Spatio-Structural Image Captioning via LLM Fine-Tuning with Depth-Enhanced Scene Graphs

Le Vinh Thuan[1], Nguyen Minh Khoa[2], Nguyen Vinh Thanh[3], and Nguyen Thi Dinh[4,]

[{1,2,3}]Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam,
[4]Ho Chi Minh City University of Industry and Trade, Ho Chi Minh City, Vietnam

[1]lvthuan23@apcs.fitus.edu.vn, [2]nmkhoa23@apcs.fitus.edu.vn, [3]23120012@student.hcmus.edu.vn, [4]dinhnt@huit.edu.vn

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

The automatic generation of natural language descriptions from images remains one of the most challenging and important tasks in computer vision and natural language processing [1, 13]. This interdisciplinary field, commonly known as image captioning, requires machines to not only perceive objects and their relationships but also to generate coherent, contextually appropriate text that captures the scene's semantic essence [1, 17]. Recent advances in deep learning, particularly the adoption of Transformer-based models and Large Language Models (LLMs) in Vision-Language Pre-training (VLP) [?, ?], have significantly improved descriptive fluency. However, a critical limitation persists: current VLP and image captioning methods primarily rely on 2D feature extraction and lack a deep, grounded understanding of the 3D structure and spatial relationships within a scene. Consequently, generated captions often fail to accurately describe position, depth, and structural context (e.g., "the car is in front of the building" without specifying the distance or depth plane), limiting their applicability in domains requiring precise spatial reasoning, such as robotics or complex scene analysis. To address this gap, Knowledge Graphs (KGs) offer a crucial pathway by providing structured semantic information and enabling reasoning capabilities [10, 16]. While previous work has used KGs to incorporate general semantic and factual knowledge [?, 19], few have effectively integrated explicit structural and geometric knowledge like scene depth and 3D spatial relationships. Furthermore, existing KG-enhanced methods often lack commonsense reasoning, which is vital for generating captions that reflect human-like understanding of causality and purpose [22]. This paper introduces a novel framework, the Depth- and Commonsense-Augmented Scene Graph Captioner (DASG-CS

Captioner), designed to generate fine-grained image descriptions with enhanced Spatio-Structural and Commonsense-Grounded awareness. We propose a comprehensive pipeline that moves beyond 2D perception by integrating state-of-the-art depth estimation into the knowledge representation stage, thereby enriching the generated captions with precise spatial details. Our method leverages the structured reasoning capabilities of Knowledge Graphs and the generative power of Fine-Tuned LLMs. The main contributions of this work are: A Novel Depth- and Commonsense-Augmented Scene Graph (DASG-CS): We propose a new KG architecture that seamlessly fuses visual features, quantitative depth information (from Depth Anything v2), and commonsense knowledge (from ConceptNet) to create a multi-faceted, structured representation of the scene. Spatially-Aware LLM Fine-Tuning Strategy: We develop an effective encoding and fine-tuning strategy for LLMs, enabling the model to explicitly utilize the Spatio-Structural and Commonsense information encoded in the DASG-CS, leading to generated captions that are demonstrably richer in depth and contextual reasoning. A Three-Stage Pipeline: We present a pipeline that integrates feature extraction, multi-source knowledge fusion, structured data encoding, and natural language generation, setting a new benchmark for spatially and contextually grounded image captioning. Extensive Evaluation with Novel Metrics: We conduct a comprehensive evaluation demonstrating the effectiveness of the DASG-CS framework, including specialized metrics to quantify the improvement in Spatial Reasoning Accuracy and Contextual Richness of the generated descriptions.

2 Related Work

Image captioning has evolved significantly from simple encoder-decoder architectures [13] to sophisticated systems incorporating spatial awareness and external knowledge [1]. Our work builds upon three critical research directions: depth-aware image understanding, knowledge graph construction for vision-language tasks, and fine-tuning language models for spatial-aware captioning.

2.1 Object Detection and Spatial Feature Extraction

Traditional image captioning systems rely primarily on 2D visual features extracted from RGB images [1,17]. The YOLO family has evolved from YOLOv1's single-stage detection [12] through multiple generations [2,3,14] to more sophisticated architectures. Modern object detectors like YOLOv11 provide real-time performance with improved accuracy, particularly for small objects and crowded scenes. While object detection has advanced significantly, most systems still focus on 2D bounding boxes without incorporating depth information for true 3D spatial understanding.

2.2 Scene Understanding and Spatial Relationships

Scene graph generation has emerged as a method to capture object relationships explicitly [19, 21]. These approaches construct structured representations of vi-

sual scenes by identifying objects and their relationships [23]. However, most scene graphs focus on semantic relationships rather than precise metric spatial relationships. Our approach aims to incorporate depth estimation to enable more accurate spatial relationship modeling.

2.3 Knowledge-Enhanced Image Captioning

Knowledge graphs provide structured representations of visual scenes and external world knowledge, enabling richer semantic understanding. Visual Genome [8] pioneered large-scale visual knowledge graphs with detailed annotations of objects, attributes, and relationships. Flickr30k [20] provide image-caption pairs widely used for training vision-language models.

Several works have explored incorporating external knowledge into image captioning. Knowledge graphs have been used for visual question answering [15] and image classification [10]. Entity linking techniques [4] help map visual detections to knowledge base entities. Knowledge-based approaches [22, 24] have shown improvements by incorporating common-sense and factual knowledge.

Most existing approaches integrate knowledge at the generation stage [16]. Our pipeline differs by constructing a comprehensive knowledge graph that fuses visual observations from datasets like Visual Genome and Flickr with external knowledge sources before the generation phase.

2.4 Natural Language Generation for Image Captioning

Natural language generation is the final component of image captioning systems. Early encoder-decoder models [7,13] used RNNs to generate captions from visual features. Attention mechanisms [1,17] improved performance by allowing models to focus on relevant image regions during generation.

More recent approaches employ transformer-based architectures [5,9] for better long-range dependency modeling. The T5 model [11] treats all NLP tasks as text-to-text problems, offering a unified framework adaptable to various generation tasks including image captioning.

Our approach leverages text-to-text transformers for generating descriptions from structured semantic representations. By fine-tuning on data that explicitly encodes spatial relationships and depth information, we aim to generate descriptions that accurately reflect 3D spatial configurations.

2.5 Research Gaps and Our Contributions

Existing image captioning systems face several limitations:

1. **Limited Spatial Awareness:** Most systems [13, 17] rely on 2D visual features without depth information for true 3D spatial understanding.
2. **Knowledge Integration:** While some works integrate external knowledge [10, 22], comprehensive frameworks that systematically combine visual observations with external knowledge graphs remain limited.

3. **Spatial Description Quality:** Generated descriptions often lack precise spatial relationships and depth-aware positioning information.

Our proposed three-system pipeline addresses these gaps:

1. **System 1 - Spatial Feature Extraction:** We combine YOLOv11 for object detection with Depth Anything V2 for monocular depth estimation, enabling extraction of objects, their relationships, and depth information.
2. **System 2 - Knowledge Graph Framework:** We construct knowledge graphs based on Visual Genome [8] and Flickr [20] datasets, enriched with external knowledge from ConceptNet and DBpedia.
3. **System 3 - Depth-Aware Caption Generation:** We fine-tune language models to generate image descriptions with emphasis on spatial and depth features, producing captions that accurately reflect 3D spatial configurations.

The key contribution is the systematic integration of depth information throughout the pipeline, from detection through knowledge graph construction to caption generation, emphasizing spatial accuracy in generated descriptions.

3 Theoretical Basis

This section presents the theoretical foundations underlying our three-stage architecture, providing formal definitions and mathematical models that support the integration of computer vision, knowledge graph reasoning, and natural language generation.

3.1 Object Detection and Spatial Analysis Framework

YOLOv11 Detection Model The YOLOv11 object detection framework can be formalized as a function $f_{YOLO} : \mathbb{R}^{H\times W\times 3} \to \mathcal{D}$, where the input image $I \in \mathbb{R}^{H\times W\times 3}$ is mapped to a detection set $\mathcal{D} = \{d_1, d_2, ..., d_n\}$. Each detection d_i is represented as:

$$d_i = (bbox_i, c_i, s_i) \tag{1}$$

where $bbox_i = (x_i, y_i, w_i, h_i)$ represents the bounding box coordinates, $c_i \in \mathcal{C}$ is the class label from the predefined class set $\mathcal{C}$, and $s_i \in [0, 1]$ is the confidence score.

Depth Estimation and Spatial Reasoning We incorporate the Depth-Anything-V2 model [18] to estimate depth information, formally defined as:

$$f_{depth} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W} \tag{2}$$

The depth map $D = f_{depth}(I)$ provides spatial context that enhances object relationship understanding. Combined with the Shapely geometry library [6], we

compute spatial relationships between detected objects using geometric operations:

$$R_{spatial}(d_i, d_j) = \{distance(d_i, d_j), overlap(d_i, d_j), relative\_position(d_i, d_j)\} \tag{3}$$

3.2 Knowledge Graph Construction Theory

Entity Linking and Knowledge Fusion Given the detection set $\mathcal{D}$ from Stage I, we define the entity linking function as:

$$f_{link} : \mathcal{D} \times \mathcal{KB} \rightarrow \mathcal{E}_{linked} \tag{4}$$

where $\mathcal{KB}$ represents the external knowledge base and $\mathcal{E}_{linked}$ is the set of linked entities. The linking process employs semantic similarity scoring:

$$sim(d_i, e_k) = \alpha \cdot sim_{text}(c_i, label(e_k)) + \beta \cdot sim_{context}(context(d_i), context(e_k)) \tag{5}$$

where α and β are weighting parameters, and sim_{text} and $sim_{context}$ represent textual and contextual similarity measures respectively.

Knowledge Graph Formalization The constructed knowledge graph is formally defined as a directed graph $KG = (V, E, R, A)$ where $V = \{v_1, v_2, ..., v_m\}$ is the set of entity vertices, $E \subseteq V \times V$ represents the edges between entities, R is the set of relation types, and $A : V \rightarrow \mathcal{P}(\mathcal{A})$ maps entities to their attribute sets.

3.3 T5-based Natural Language Generation Framework

Text-to-Text Transfer Learning Our approach leverages the T5 (Text-to-Text Transfer Transformer) model [11] for generating natural language descriptions from structured semantic representations. The T5 framework treats all NLP tasks as text-to-text problems:

$$f_{T5} : \mathcal{S}_{semantic} \rightarrow \mathcal{T}_{text} \tag{6}$$

where $\mathcal{S}_{semantic}$ represents the semantic input derived from the knowledge graph and $\mathcal{T}_{text}$ is the generated natural language output.

Semantic Representation Encoding The knowledge graph entities and relationships are encoded into a structured semantic representation $S_{semantic}$ that serves as input to the T5 model:

$$S_{semantic} = encode(KG, R_{spatial}, C_{context}) \tag{7}$$

where $C_{context}$ represents the contextual information derived from the enriched data analysis.

Attention Mechanism for Semantic Focus The T5 model employs multi-head self-attention to focus on relevant semantic elements:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where Q, K, and V represent query, key, and value matrices respectively, derived from the semantic representation.

3.4 Contextual Reasoning and Inference Theory

Enriched Data Analysis Our approach performs contextual reasoning using enriched data from the knowledge graph without explicit relationship modeling. The contextual inference function is defined as:

$$f_{inference} : \mathcal{E}_{enriched} \rightarrow \mathcal{I}_{context} \tag{9}$$

where $\mathcal{E}_{enriched}$ represents the enriched entity set and $\mathcal{I}_{context}$ denotes the inferred contextual information.

Semantic Coherence Optimization To ensure semantic coherence in the generated descriptions, we optimize the following objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{generation} + \lambda_1 \mathcal{L}_{coherence} + \lambda_2 \mathcal{L}_{semantic} \tag{10}$$

where $\mathcal{L}_{generation}$ is the standard language modeling loss, $\mathcal{L}_{coherence}$ measures textual coherence, $\mathcal{L}_{semantic}$ ensures semantic consistency with the input knowledge graph, and λ_1, λ_2 are regularization parameters.

3.5 Multi-Modal Integration Theory

Cross-Modal Alignment The integration of visual features, spatial information, and semantic knowledge requires cross-modal alignment. We define the alignment function as:

$$f_{align} : \mathcal{F}_{visual} \times \mathcal{F}_{spatial} \times \mathcal{F}_{semantic} \rightarrow \mathcal{F}_{unified} \tag{11}$$

where $\mathcal{F}_{visual}$, $\mathcal{F}_{spatial}$, and $\mathcal{F}_{semantic}$ represent visual, spatial, and semantic feature spaces respectively, and $\mathcal{F}_{unified}$ is the unified representation space.

Information Fusion Strategy The multi-modal information fusion employs weighted combination of feature representations:

$$\mathcal{F}_{unified} = \sum_{i=1}^{3} w_i \cdot \phi_i(\mathcal{F}_i) \tag{12}$$

where ϕ_i represents the transformation function for each modality, and w_i are learned weights that adapt to the relative importance of each information source.

This theoretical framework provides the mathematical foundation for our three-stage architecture, ensuring principled integration of computer vision, knowledge representation, and natural language generation components.

4 Proposed Approach

4.1 Overall Architecture

Our proposed architecture consists of three main stages: (1) Spatial Feature Extraction, (2) Knowledge Graph Construction, and (3) Depth-Aware Caption Generation. Each stage is designed to progressively enrich the input image data with spatial and semantic information, culminating in the generation of detailed and contextually relevant captions.

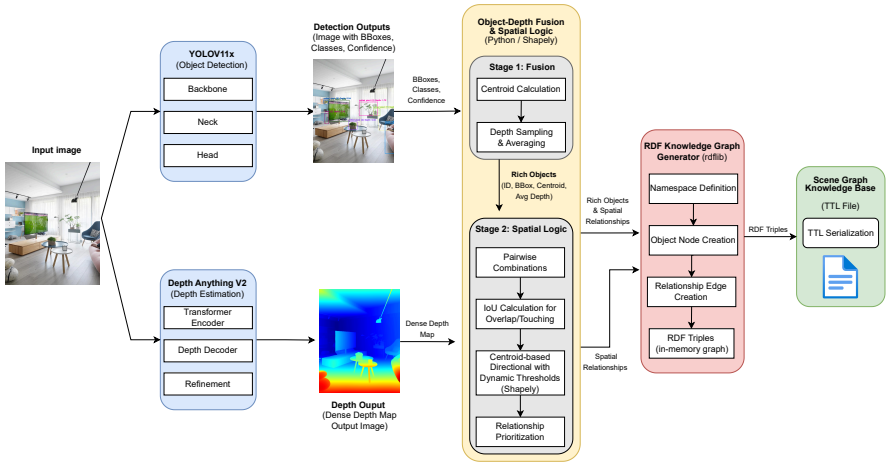

Fig. 1. Overall Architecture of the Proposed Approach

4.2 Stage 1: Initial Visual-Spatial Knowledge Generation

The first stage of our proposed architecture focuses on generating a preliminary, intrinsic Knowledge Graph (KG) that encodes both the semantic entities and their derived spatial-structural relationships within the input image. This initial graph, denoted as the Internal Knowledge Graph ($\mathbf{KG}_{\text{Int}}$), serves as the foundational, image-grounded structure before external knowledge enrichment.

As illustrated in Figure 1, Stage 1 encompasses the feature extraction, object-depth fusion, and the construction of the $\mathbf{KG}_{\text{Int}}$.

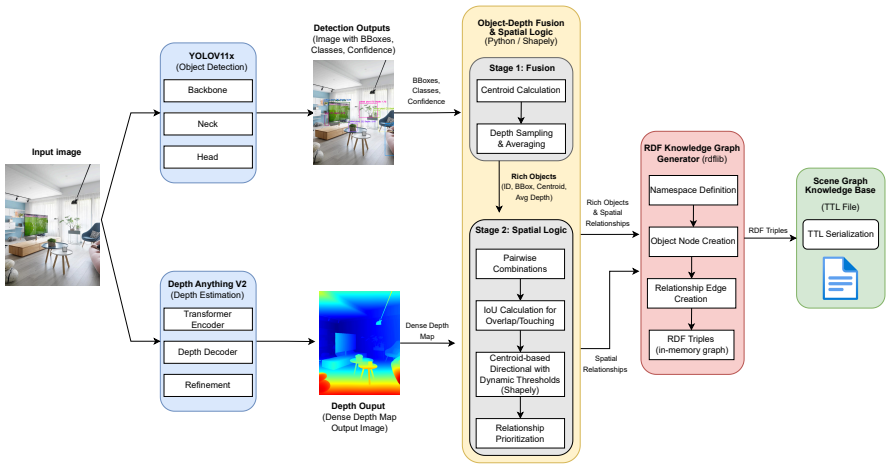

Fig. 2. Stage 1 architecture: Initial Visual-Spatial Knowledge Generation

Object and Depth Feature Extraction We employ a dual-branch feature extraction mechanism.

1. **Object Detection:** The **YOLOv11x** model is used to detect objects and extract attributes. The output is a set of detections $D = \{d_i\}_{i=1}^{n}$, where each d_i includes the bounding box B_i, class label C_i, and confidence score S_i.
2. **Depth Estimation:** The **Depth Anything V2** model generates a high-resolution, dense Depth Map $D_{\text{map}} \in \mathbb{R}^{H \times W}$. This map provides the crucial relative depth information necessary for 3D spatial reasoning.

Object-Depth Fusion and Spatial Logic The Fusion and Spatial Logic modules convert the raw visual and depth data into structured relationships.

1. **Quantification and Fusion:** The bounding boxes B_i are aligned with D_{map}. We compute the Average Relative Depth ($\bar{d}_i$) for each object O_i within its bounding box. This step yields a set of Rich Objects R, where each object entity is augmented with its spatial center and $\bar{d}_i$. The $\bar{d}_i$ values are used exclusively for internal calculation of relative depth and are not stored as absolute values in the final KG.
2. **Spatial Predicate Generation:** The core of this stage is the derivation of new spatial predicates. By comparing the difference in average depth $\mathbf{\Delta d_{i,j}} = \bar{d}_i - \bar{d}_j$ between object pairs (O_i, O_j), we generate a Novel Set of Qualitative 3D Spatial Predicates ($\mathcal{P}_{3D}$):

   $\mathcal{P}_{3D} \subset \{\text{is\_closer\_than}, \text{is\_farther\_than}, \text{on\_same\_depth\_plane\_as}, \text{in\_immediate\_foreground}, \text{etc.}\}$

   This logic uses dynamic thresholds to convert the continuous $\Delta d_{i,j}$ into discrete, robust relational triplets. Traditional 2D relationships (overlap, adja-

cency) are computed alongside these 3D predicates (using methods like IoU and Centroid-based Directional logic).

Internal Scene Graph Construction The resulting Rich Objects and the generated 2D/3D relationships are fed into the RDF Knowledge Graph Generator (utilizing libraries such as `rdflib`). This module performs:

1. **Namespace Definition** and **Object Node Creation** (from R).
2. **Relationship Edge Creation** (from $\mathcal{P}_{3D}$ and 2D predicates).

This process finalizes the Internal Knowledge Graph ($\mathbf{KG}_{\text{Int}}$), a set of RDF Triples that accurately models the scene's visual and spatial configuration. This $\mathbf{KG}_{\text{Int}}$ is then serialized (e.g., into a TTL file) as the output of Stage 1, ready for external knowledge enrichment in Stage 2.

4.3 Stage 2: DASG-CS Construction and Multi-Source Fusion

The primary objective of Stage 2 is the construction of the Depth- and Commonsense-Augmented Scene Graph (**DASG-CS**) by integrating the Internal Knowledge Graph ($\mathbf{KG}_{\text{Int}}$) with external knowledge sources. This stage addresses the critical limitation of purely visual KGs by introducing contextual reasoning capabilities. The architecture for this stage is depicted in Figure 3.

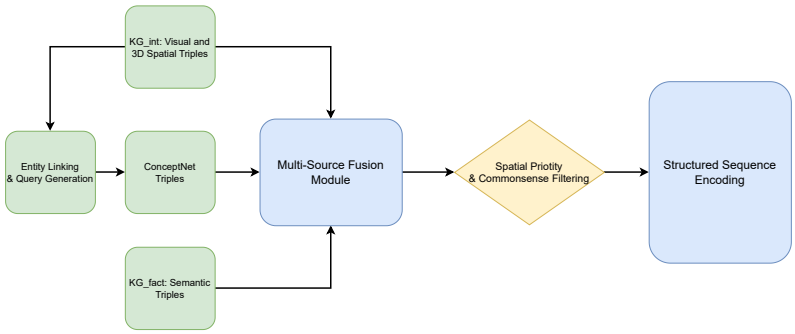

Fig. 3. Architecture of the DASG-CS Construction and Fusion Module (Stage 2)

Knowledge Source Acquisition Three distinct knowledge streams are prepared for fusion:

1. **Internal Visual and 3D Spatial Triples ($\mathbf{KG}_{\text{Int}}$):** This is the output from Stage 1, containing robust 2D semantic triples and our novel $\mathcal{P}_{3D}$ set of qualitative 3D spatial predicates.

2. **Factual/Semantic Triples ($\mathbf{KG_{Fact}}$):** This stream incorporates standard semantic relationships retrieved from large-scale visual knowledge bases (e.g., Visual Genome and Flickr), providing general object-to-object semantic links.
3. **Commonsense Triples ($\mathbf{KG_{CS}}$):** This knowledge is acquired through a two-step process: **Entity Linking & Query Generation** uses the object entities (Nodes) from $\mathbf{KG}_{\text{Int}}$ as queries to retrieve relevant triples from the **ConceptNet** semantic network. This yields $\mathbf{KG}_{\text{CS}}$, which enriches the graph with human-like understanding of context, purpose, and causality.

Multi-Source Fusion and Graph Consolidation The three knowledge streams ($\mathbf{KG}_{\text{Int}}$, $\mathbf{KG}_{\text{Fact}}$, $\mathbf{KG}_{\text{CS}}$) converge into the Multi-Source Fusion Module.

1. **Entity Alignment:** The module first ensures all entities (subjects and objects) are correctly aligned and unified across the three sources.
2. **Consolidation:** The module performs a set union operation to form the raw **DASG-CS**:

$$\textbf{DASG-CS} = \mathbf{KG}_{\text{Int}} \cup \mathbf{KG}_{\text{Fact}} \cup \mathbf{KG}_{\text{CS}}$$

3. **Conflict Resolution:** During consolidation, $\mathbf{KG}_{\text{Int}}$ (the direct visual evidence) is assigned the highest priority to resolve any potential conflicts or inconsistencies arising from general/abstract knowledge found in $\mathbf{KG}_{\text{Fact}}$ or $\mathbf{KG}_{\text{CS}}$.

Spatial Priority and Commonsense Filtering The raw **DASG-CS** is typically redundant and high-dimensional. We introduce a filtering mechanism (represented by the diamond shape in Figure 3) to retain only the most impactful triples for caption generation:

1. **Spatial Priority Filtering:** Triples containing the Novel 3D Spatial Predicates ($\mathcal{P}_{3D}$) are automatically prioritized and flagged. This ensures the LLM's final output is strongly grounded in the scene's structural awareness.
2. **Commonsense Filtering:** $\mathbf{KG}_{\text{CS}}$ triples are filtered based on a computed Relevance Score (e.g., semantic similarity between the triple and the image context) to remove weak or generic commonsense links that might otherwise introduce noise.

This filtering step yields a set of high-quality, non-redundant, and prioritized triples $\mathcal{T}_{\text{opt}}$, ready for the final encoding.

Structured Sequence Encoding The final step of Stage 2 is the Structured Sequence Encoding. The optimized set of triples $\mathcal{T}_{\text{opt}}$ is serialized into a single input sequence S_{prompt} for the LLM. This encoding uses specific delimiter tokens (e.g., `[3D_START]`, `[CS_FACT]`) to clearly distinguish the type of knowledge, allowing the LLM in Stage 3 to explicitly learn to weight and utilize each semantic and spatial component effectively.

4.4 Stage 3: Spatially-Aware LLM Fine-Tuning

The final stage of our pipeline leverages the structured knowledge encoded in S_{prompt} (the output of Stage 2) to fine-tune a Large Language Model for generating spatially-grounded and contextually rich image captions. This stage is critical for translating the multi-faceted knowledge graph into natural, human-readable descriptions that accurately reflect the 3D spatial structure and commonsense context of the scene. The architecture of Stage 3 is illustrated in Figure 4.

Model Selection and Architecture We employ a Text-to-Text Transfer Transformer (T5) [11] as our base language model due to its unified text-to-text framework and strong performance on various NLP tasks. The T5 architecture treats caption generation as a sequence-to-sequence translation task:

$$P(Y|S_{\text{prompt}}) = \prod_{t=1}^{T} P(y_t|y_{<t}, S_{\text{prompt}}; \theta) \tag{13}$$

where $Y = \{y_1, y_2, ..., y_T\}$ is the target caption sequence, S_{prompt} is the structured input from Stage 2, and θ represents the model parameters.

Input Representation and Tokenization The structured sequence S_{prompt} is formatted with special delimiter tokens to guide the model's attention to different knowledge types:

```
[3D_START] car is_closer_than building [3D_END]
[2D_START] person on sidewalk [2D_END]
[CS_START] car UsedFor transportation [CS_END]
[FACT_START] building IsA structure [FACT_END]
```

This explicit structuring allows the model to learn weighted importance of spatial vs. semantic vs. commonsense knowledge during fine-tuning.

Training Objective and Loss Functions We design a multi-component loss function to optimize both generation quality and spatial-semantic alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda_{\text{spatial}}\mathcal{L}_{\text{spatial}} + \lambda_{\text{cs}}\mathcal{L}_{\text{cs}} \tag{14}$$

where:

1. **Generation Loss** ($\mathcal{L}_{\text{gen}}$): Standard cross-entropy loss for caption generation:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log P(y_t^*|y_{<t}^*, S_{\text{prompt}}; \theta) \tag{15}$$

where y_t^* denotes the ground truth token at position t.

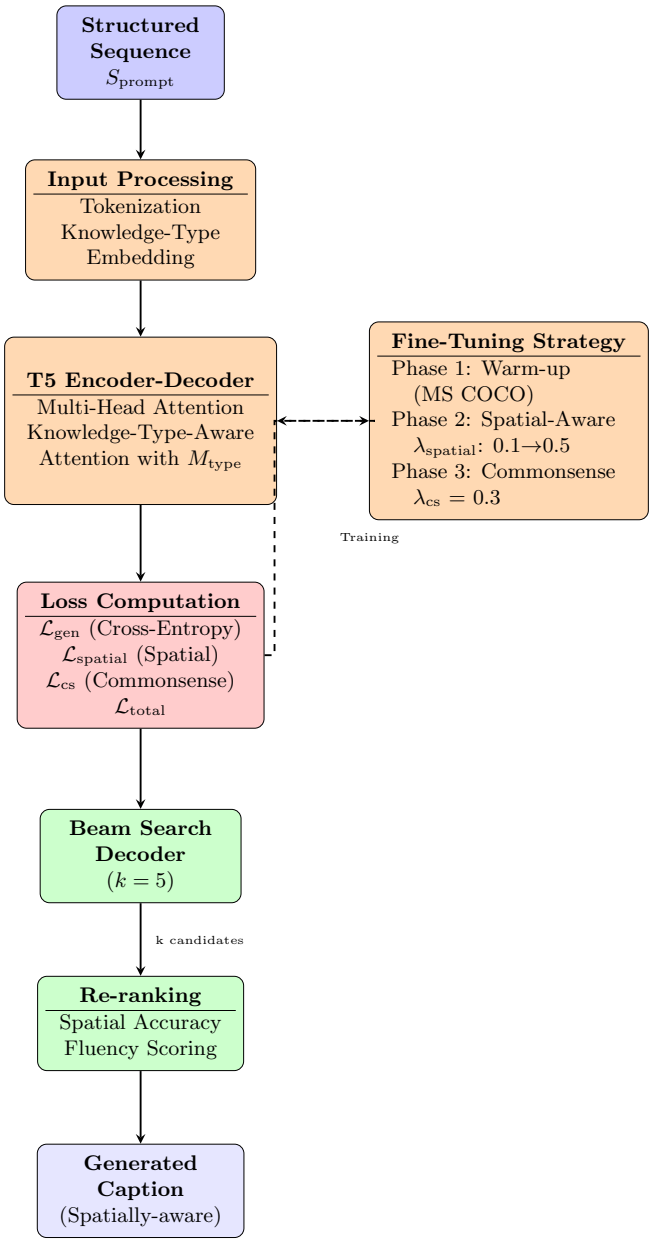

Fig. 4. Architecture of the Spatially-Aware LLM Fine-Tuning Module (Stage 3)

2. **Spatial Consistency Loss** ($\mathcal{L}_{\text{spatial}}$): Enforces that generated spatial terms (e.g., "in front of", "behind", "closer") align with the 3D spatial predicates

in $\mathcal{P}_{3D}$. This is implemented as a binary classification loss over predicted spatial relationships:

$$\mathcal{L}_{\text{spatial}} = -\sum_{i,j} [r_{ij} \log \hat{r}_{ij} + (1 - r_{ij}) \log(1 - \hat{r}_{ij})] \tag{16}$$

where r_{ij} indicates ground truth spatial relationship between objects i and j, and $\hat{r}_{ij}$ is the model's prediction.

3. **Commonsense Grounding Loss** ($\mathcal{L}_{\text{cs}}$): Encourages the model to incorporate relevant commonsense knowledge from $\mathbf{KG}_{\text{CS}}$ by measuring semantic similarity between generated text and commonsense triples:

$$\mathcal{L}_{\text{cs}} = 1 - \frac{1}{|\mathbf{KG}_{\text{CS}}|} \sum_{t \in \mathbf{KG}_{\text{CS}}} \text{sim}_{\text{semantic}}(Y, t) \tag{17}$$

The hyperparameters λ_{spatial} and λ_{cs} control the relative importance of spatial accuracy and commonsense incorporation.

Fine-Tuning Strategy Our fine-tuning procedure consists of three phases:

1. **Warm-up Phase:** We first fine-tune on standard image captioning datasets (e.g., MS COCO [?]) to maintain general caption generation capability. This prevents catastrophic forgetting of linguistic fluency.
2. **Spatial-Aware Phase:** We introduce the spatial consistency loss $\mathcal{L}_{\text{spatial}}$ and fine-tune on depth-annotated image-caption pairs, gradually increasing λ_{spatial} from 0.1 to 0.5 over training epochs.
3. **Commonsense Integration Phase:** We activate $\mathcal{L}_{\text{cs}}$ and fine-tune on the complete **DASG-CS** representation, setting $\lambda_{\text{cs}} = 0.3$ to balance factual grounding with creative description.

Attention Mechanism Enhancement To explicitly leverage the structured input format, we enhance T5's standard multi-head attention mechanism with a knowledge-type-aware attention layer. This layer computes separate attention weights for different knowledge types (3D spatial, 2D semantic, commonsense, factual):

$$\text{Attention}_{\text{type}}(Q, K_{\text{type}}, V_{\text{type}}) = \text{softmax}\left(\frac{QK_{\text{type}}^T}{\sqrt{d_k}} + M_{\text{type}}\right) V_{\text{type}} \tag{18}$$

where M_{type} is a learned mask matrix that adjusts attention weights based on knowledge type, allowing the model to prioritize spatial predicates when generating positional descriptions.

Decoding and Post-Processing During inference, we employ beam search decoding with beam size $k = 5$ to generate diverse caption candidates. We then apply a re-ranking mechanism based on:

$$\text{Score}(Y) = \log P(Y|S_{\text{prompt}}) + \alpha \cdot \text{Spatial\_Acc}(Y) + \beta \cdot \text{Fluency}(Y) \quad (19)$$

where $\text{Spatial\_Acc}(Y)$ measures the consistency between generated spatial terms and $\mathcal{P}_{3D}$, and $\text{Fluency}(Y)$ is computed using a pre-trained language model (e.g., GPT-2) to ensure linguistic naturalness.

The highest-scoring caption is selected as the final output, ensuring both spatial accuracy and linguistic quality.

4.5 Implementation Details and Dataset Preparation

5 Evaluation

5.1 Environment and Experiment Data

6 Future Work

Acknowledgments

The authors would like to thank the Faculty of Information Technology, Ho Chi Minh City University of Industry and Trade, and University of Science, VNU-HCM, which are sponsors of this research. We also thank anonymous reviewers for their helpful comments on this paper.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6077–6086 (2018)
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Cao, C., Song, Y., He, J.: An overview of yolo series: From yolov1 to yolov8. Journal of Physics: Conference Series **2685**(1), 012002 (2023)
4. Chen, Y., Ma, L., Huang, Y., Zhang, S., Qian, Y.: A survey of knowledge graph-based entity linking from text to knowledge graph. IEEE Access **9**, 13000–13020 (2021)
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10578–10587 (2020)
6. Gillies, S.: Shapely: manipulation and analysis of planar geometric objects (2007), python package, available at https://pypi.org/project/Shapely/

7. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3128–3137 (2015)
8. Krishna, R., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
9. Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8928–8937 (2019)
10. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: Using knowledge graphs for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2673–2681 (2017)
11. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
13. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3156–3164 (2015)
14. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies for real-time object detection. arXiv preprint arXiv:2207.02696 (2022)
15. Wang, P., Wu, Q., Shen, C., van den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **40**(10), 2413–2427 (2018)
16. Wu, Q., Shen, C., Wang, P., Dick, A., van den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1367–1381 (2017)
17. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 2048–2057 (2015)
18. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv preprint arXiv:2406.09414 (2024)
19. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: European Conference on Computer Vision (ECCV). pp. 684–699 (2018)
20. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
21. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph generation with global context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5831–5840 (2018)
22. Zhao, W., Wu, B., Ma, S.: Knowledge enhanced fine-grained image captioning. In: ACM International Conference on Multimedia (MM). pp. 1631–1640 (2021)
23. Zhong, Y., Wang, L., Chen, J., Yu, D., Li, Y.: Comprehensive image captioning via scene graph decomposition. In: European Conference on Computer Vision (ECCV). pp. 211–229 (2020)
24. Zhu, Y., Yang, Z., Salakhutdinov, R., Xing, E.P.: Incorporating commonsense knowledge into image captioning via graph convolutional networks. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9008–9017 (2019)