations will be released upon acceptance. No personally identifiable information appears in any image.

Question design constraints. For each image, the four questions are written under three operational constraints: **(i) Option-order invariance**—Q_1 and Q_2 ask the same question with the correct

answer at different option positions, distinguishing models that ground in the image from those exploiting positional bias; **(ii) Multi-perspective diversification**—Q_3 and Q_4 probe visual aspects semantically orthogonal to Q_1/Q_2 and to each other, so a model relying on a single global feature cannot answer all four; and **(iii) Balanced answer distribution**—the correct option is uniformly distributed over $\{A, B, C, D\}$ (180 each). We verify (i)–(ii) via pairwise CLIP text similarities (Appendix D).

Scoring protocol. For each image, a model answers all four questions and is awarded an *image-level score* of 1 if it answers at least n of them correctly, 0 otherwise. We report results at all four thresholds $n \in \{1, 2, 3, 4\}$, with larger n corresponding to stricter consistency; the reported number is the count of images (out of 180) satisfying the threshold. Random guessing yields $\sim$123, 47, 9, and 0.7 images at $n = 1, 2, 3, 4$, making $n = 3, 4$ the most discriminative thresholds. MM4 complements rather than replaces large-scale benchmarks: it measures per-image instruction sensitivity, a property those benchmarks structurally cannot.

4 Experiments

4.1 Experimental Settings

Training Setup. All experiments use the LLaVA-1.5 training framework with the same open-source data (558K alignment + 665K instruction-tuning samples) and the same **CLIP-Large-336** (Radford et al., 2021) vision backbone, so methods differ only in how visual features are conditioned on text. We evaluate iGVLM with **Vicuna-7B/13B** (Chiang et al., 2023) and **Qwen2.5-3B** (Yang et al., 2024) backbones on 8 NVIDIA A100 GPUs. Training a 7B iGVLM incurs $\sim$**1.1**$\times$ GPU hours over LLaVA-1.5, reflecting the lightweight nature of the proposed modulation.

Evaluation Benchmarks. Our primary evaluation is on MMStar (Chen et al., 2024b), a vision-dependent benchmark designed to minimize data leakage. We additionally report VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), POPE (Li et al., 2023), VizWiz (Gurari et al., 2018), and ScienceQA (Lu et al., 2022) to assess open-ended understanding, hallucination robustness, and scientific reasoning.

Baselines. We adopt **LLaVA-1.5** (Liu et al., 2024) as the primary baseline (open data, weights, and code). All baselines are trained under identical data, optimization, and vision backbone as iGVLM, so performance differences arise solely from visual encoding strategies. On Vicuna backbones we additionally compare with **QA-ViT** (Ganz et al., 2024) (query-aware modulation) and **DyFo** (Li et al., 2025) (expert-guided search); QA-ViT and DyFo are not evaluated on Qwen2.5-3B due to architectural compatibility.

4.2 Results on MMStar

We evaluate iGVLM on MMStar under three representative settings. For Vicuna-based models (7B and 13B), we compare iGVLM with both the static baseline LLaVA-1.5 and two instruction-aware modulation methods, QA-ViT and DyFo. For the Qwen2.5-3B backbone, we compare against the corresponding LLaVA-1.5 model due to architectural compatibility. All methods are evaluated using identical training data, vision backbones, and inference configurations to ensure a controlled comparison.

As shown in Table 1, iGVLM achieves the best overall performance across all backbones. On Vicuna-7B, iGVLM improves the average MMStar score by **+4.4** points over LLaVA-1.5, outperforming both QA-ViT (+1.2) and DyFo (+2.7). Notably, the gains are most pronounced on instruction-sensitive and fine-grained reasoning dimensions, including Instance Reasoning (IR), Logical Reasoning (LR), and Science & Technology (ST), where iGVLM consistently surpasses both baselines. A similar trend is observed on Vicuna-13B: iGVLM achieves an average improvement of **+3.6** points, exceeding QA-ViT (+3.1) and DyFo (+1.8), with particularly strong gains on Fine-grained Perception (FP) and ST. These results indicate that explicitly conditioning the utilization of visual features enables more effective instruction-aware reasoning than single-path modulation (QA-ViT) or expert-guided search (DyFo).

In addition to accuracy, iGVLM maintains a favorable efficiency profile. Despite achieving higher overall performance, iGVLM preserves throughput comparable to LLaVA-1.5, with only a modest reduction (13.5$\rightarrow$11.1 for Vicuna-7B and 9.8$\rightarrow$8.6 for Vicuna-13B). In contrast, DyFo incurs a severe efficiency penalty due to repeated expert-guided search, reducing throughput by more than **20**$\times$ (13.5$\rightarrow$0.49 for Vicuna-7B and 9.8$\rightarrow$0.47 for Vicuna-13B). QA-ViT maintains efficiency similar to LLaVA-1.5, but achieves only limited accuracy

Method	Venue	CP	FP	IR	LR	ST	MA	Avg.	Δ	Thpt
Vicuna-7B Backbone										
LLaVA-1.5	CVPR'24	58.8	24.0	38.8	24.0	13.6	22.8	30.3	–	13.5
QA-ViT	CVPR'24	**60.0**	28.0	40.4	25.6	14.0	21.2	31.5	+1.2	12.9
DyFo	CVPR'25	54.8	**28.4**	38.0	25.6	**24.0**	**27.2**	33.0	+2.7	0.49
iGVLM	Ours	59.6	28.0	**41.2**	**32.8**	19.6	26.8	**34.7**	**+4.4**	11.1
Vicuna-13B Backbone										
LLaVA-1.5	CVPR'24	**58.8**	28.0	41.6	24.4	18.4	25.6	32.8	–	9.8
QA-ViT	CVPR'24	57.6	32.8	43.2	**28.8**	21.2	**31.6**	35.9	+3.1	10.2
DyFo	CVPR'25	55.6	32.0	**44.4**	27.2	22.0	26.4	34.6	+1.8	0.47
iGVLM	Ours	57.6	**37.6**	43.2	27.2	**22.4**	30.8	**36.4**	**+3.6**	8.6
Qwen2.5-3B Backbone										
LLaVA-1.5	CVPR'24	8.4	8.4	16.4	24.4	16.8	26.4	16.8	–	12.6
iGVLM	Ours	**20.0**	**13.2**	**19.2**	**28.4**	**21.6**	25.6	**21.3**	**+4.5**	10.9

Table 1: Comparison among dynamic visual perception methods on **MMStar**, evaluated across six perception dimensions (CP, FP, IR, LR, ST, MA), average accuracy (Avg.), and throughput (it/s). Our **iGVLM** consistently outperforms prior approaches under both **Vicuna** and **Qwen2.5** backbones, achieving the best overall balance between accuracy and efficiency. Δ indicates improvement over LLaVA-1.5 under the same backbone. **Bold** and underlined values indicate the best and second-best results, respectively.

gains, highlighting the trade-off between conditioning strength and computational cost in existing approaches.

Results on Qwen2.5-3B further confirm the generality of iGVLM. Compared to LLaVA-1.5, iGVLM improves the average MMStar score from 16.8 to **21.3** (+4.5), with consistent gains across all capability dimensions. Taken together, iGVLM achieves a better accuracy-efficiency balance than prior modulation methods under controlled comparison.

4.3 Results on MM4

We evaluate iGVLM on MM4, which measures whether a model adapts its visual perception consistently across four semantically distinct questions grounded in the same image. We report results at all four thresholds $n \in \{1, 2, 3, 4\}$, where larger n corresponds to stricter consistency.

Reference Points: Closed-Source Systems. For context, Table 2 includes GPT-4o (OpenAI, 2024), GPT-4V, and Qwen2.5-VL-max (Qwen Team, 2025) as reference points rather than controlled comparisons, since they differ from our models in training data, framework, and scale. They confirm that MM4 is not saturated by current closed-source systems—even GPT-4o reaches only 58/180 images at $n = 4$.

Controlled Comparisons against LLaVA-1.5. The controlled comparisons in this work are between iGVLM and LLaVA-1.5 under identical training data, vision backbones, and language models. At the most discriminative threshold $n = 4$, iGVLM improves over LLaVA-1.5 from 13→17 (7B), 18→23 (13B), and 27→29 (Qwen2.5-3B), with similar gaps at $n = 3$ (e.g., 18→23 for 13B). Because random guessing yields only ∼0.7 images at $n = 4$ and ∼9 at $n = 3$, these improvements lie well outside MM4's noise floor at the strictest thresholds.

Multi-Query Consistency. For all models, the per-image score decreases monotonically as n increases (Table 2); we are interested in the *rate* of this decay, as a model that genuinely re-attends per query should degrade more slowly than one driven by a single static feature. Under matched backbones, iGVLM retains a larger fraction of its single-question accuracy as the consistency requirement tightens: the n=4 to n=1 ratio is 14.3% vs. 11.2% for the 13B pair and 10.6% vs. 8.3% for the 7B pair—precisely the property MM4 is designed to measure.

Cross-Backbone Observation. iGVLM-3B (Qwen2.5) achieves a higher $n = 4$ score than iGVLM-13B (Vicuna) (29 vs. 23). We interpret this as a backbone-family effect rather than an effect of our method: the LLaVA-1.5-3B baseline itself already exceeds LLaVA-1.5-13B at $n = 4$ (27 vs. 18), reflecting the stronger instruction-following of Qwen2.5 over Vicuna. The methodologically relevant comparison is therefore *within* each backbone family, where

Figure 3: **Representative Examples from MM4.** Deceptively simple questions demand multiperspective reasoning. Correct answers (highlighted) are uniformly distributed across options to prevent positional bias. More examples can be found in Appendix E.

iGVLM consistently improves over its LLaVA-1.5 counterpart, compatible with the within-family scaling in Appendix 7.

Qualitative Analysis. Figure 3 illustrates how iGVLM-13B more accurately localizes instruction-relevant regions across queries. In the science-diagram case (water cycle), iGVLM-13B correctly distinguishes visually similar stages (*precipitation*, *transpiration*, *surface runoff*) across all four questions, while LLaVA-1.5-13B and QA-ViT-13B answer the first correctly but fail to re-locate the relevant region as the query shifts. In the food-scene case, iGVLM-13B correctly identifies both the spatial location of the green leaves and the missing ingredient—queries requiring different image regions—while the baselines repeat one answer across queries that demand different regions.

4.4 Other Benchmarks

To verify that instruction-guided visual modulation does not compromise general-purpose multimodal capability, we further evaluate **iGVLM** on VQAv2, GQA, POPE, VizWiz, and ScienceQA-IMG (Table 3), covering open-ended reasoning, hallucination robustness, real-world grounding, and scientific QA.

Across all three backbones, iGVLM matches or improves over LLaVA-1.5 on every benchmark, with no systematic degradation. The clearest gains arise on benchmarks that require selective visual attention: VizWiz (+2.5 / +1.7 / +2.7 over the 7B / 13B / 3B baselines, respectively), where real-world images often contain irrelevant visual content, and POPE (+0.5 / +0.7 on 7B / 13B), where hallucination errors typically stem from over-reliance on a single dominant feature. ScienceQA-IMG also improves on Vicuna-7B (+3.1) and Qwen2.5-3B (+0.8). Open-ended VQAv2 and GQA, which depend less on instruction-conditioned attention, remain comparable. DyFo is omitted from VQAv2, GQA, and VizWiz because its MCTS-based search is designed for selective-answer settings and does not directly apply to these tasks.

Overall, iGVLM behaves as a drop-in enhancement: it improves on benchmarks where instruction conditioning should help, and does no harm on those where it should not.

n	GPT-4o	Qwen2.5-vl-max	GPT-4v	iGVLM-13B	QA-ViT-13B	LLaVA1.5-13B	iGVLM-7B	LLaVA1.5-7B	QA-ViT-7B	LLaVA-1.5-3B	iGVLM-3B
1	170	**172**	161	161	163	161	161	157	162	165	164
2	138	**149**	120	117	115	113	109	113	108	130	124
3	107	**114**	66	71	68	77	58	61	59	82	78
4	**58**	52	24	23	21	18	17	13	11	27	29

Table 2: **Results on MM4.** Each cell reports the number of images (out of 180) for which the model answered at least n of the image's four questions correctly. Larger n corresponds to stricter instruction-consistency requirements. Random guessing yields $\sim$123, 47, 9, 0.7 at $n = 1, 2, 3, 4$.

Method	LLM	VQAv2	GQA	POPE	VisWiz	SQA
LLaVA-1.5	Vicuna-7B	78.5	62.0	85.4	50.0	66.8
QA-ViT	Vicuna-7B	79.0	62.2	85.7	51.3	68.1
DyFo	Vicuna-7B	-	-	84.8	-	65.5
iGVLM	Vicuna-7B	**79.1**	**62.8**	**85.9**	**52.5**	**69.9**
LLaVA-1.5	Vicuna-13B	80.0	63.3	85.4	53.6	71.6
QA-ViT	Vicuna-13B	79.9	63.2	85.8	**56.0**	71.3
DyFo	Vicuna-13B	-	-	84.6	-	71.6
iGVLM	Vicuna-13B	**80.2**	**63.3**	**86.1**	55.3	71.5
LLaVA-1.5	Qwen2.5-3B	77.7	62.4	84.5	50.7	72.2
iGVLM	Qwen2.5-3B	77.2	61.7	**84.6**	**53.4**	**73.0**

Table 3: Comparison with existing VLMs on other benchmarks including VQAv2, GQA, POPE, VisWiz, and SQA. Our **iGVLM** consistently achieves higher accuracy across all backbones, confirming its strong generalization beyond MMStar and MM4.

Method	Params(B)	MM4	MMStar	VQAv2	Viswiz
LLaVA1.5-13B	13.35	19	32.8	80.0	53.6
iGVLM-13B	13.78	**23**	**36.4**	**80.2**	**55.3**
-w/o AdaLN	13.78	22	35.1	80.2	53.5
-w/o FFN	13.78	17	34.1	80.1	54.7
-w/o Pure	13.48	5	27.3	60.2	37.9

Table 4: Component ablation of **iGVLM**. We analyze the effect of removing key modules including AdaLN, Zero-FFN, and the static branch. Performance consistently drops, confirming that instruction-guided modulation, Zero-FFN fusion, and dual-branch structure are all critical for iGVLM's effectiveness.

4.5 Ablation Study

Effect of Architectural Components. We ablate the three key components of **iGVLM**: AdaLN, the Zero-FFN fusion adapter, and the static branch. As shown in Table 4, each removal degrades performance on both MMStar and MM4: **w/o AdaLN** loses 1.3 MMStar points, confirming that layer-wise instruction conditioning matters; **w/o FFN** drops further, showing that gradual fusion is needed to avoid disrupting pretrained features; and **w/o Pure** (removing the static branch) causes the largest collapse (MMStar $36.4 \rightarrow 27.3$, MM4 $23 \rightarrow 5$), confirming that the dual-branch decoupling is the primary source of stability.

Comparison with Alternative Modulation Designs. We compare iGVLM against two variants

Method	MMStar	VQAv2	GQA	VisWiz	POPE
iGVLM	34.7	79.1	62.8	69.9	85.9
iGVLM-MoF	32.0	79.3	62.7	68.8	85.4
iGVLM-Cross	33.0	79.4	62.8	68.6	85.8

Table 5: Comparison of **iGVLM** variants on multiple benchmarks. **iGVLM-MoF** and **iGVLM-Cross** replace AdaLN with interleaved or cross-attention modulation, but both yield lower accuracy, highlighting the efficiency of our AdaLN-based design.

in Table 5: **iGVLM-MoF** (Tong et al., 2024) interleaves static and dynamic tokens, and **iGVLM-Cross** (Peebles and Xie, 2023) replaces AdaLN with cross-attention. Both underperform iGVLM—MoF weakens the static/dynamic separation, while cross-attention adds computational overhead without improving instruction consistency—indicating that AdaLN-based modulation offers a more effective and efficient conditioning mechanism.

We additionally analyze scaling behavior across four model sizes (1.5B–13B) in Appendix C.

5 Conclusion

We presented **iGVLM**, an instruction-guided vision encoder built on a single design principle: architecturally decoupling representation preservation from instruction-conditioned modulation via a frozen static branch and a dynamic AdaLN-modulated branch fused through a zero-initialized projection. Across LLM backbones from 1.5B to 13B (with 1.5B detailed in Appendix C) and six benchmarks under a controlled setting, iGVLM consistently improves instruction sensitivity over both in-place modulation and search-based methods while preserving general-purpose performance and $\sim$80% throughput of LLaVA-1.5. We additionally introduced **MM4**, a diagnostic probe of per-image multi-query instruction sensitivity that complements existing isolated-query benchmarks. We hope the decoupling principle and the diagnostic probe inform future design of instruction-aware multimodal models.

Limitations

Single training framework. All experiments in this paper are conducted within the LLaVA-1.5 training framework, sharing its data composition, two-stage training recipe, and connector design. This was a deliberate choice: our goal is to isolate the effect of instruction-guided visual encoding under a controlled setting, and varying the framework simultaneously with the encoder would confound the comparison against QA-ViT, DyFo, and LLaVA-1.5 itself. As a consequence, the gains we report should be interpreted as evidence for the *decoupling design principle* under one well-studied framework, rather than as a general claim about iGVLM's behavior under all modern VLM training recipes. Extending iGVLM to more recent frameworks with native dynamic-resolution encoders (e.g., LLaVA-NeXT, Qwen2-VL) is an immediate next step, and we expect the dual-branch design to remain compatible since it does not alter the underlying patch-token interface.

Limited model scale. Due to compute constraints, our experiments span LLM backbones from 1.5B to 13B parameters (Qwen2.5-1.5B/3B, Vicuna-7B/13B). We do not evaluate iGVLM at the 30B+ scale, where instruction-following capability tends to improve sharply and the relative contribution of explicit visual instruction conditioning may diminish. The monotonic improvement observed across our four scales (Appendix Table 7) suggests the proposed mechanism remains beneficial as backbones grow, but we cannot extrapolate this trend beyond 13B with confidence. Validation at larger scales is left to future work.

Scale of the MM4 benchmark. MM4 contains 180 images and 720 QA pairs, which is substantially smaller than benchmarks such as MMStar or MMBench. This is by design: MM4 is intended as a controlled diagnostic probe rather than a leaderboard, and its per-image multi-query construction requires manual writing and verification of four mutually constrained questions per image, which limits its practical scale. The current size is sufficient for the comparisons reported in this paper—random guessing yields only $\sim$0.7 images at the strictest threshold ($n = 4$), and the gaps between methods at $n = 3, 4$ are well outside this noise floor—but it is not large enough to support fine-grained subgroup analysis (e.g., per-category breakdowns at the strictest thresholds, or correlation studies between model properties and instruction sensitivity). We plan to expand MM4 in a future release while preserving its construction protocol.

Dataset licensing. The MM4 images were collected from publicly available web sources under permissive use, but we did not perform exhaustive per-image license verification at collection time. Prior to public release, we will conduct a full license audit, remove any images whose redistribution status is unclear, and publish per-image URLs together with attribution information. We will additionally provide a takedown mechanism for any rights holder who wishes to have their image removed from MM4.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, volume 15075, pages 370–387.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? In *NeurIPS*.

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. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2023. Can vision-language models think from a first-person perspective? *CoRR*, abs/2311.15596.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben-Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In *CVPR*, pages 13861–13871.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617.

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

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. 2025. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *CVPR*, pages 9098–9108.

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

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pages 26763–26773.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*, pages 26286–26296.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, abs/2403.05525.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

OpenAI. 2024. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *CVPR*, pages 4195–4205.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *AAAI*, pages 3942–3951.

Alibaba Group Qwen Team. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

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. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763.

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

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. 2024. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, pages 12490–12500.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, volume 13668, pages 146–162.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*.

Table 6: Hyperparameters of iGVLM are the same as the original LLaVA-1.5, except the learning rate in pre-training.

Hyperparameter	Pretrain	Finetune
Batch size	256	128
Lr	6e-4	2e-5
Lr schedule	cosine decay	
Lr warmup ratio	0.03	
Weight decay	0	
Epoch	1	
Optimizer	AdamW	
DeepSpeed stage	2	3

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Shuoxi Zhang, Hanpeng Liu, Stephen Lin, and Kun He. 2024. You only need less attention at each stage in vision transformers. In *CVPR*, pages 6057–6066.

A Hyperparameters

Our iGVLM follows the hyperparameter configuration of LLaVA-1.5. Since iGVLM employs a dual-branch vision encoder, we set the learning rate to 6e-4 during the pre-training stage. The hyperparameters for the first-stage vision–language alignment pre-training and the second-stage instruction tuning are provided in Table 6.

B MM4 Image Sources and License Information

A complete list of per-image source URLs and attribution information will be released together with the benchmark upon paper acceptance. At collection time, we restricted image selection to publicly accessible web sources whose terms of use permit research redistribution.

C Scaling Behavior

Table 7 reports iGVLM across four scales. Within each backbone family, MMStar improves monotonically (Qwen2.5: 1.5B $\rightarrow$ 3B: 19.7 $\rightarrow$ 21.3; Vicuna: 7B $\rightarrow$ 13B: 34.7 $\rightarrow$ 36.4); the smallest iGVLM-1.5B already exceeds its LLaVA-1.5B counterpart (19.7 vs. 17.1), indicating that the proposed modulation is beneficial across scales. The lower MM4 score of iGVLM-1.5B (16 vs. 19) suggests, however, that consistent multi-query reasoning requires sufficient language capacity. As noted in Section 4.3, the higher MM4 score of iGVLM-3B over iGVLM-13B reflects Qwen2.5's stronger instruction-following over Vicuna, not a scaling reversal of our method.

Model	LLM	MMStar (Avg.)	MM4 (Score)
LLaVA-1.5B	Qwen2.5-1.5B	17.1	19
iGVLM-1.5B	Qwen2.5-1.5B	19.7	16
iGVLM-3B	Qwen2.5-3B	21.3	29
iGVLM-7B	Vicuna-7B	34.7	17
iGVLM-13B	Vicuna-13B	36.4	23

Table 7: Scaling performance of iGVLM on MMStar and MM4 benchmarks under different model sizes and LLM backbones.

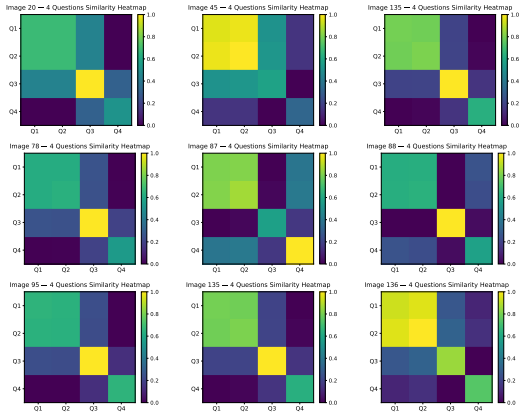


Figure 4: Different questions similarity heatmap.

D Analyzing Question Diversity via CLIP Text Embeddings

During the construction of the MM4 benchmark, we enforce strict criteria to ensure sufficient distinction between question pairs (Q_1, Q_2) and (Q_3, Q_4). To further verify the semantic separation achieved by these rules, we extract feature representations for all questions using the CLIP text encoder, compute pairwise cosine similarities, and visualize the results. We randomly select nine representative images for illustration. As shown in Figure 4, each heatmap consistently exhibits three distinct color clusters along both rows and columns, indicating that our rule-based construction effectively maintains semantic diversity and minimizes redundancy among questions.

E More Cases of MM4

We present additional examples from MM4 to provide a more intuitive demonstration of the importance of question-aware understanding for vision-

language models.

Case-1:

Q-1: What color dominates the flowers in the foreground? (Type: CP)

Options: **A**: Blue, **B**: Yellow, **C**: Red, D: Purple

Q-2: What color dominates the flowers in the foreground? (Type: CP)

Options: **A**: Yellow, B: Purple, **C**: Red, **D**: Blue

Q-3: What detail is noticeable about the ground surrounding the flowers? (Type: FP)

Options: **A**: It is fully paved, **B**: It is covered in grass, C: It is partially rocky, **D**: It is sandy

Q-4: What might be a practical reason for the arrangement of flowers? (Type: IR)

Options: A: To attract pollinators, **B**: To prevent erosion, **C**: To block noise, **D**: To create shade

Answers:

GPT-4o: D, B, C, A **QWen2.5-vl-max**: D, B, C, A **QWen2.5-vl-plus**: **C**, B, C, A
Doubao-vision-pro-32k: **C**, **A**, **B**, A **GPT-4v**: **C**, B, **B**, A **iGVLM-13B**: D, B, C, A
QA-ViT-13B: D, B, C, A **LLaVA-1.5-13B**: D, B, C, A

Case-2:

Q-1: What is the age range of the person wearing the yellow T-shirt? (Type: IR)

Options: **A**: 0-2 years old, **B**: 20-25 years old, C: 4-6 years old, **D**: 60-65 years old

Q-2: What is the age range of the person wearing the yellow T-shirt? (Type: IR)

Options: A: 4-6 years old, **B**: 20-25 years old, **C**: 0-2 years old, **D**: 60-65 years old

Q-3: What's the person in the white T-shirt doing? (Type: IR)

Options: **A**: Reading a book, B: Teaching, **C**: Drawing, **D**: Writing

Q-4: What's on the wall in the upper right corner of the picture? (Type: FP)

Options: **A**: Star posters, **B**: People photos, **C**: Calendar, D: Animal stickers

Answers:

GPT-4o: C, A, B, D **QWen2.5-vl-max**: C, A, B, D **QWen2.5-vl-plus**: C, A, B, D
Doubao-vision-pro-32k: C, A, B, **A** **GPT-4v**: **B**, **B**, B, **B** **iGVLM-13B**: C, A, B, D
QA-ViT-13B: C, A, B, **C** **LLaVA-1.5-13B**: C, **C**, B, D

Case-3:

Q-1: What is the man in the picture doing? (Type: IR)

Options: **A**: Running, **B**: Cycling, **C**: Skating, D: Skiing

Q-2: What is the man in the picture doing? (Type: IR)

Options: **A**: Running, **B**: Cycling, C: Skiing, **D**: Skating

Q-3: What is the direction of motion of the dog in the diagram? (Type: IR)

Options: A: In the same direction as the man's movement, **B**: In the opposite direction of the man's movement, **C**: Movement toward the camera, **D**: Unable to determine direction

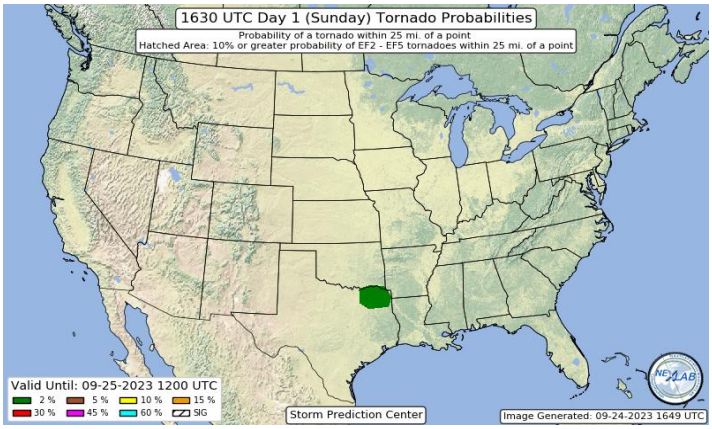
Q-4: Which option does not match the man's appearance? (Type: FP)

Options: **A**: He's wearing gloves, B: He's wearing a mask, **C**: He's carrying a backpack, **D**: He's wearing blue pants

Answers:

GPT-4o: D, C, A, B **QWen2.5-vl-max**: D, C, A, B **QWen2.5-vl-plus**: D, C, **B**, B
Doubao-vision-pro-32k: D, C, A, **A** **GPT-4v**: D, C, A, B **iGVLM-13B**: D, C, A, B
QA-ViT-13B: D, C, **B**, B **LLaVA-1.5-13B**: D, C, **B**, B

Case-4:



Q-1: Which place is the green area on the map? (Type: FP)

Options: A: Texas, **B**: New Mexico, **C**: Arkansas, **D**: Mississippi

Q-2: Which place is the green area on the map? (Type: FP)

Options: **A**: Mississippi, **B**: New Mexico, **C**: Arkansas, D: Texas

Q-3: Which country is this? (Type: CP)

Options: **A**: Canada, **B**: Mexico, C: United States, **D**: China

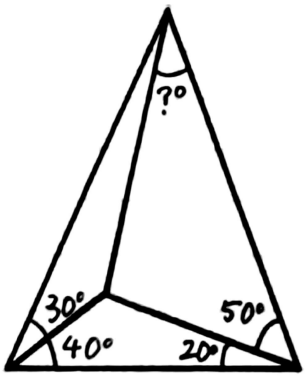
Q-4: Where is the green area marked on the map located? (Type: IR)

Options: **A**: North, B: South, **C**: Southwest, **D**: East

Answers:

GPT-4o: A, D, C, **C** **QWen2.5-vl-max**: **C**, **C**, C, **C** **QWen2.5-vl-plus**: **C**, **C**, C, **C**
Doubao-vision-pro-32k: **C**, **C**, C, **C** **GPT-4v**: **B**, **C**, C, **C** **iGVLM-13B**: A, D, C, **C**
QA-ViT-13B: **B**, D, C, **C** **LLaVA-1.5-13B**: **B**, **B**, C, **C**

Case-5:



Q-1: What is the answer in the picture? (Type: MA)

Options: A: 30, **B**: 40, **C**: 60, **D**: 10

Q-2: What is the answer in the picture? (Type: MA)

Options: **A**: 10, **B**: 40, **C**: 60, D: 30

Q-3: How many obtuse angles are in the picture? (Type: MA)

Options: **A**: 4, **B**: 3, C: 2, **D**: 1

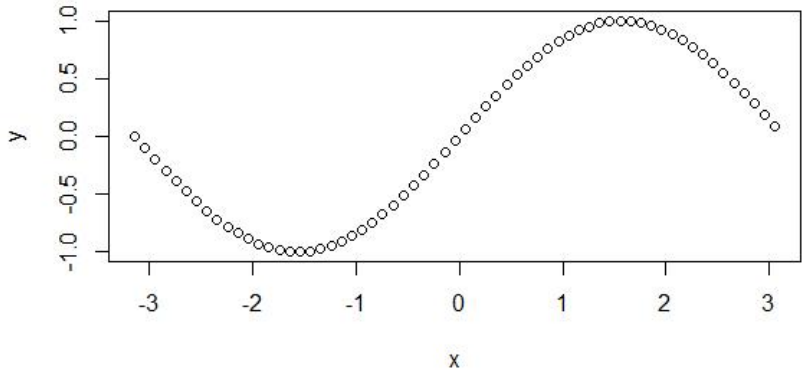
Q-4: How many triangles are in the picture? (Type: MA)

Options: **A**: 5, B: 4, **C**: 6, **D**: 3

Answers:

GPT-4o: **C**, **C**, **B**, **A** **QWen2.5-vl-max**: **C**, **C**, C, **C** **QWen2.5-vl-plus**: **C**, **C**, **D**, **C**
Doubao-vision-pro-32k: A, **B**, **A**, **A** **GPT-4v**: **C**, D, **D**, **A** **iGVLM-13B**: **B**, **B**, C, B
QA-ViT-13B: **B**, **B**, C, **D** **LLaVA-1.5-13B**: A, D, **D**, **D**

Case-6:



Q-1: What is the range of this function? (Type: MA)

Options: A: [-1, 1], **B**: [0, 2], **C**: [0, 1], **D**: [-1, 0]

Q-2: What is the range of this function? (Type: MA)

Options: **A**: [0, 2], B: [-1, 1], **C**: [0, 1], **D**: [-1, 0]

Q-3: What is the parity of this function? (Type: MA)

Options: **A**: Neither odd nor even, **B**: Even function, C: Odd function, **D**: Both even and odd function

Q-4: What is the monotonicity of this function? (Type: MA)

Options: **A**: Monotonically increasing, **B**: Monotonically decreasing, **C**: First monotonically increasing, then decreasing, and finally increasing again, D: First monotonically decreasing, then decreasing, and finally decreasing

Answers:

GPT-4o: A, B, **B**, **C** **QWen2.5-vl-max**: A, B, **B**, **C** **QWen2.5-vl-plus**: A, B, **A**, **C**
Doubao-vision-pro-32k: **C**, **C**, **A**, **C** **GPT-4v**: A, B, **B**, **C** **iGVLM-13B**: A, B, **A**, **C**
QA-ViT-13B: **C**, B, C, **C** **LLaVA-1.5-13B**: **C**, B, C, **C**