Structural Graph Probing of Vision–Language Models

Supplementary Material

8. Additional Experimental Setup

Models. We evaluate three representative vision–language models: InternVL3-1B [52], Qwen2.5-VL Instruct-3B [6], and LLaVA-1.5-7B [29]. These models span both instruction-tuned and general-purpose VLM families. InternVL3 uses an EVA-style visual encoder [41], Qwen2.5-VL adopts a dynamic-resolution visual frontend, and LLaVA-1.5 uses CLIP ViT-L/14 as its visual encoder. For each model, we extract layerwise hidden activations from the multimodal transformer backbone after projected visual tokens are concatenated with tokenized text embeddings, and construct correlation graphs from these hidden states.

Datasets. We evaluate predictability on seven benchmarks: CLEVR [21], TDIUC [22], MMMU [47], MMMU-Pro [48], BLINK [16], EMMA [18], and MHaluBench [10].

For CLEVR, we adapt the task to object counting using 10,000 image–question pairs, where the model is asked to count the number of visible objects in a synthetic scene. For TDIUC, we use the sports recognition subset containing 4,634 image–question pairs across six activity categories: baseball, surfing, skiing, tennis, frisbee, and skateboarding. For MHaluBench, we use the validation split with 2,110 examples, consisting of 1,055 hallucinating and 1,055 non-hallucinating responses. Each sample is formed by concatenating a question with either a faithful or hallucinated answer. CLEVR, TDIUC, and MHaluBench are emphasized in the main paper because they provide focused tests of numerical grounding, semantic discrimination, and multimodal factual consistency, respectively. Results on MMMU, MMMU-Pro, BLINK, and EMMA are included to assess broader generality.

Training Details. For each dataset, we randomly split examples into 80% training and 20% test sets. We train both a linear probe and a GCN probe on each layer's graph representation using cross-entropy loss for classification and standard regression objectives for numerical prediction. Optimization is performed with Adam [23], and we report the best test performance across epochs. All experiments are conducted on a single NVIDIA L40S GPU.

Text-only Baselines for Hallucination. For hallucination detection on MHaluBench, we construct two text-only baselines using `word2vec` [32, 53]: the mean embedding of the question–answer prompt and the token count of the prompt. Separate linear classifiers are trained on these features to quantify how much hallucination is predictable from shallow textual information alone before introducing graph-based multimodal representations.

Table 5. **Precision and recall for graph probing across multimodal benchmarks.** Companion to Table 1, reporting Precision and Recall for linear and graph-based probes on TDIUC, CLEVR, MMMU, MMMU-Pro, BLINK, and EMMA. Best result in each model–dataset pair is in **bold**.

Dataset	InternVL3-1B				Qwen2.5-VL-3B				LLaVA-1.5-7B			
	Linear		GCN		Linear		GCN		Linear		GCN	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
TDIUC	0.879	0.839	**0.961**	**0.957**	0.932	0.929	**0.970**	**0.970**	**0.964**	**0.967**	0.945	0.949
CLEVR	0.980	0.980	**0.993**	**0.993**	0.919	0.919	**0.962**	**0.963**	0.619	0.594	**0.712**	**0.684**
MMMU	**0.270**	0.288	0.231	**0.300**	0.226	0.279	**0.337**	**0.337**	0.320	**0.324**	**0.389**	0.261
MMMU-Pro	**0.368**	0.280	0.272	**0.323**	0.157	0.256	**0.299**	**0.300**	0.224	0.260	**0.306**	**0.306**
BLINK	0.547	0.541	**0.591**	**0.590**	0.544	0.544	**0.564**	**0.564**	**0.653**	**0.650**	0.592	0.591
EMMA	0.160	0.288	**0.332**	**0.322**	0.167	0.247	**0.316**	**0.324**	0.359	0.267	**0.386**	**0.296**