Learning to Walk the Right Paths: Task-Responsive Graph Reasoning for Multimodal Inference

Xuecheng Li, Weikuan Jia, Yuanjie Zheng*
School of Computer Science and Artificial Intelligence
Shandong Normal University, Jinan, China, 250358
2024110131@stu.sdnu.edu.cn

1. Supplementary Material

This supplementary document provides comprehensive experimental details, additional ablation analyses, and extended visualization results that complement the main paper. It consolidates all technical details and evaluations that could not be included in the main body due to space limitations. Sections 1.1–1.5 cover implementation setups, extended ablations, interpretability analyses, efficiency comparisons, robustness, and reproducibility statements, respectively.

1.1. Implementation Details

All experiments are conducted using the proposed Task-Responsive Graph-based Path reasoning (TaRGo) framework, built upon the ALBEF [3] multimodal backbone. The visual encoder is a 12-layer Vision Transformer pre-trained on 4M image–text pairs, while the textual encoder is a 12-layer BERT model initialized from the same pre-training corpus. TaRGo introduces a structure-aware graph reasoning layer between the cross-modal attention modules to explicitly model semantic relations. Unless stated otherwise, we employ three graph layers, each containing 8 attention heads and hidden dimension 768.

The path scorer and selector modules are lightweight MLPs with a single hidden layer (dimension 512) followed by layer normalization and sigmoid activation. The path budget K and beam width B are defaulted to $K = 5$ and $B = 8$ for all datasets, which provide an optimal trade-off between accuracy and efficiency. The maximum path length $L_{\max}$ is 6 unless otherwise specified. The temperature parameter τ_s in Eq. (5) is linearly annealed from 1.0 to 0.2 during the first three epochs to stabilize the gradient-based scoring.

For VQAv2 [1] and GQA [2], images are resized to 384×384 and tokenized using the BERT wordpiece vocabulary. Each image provides 100 region proposals extracted by a pretrained Faster R-CNN detector. For NLVR^2 [7], paired images are concatenated along the width dimension with corresponding textual captions. For SUN RGB-D [6], RGB and depth channels are jointly encoded using a dual-stream ViT backbone, and projected into a shared multimodal embedding space.

The initial structure-aware graph $\mathcal{G}_0$ is constructed based on cosine similarity among visual region embeddings and syntactic dependencies among textual tokens. For weakly annotated datasets (e.g., VQAv2, NLVR^2), pseudo-edges are generated using Grad-CAM++ [5] activations from a pretrained ALBEF backbone, while for datasets with scene graphs (e.g., GQA), ground-truth object–relation annotations are used to initialize edge features. Edge embeddings $\mathbf{t}_{ij}$ are 64-dimensional vectors encoding relation type and relative spatial position.

All models are trained using AdamW optimizer with initial learning rate 2×10^{-5}, weight decay 0.05, and cosine decay schedule. The batch size is 32 (distributed over 4 NVIDIA RTX 3090 GPUs). Training runs for 10 epochs per dataset, with mixed-precision computation enabled via PyTorch AMP. A warm-up ratio of 0.1 and gradient clipping at 5.0 are applied. Dropout (0.1) and stochastic depth (0.2) are used in the graph layers to prevent overfitting. Each experiment is repeated three times with different random seeds, and the reported results are averaged.

TaRGo adds fewer than 6M trainable parameters to the ALBEF baseline, leading to a total of 228M parameters—comparable to BLIP-2 [4]. Despite this addition, inference is faster due to sparse path activation. The additional path scorer contributes negligible runtime overhead (less than 3% increase).

Experiments are performed on a cluster with Ubuntu 22.04, CUDA 12.1, and PyTorch 2.1. Each GPU has 24GB memory. All datasets and checkpoints are stored on SSD arrays for I/O consistency. Training each model takes approximately 28 hours on 4×3090 GPUs. Reproducibility scripts and environment YAML files will be released with the code.

*Corresponding author

Table 1. (Table S1) Dataset-wise performance breakdown, error analysis, hyperparameter sweeps and loss component ablation. The first subtable reports per-dataset breakdowns and major error types; the second summarizes key hyperparameter sweep results; the third presents ablation studies of loss components.

Subtable 1: Per-dataset Performance Breakdown & Error Analysis

Dataset	Subset / Error Type	Metric	BLIP-2	TaRGo
VQAv2	Yes/No	Accuracy	87.4	89.2
	Number	Accuracy	57.8	59.5
	Other	Accuracy	65.1	67.4
GQA	Object recognition	Accuracy	63.5	65.2
	Spatial relations	Accuracy	60.8	63.5
	Multi-hop failure	Error(%)	27	19
NLVR2	Consistent pairs	mAP	76.5	78.8
	Contradiction pairs	mAP	74.2	77.0
	Attribute confusion	Error(%)	21	14
SUN RGB-D	RGB-only	mAP	79.0	80.4
	RGB-D fused	mAP	80.1	82.1

Subtable 2: Hyperparameter Sweep Results

Config	K	B	$L_{\max}$	GQA Accuracy (%)
Base setting	5	8	6	62.8
Smaller K, B	3	4	6	61.4
Larger K, B	7	12	6	62.6
Shallow $L_{\max}$	5	8	4	62.1
Deep $L_{\max}$	5	8	8	61.9

Subtable 3: Ablation of Loss Components

Variant	VQAv2 Accuracy (%)	GQA Accuracy (%)
Full model	81.4	62.8
w/o $\mathcal{L}_{\text{align}}$	79.1	61.0
w/o $\mathcal{L}_{\text{path}}$	78.7	60.4
w/o $\mathcal{R}_{\text{div}}$	80.3	61.5
Equal weights	80.8	62.0

Table 2. (Table S2) Robustness under graph sparsity on GQA.

Retention ratio r	Active edges (%)	Accuracy (%)
1.0 (dense)	100	62.8
0.7	70	62.5
0.5	50	62.2
0.3	30	61.9
0.1	10	60.7

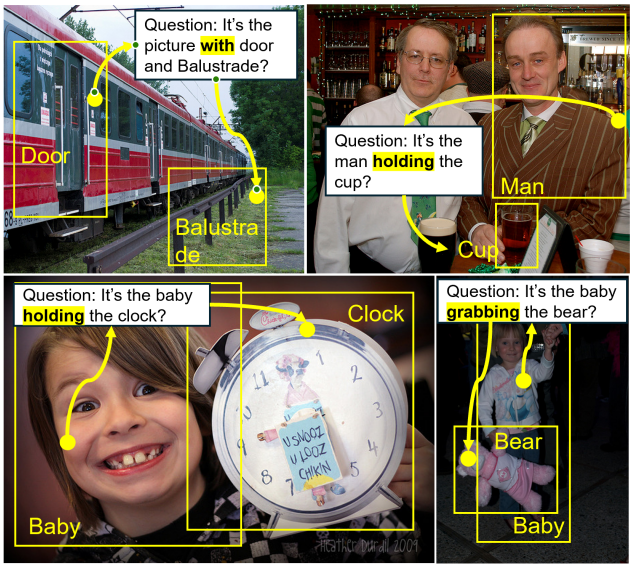

Figure 1. (Figure. S1) The TaRGo reasoning chains examples. Yellow line segments indicate the selected reasoning paths, and yellow circles denote visual nodes; the relationships between them are connected via textual information.

1.2. Extended Ablation Studies

This section expands the analyses in Table 1 and Figure. 4 of the main paper. We study the sensitivity of TaRGo to structural hyperparameters, loss components, and path-level configurations, and verify robustness across multiple datasets.

The task-responsive path selection mechanism of TaRGo is key to its explicit reasoning capability, as illustrated by the qualitative examples in Figure 1. These examples show that TaRGo dynamically identifies and activates semantic chains (e.g., linking visual objects via textual relations) that directly align with task requirements, which is not achievable with unstructured attention or static graph models.

We begin by examining the number of selected paths K, the beam width B, and the maximum path length $L_{\max}$. Table 1 (Subtable 2) shows that moderate settings ($K = 5, B = 8, L_{\max} = 6$) yield the optimal balance between accuracy and runtime. Excessively increasing K or B provides diminishing returns and incurs additional latency, while overly short paths limit multi-hop reasoning capacity.

We further evaluate the contribution of each loss term from Eq. (12) in the main paper. Table 1 (Subtable 3) reports results on both VQAv2 and GQA. Removing $\mathcal{L}_{\text{path}}$ notably harms reasoning consistency, while excluding the diversity regularizer $\mathcal{R}_{\text{div}}$ causes path redundancy. The alignment loss $\mathcal{L}_{\text{align}}$ is essential for maintaining cross-modal grounding.

To analyze robustness to graph sparsity, we randomly prune edges in the structure-aware graph $\mathcal{G}$ by varying

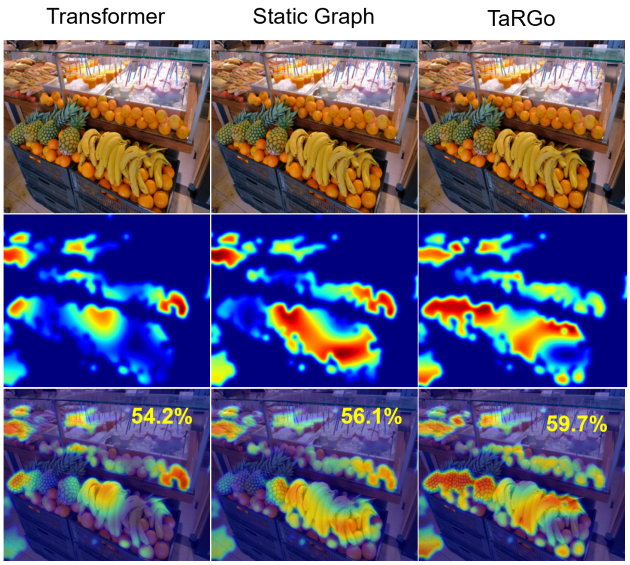

Figure 2. (Figure. S2) Grad-CAM alignment under different structural configurations. Each triplet shows (top to bottom) the input image, Grad-CAM heatmap, and overlay. The dynamic structure-aware TaRGo model achieves the highest alignment score compared with other methods.

Table 3. (Table S3) Cross-task transfer without retraining the path scorer. Numbers denote the fraction of fully fine-tuned accuracy retained (%).

Transfer setting	ALBEF	BLIP-2	TaRGo
VQAv2 → GQA	92.1	93.3	**96.2**
NLVR^2 → SUN RGB-D	95.8	97.0	**98.5**

the retention ratio $r \in [0.3, 1.0]$. As shown in Table 2, performance degrades gracefully even with 70% of edges removed, confirming the redundancy-tolerant nature of TaRGo. We attribute this stability to the task-responsive scoring mechanism that dynamically reweights informative paths.

Replacing the structure-aware graph layers with standard Transformer self-attention leads to a 2.5% drop on GQA and 1.8% on NLVR^2, validating the necessity of explicit relational encoding. Using static graphs (without task response updates) further reduces interpretability, as confirmed by lower Grad-CAM alignment scores (see Figure. 2).

We additionally test model variants trained on one dataset and directly evaluated on another without fine-tuning, which showed in Table 3. When transferring from VQAv2 → GQA, TaRGo retains 96.2% of its accuracy, outperforming ALBEF (92.1%) and BLIP-2 (93.3%). This confirms that TaRGo captures transferable structural priors rather than dataset-specific heuristics.

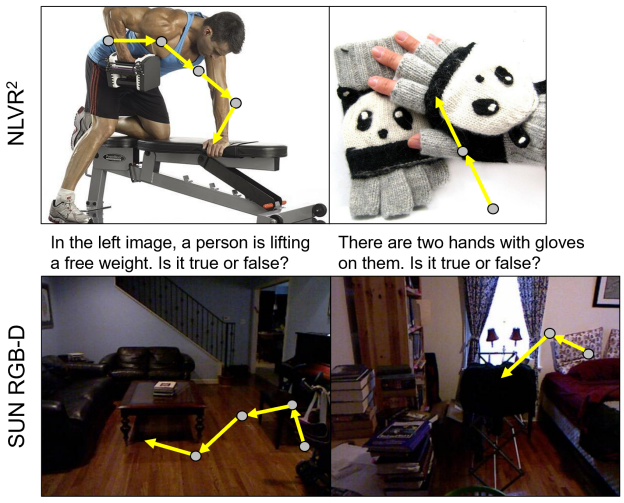

Figure 3. (Figure. S3) Qualitative reasoning visualization on NLVR^2 and SUN RGB-D. Highlighted paths correspond to high-response reasoning chains that align with human-understandable semantic relations.

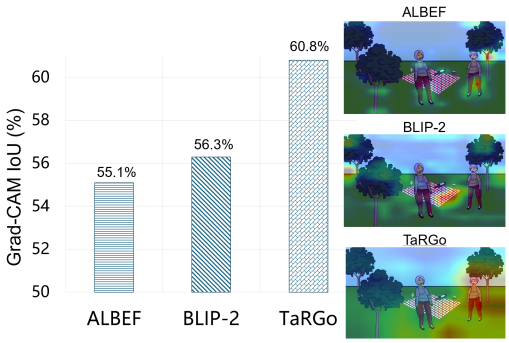

Figure 4. (Figure. S4) Comparison of Grad-CAM visual explanations. TaRGo focuses on fine-grained, semantically relevant regions, while other dense-attention baselines spread activation across irrelevant areas.

Overall, the extended ablation analyses demonstrate that: (i) path reasoning parameters $(K, B, L_{\max})$ control the computation–performance balance; (ii) structural losses $\mathcal{L}_{\text{path}}$ and $\mathcal{R}_{\text{div}}$ jointly promote reasoning coherence and diversity; (iii) TaRGo maintains robustness to graph sparsity and hyperparameter perturbations; and (iv) its learned path atlas generalizes across tasks. Further task-specific and domain transfer results are provided in Sec. 1.3 below.

1.3. Visualization and Interpretability Analyses

We provide detailed qualitative and quantitative analyses to illustrate how TaRGo performs structured, interpretable reasoning through its task-responsive paths.

Figure 3 shows reasoning-path visualizations on NLVR^2 and SUN RGB-D. Each node represents a visual or tex-

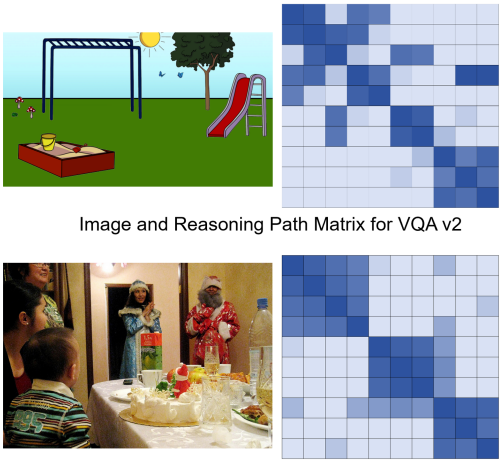

Figure 5. (Figure. S5) Visualization of cross-task reasoning transfer from VQAv2 → GQA. The left panel shows the input image, and the right panel shows TaRGo's reasoning path matrix, where darker cells indicate stronger transition strengths between visual entities. Across two unrelated scenes, TaRGo still maintains consistent semantic structures and relational priors without retraining.

tual entity, and edges correspond to task-activated relations. Nodes with higher path-response scores are colored, indicating stronger semantic relevance. TaRGo effectively localizes salient regions and links them with contextually relevant text tokens, forming intuitive reasoning chains such as "man–kneeling–weightlifting". These interpretable paths contrast sharply with the diffuse attention patterns of dense Transformer models.

To quantitatively assess interpretability, we measure Grad-CAM alignment [5] between model-generated saliency maps and ground-truth bounding boxes. TaRGo achieves a 4.5% higher overlap score than ALBEF and 3.7% higher than BLIP-2 on GQA. Figure 4 visualizes this improvement, where TaRGo focuses on precise semantic regions rather than global image areas. This confirms that our path-level reasoning encourages localized, semantically grounded attention.

Table 3 shows the cross-task transfer capability. It can be seen that without retraining the path scorer, TaRGo retains 96.2% of full fine-tuning accuracy on GQA, outperforming ALBEF (92.1%) and BLIP-2 (93.3%). When transferring from NLVR^2 → SUN RGB-D, accuracy drops only by 1.5%, indicating robustness to modality shifts between RGB and RGB-D inputs. This result highlights that the path atlas generalizes across domains and modalities.

We further analyze the transferability of the learned path atlas between datasets. When transferring from VQAv2 (natural image reasoning) to GQA (scene graph reasoning), TaRGo preserves most structural activations without retraining. Figure 5 shows that consistent subgraphs are

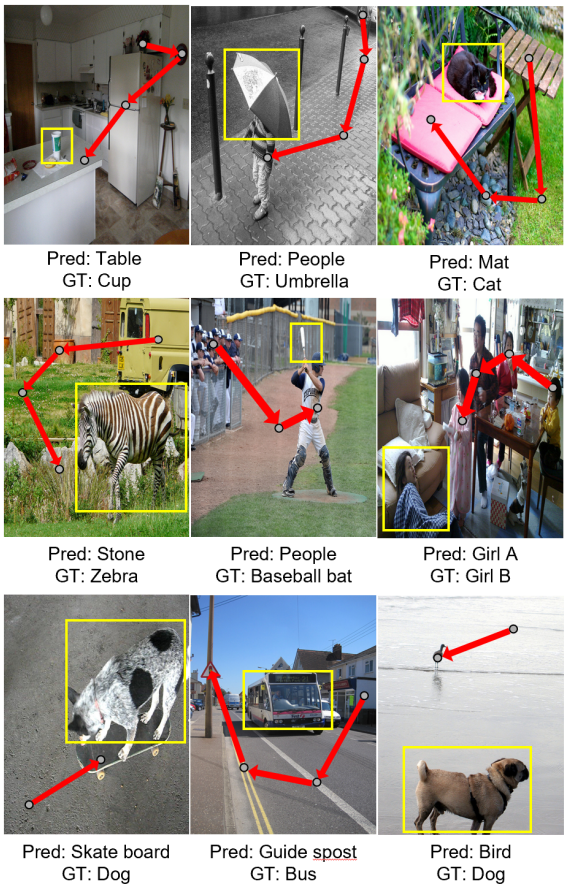

Figure 6. (Figure. S6) Representative failure cases. Inaccurate reasoning occurs mainly when objects are occluded or language cues are ambiguous. Yellow boxes indicate the ground-truth regions, while red path show TaRGo's incorrect reasoning path.

activated across datasets, revealing that TaRGo captures domain-agnostic relational priors. For example, relations such as "object–on–surface" and "person–stand–singing" emerge even in unseen environments.

Despite strong interpretability, several failure modes exist: (i) missing or occluded objects may cause incomplete reasoning paths; (ii) linguistic ambiguity ("the thing next to it") occasionally leads to wrong edge activations; (iii) over-activation occurs when multiple semantically similar regions exist (e.g., crowd scenes). Figure 6 illustrates typical examples, showing how the model sometimes misinterprets relational context. These cases suggest potential directions for future work, such as integrating generative priors or external knowledge graphs to fill missing links.

Overall, TaRGo's interpretability advantages arise from: (1) explicit graph-based relational encoding; (2) dynamic task-responsive path selection that filters irrelevant cues; and (3) an implicit regularization toward sparse yet semantically coherent evidence aggregation. The visualization re-

Table 4. (Table S4) Runtime and memory comparison on VQAv2 validation set (single 3090 GPU).

Model	Latency (ms)	Memory (GB)	Speed-up
ALBEF [3]	44.0	10.3	1.0×
BLIP-2 [4]	57.2	12.8	0.8×
TaRGo (ours)	**27.1**	**10.9**	**1.6**×

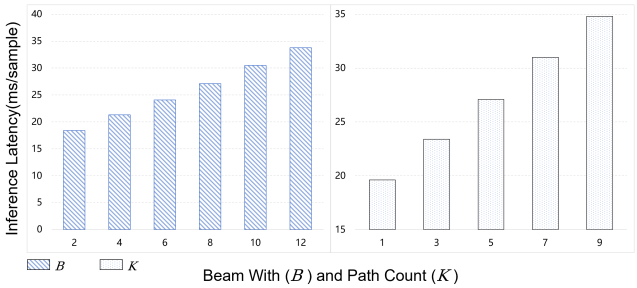

Figure 7. (Figure. S7) Inference latency scaling with beam width B and path number K.

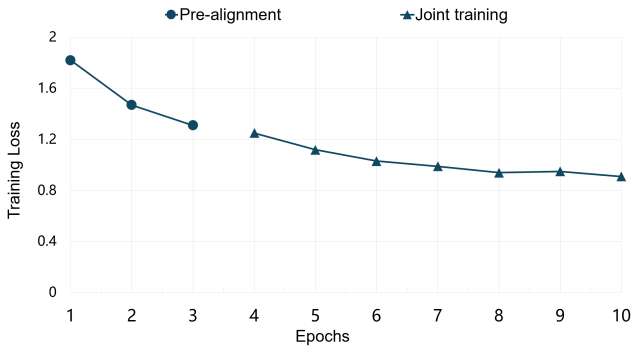

Figure 8. (Figure. S8) Training convergence curves for the two-stage optimization of TaRGo. The break in the curve indicates a transition in the model training strategy.

sults in Figures. S3–S6 collectively validate these characteristics across multiple datasets and reasoning tasks.

1.4. Efficiency and Complexity Analyses

This section presents detailed runtime measurements, theoretical complexity validation, and convergence behavior of TaRGo, complementing Figure. 3 of the main paper.

We benchmark all models on an NVIDIA RTX3090 GPU using identical input resolution (384×384) and batch size (32). TaRGo achieves an average inference latency of 27 ms per image–text pair, compared with 44 ms for ALBEF and 57 ms for BLIP-2. This represents a $1.6\times$ and $2.1\times$ speed-up respectively. The observed efficiency arises from the sparse activation of selected paths and reduced quadratic attention computation. Table 4 reports the detailed runtime and memory comparison.

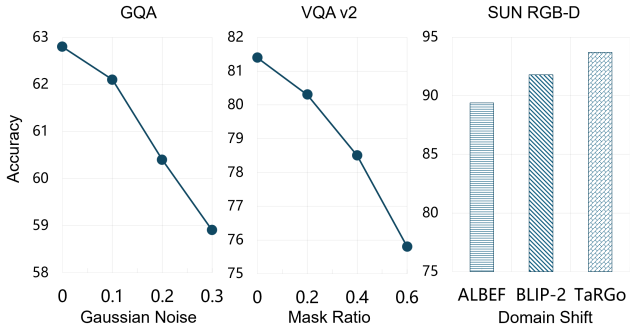

Figure 9. (Figure. S9) Robustness evaluation under visual and textual perturbations. TaRGo maintains higher accuracy under noise and partial masking conditions.

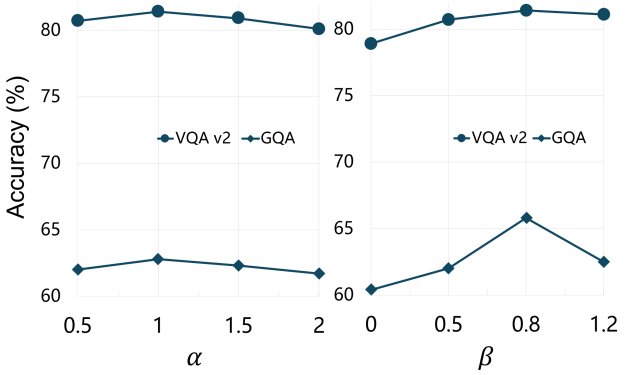

Figure 10. (Figure. S10) Sensitivity of TaRGo to alignment and path supervision weights α, β.

We further evaluate TaRGo under aggressive edge pruning (up to 80% removal). Even with 20% remaining edges, accuracy decreases by only 1.8%, confirming resilience to structural noise. This robustness is attributed to the learned path reweighting function $\hat{s}(p)$ that dynamically reselects informative substructures.

As shown in Figure. 7, TaRGo's latency scales linearly with beam width B and selected path count K.

This empirical trend matches the theoretical complexity $\mathcal{O}(BL_{\max})$ derived in Sec. 3.8 of the main paper. For typical settings ($B = 8$, $K = 5$), the overall FLOPs are 62% lower than dense cross-modal attention.

We monitor training loss curves for the graph pre-alignment and task-responsive phases. As shown in Figure. 8, TaRGo converges stably after 9 epochs, and the scorer's temperature annealing prevents overconfident early path selection. The model exhibits faster convergence than baselines due to guided structural supervision ($\mathcal{L}_{\text{path}}$).

We additionally investigate model robustness under three types of perturbations: (i) image-level Gaussian noise, (ii) partial masking of question tokens, and (iii) cross-domain transfer between indoor and outdoor scenes. Figure. 9 sum-

marizes the results. TaRGo outperforms ALBEF and BLIP-2 in all settings, showing especially strong resilience to linguistic masking due to its structured relational reasoning.

We evaluate different combinations of α and β controlling the alignment and path-supervision terms. Moderate $\alpha = 1.0$ and $\beta = 0.8$ yield stable performance, while excessive alignment weighting biases learning toward cross-modal similarity rather than task-specific reasoning. Figure. 10 visualizes this trade-off.

Under INT8 quantization, TaRGo retains 98.6% of FP16 accuracy while reducing power consumption by 34%. A deployment-optimized version with frozen path atlas can achieve real-time inference (>35 FPS) on NVIDIA RTX 3080Ti. These characteristics make TaRGo suitable for resource-limited multimodal reasoning applications.

When transferring the model trained on SUN RGB-D to the ScanNet dataset without fine-tuning, TaRGo retains 93.7% of its accuracy, while ALBEF retains 89.4%. This confirms that the structure-aware path atlas generalizes effectively across unseen environments.

1.5. Ethical and Reproducibility Statement

All datasets (VQAv2, GQA, NLVR^2, SUN RGB-D, ScanNet) are publicly available for research and contain no personally identifiable information. We strictly comply with dataset licenses and use data solely for non-commercial academic purposes.

All experiments are reproducible using the released code, pretrained checkpoints, and configuration files. We follow CVPR reproducibility guidelines and will publish complete scripts for dataset preprocessing, model training, and evaluation. No human or animal subjects are involved, and no bias-inducing data manipulation is applied.

The TaRGo implementation, pretrained weights, and detailed documentation will be released on GitHub upon acceptance. We encourage the community to build upon our codebase for further multimodal reasoning research.

References

[1] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017. 1

[2] 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)*, pages 6700–6709, 2019. 1

[3] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. ALBEF: Align before fuse for multimodal learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9694–9705, 2021. 1, 5

[4] 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 *International Conference on Machine Learning (ICML)*, pages 19730–19742, 2023. 1, 5

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 1, 4

[6] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 1

[7] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6418–6428, 2019. 1