MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

NGUYEN VAN THINH

DEVELOPING DEEP LEARNING NETWORK–BASED METHODS FOR IMAGE CAPTIONING

SUMMARY OF DISSERTATION ON COMPUTER

Major: Computer Science

Code: 9 48 01 01

Ha Noi - 2026

The dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy Science and Technology

Supervisors:

1. Supervisor 1: Assoc. Prof. Dr. Tran Van Lang, Ho Chi Minh City University of Foreign Languages - Information Technology
2. Supervisor 2: Dr. Van The Thanh, Ho Chi Minh City University of Education

Referee 1: Dr. Nguyen Nhu Son

Referee 2: Assoc. Prof. Dr. Le Kim Hung

Referee 3: Dr. Le Quang Minh

The dissertation is examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at at 14:00, April 23, 2026.

The dissertation can be found at:

1. Graduate University of Science and Technology Library
2. National Library of Vietnam

INTRODUCTION

1. Motivation and Significance

Image captioning lies at the intersection of computer vision and natural language processing: models must both *interpret* visual content and *verbalise* it in fluent, coherent sentences. Despite advances in multimodal deep learning, three limitations persist: (i) insufficient exploitation of deep visual features; (ii) weak modelling and use of inter-object relations; and (iii) ineffective integration of external knowledge (e.g., ConceptNet, Abstract Meaning Representation - AMR, large language models - LLMs), yielding captions with limited semantic depth. Developing deep neural approaches that jointly leverage visual cues, relational structure, and external semantic/AMR knowledge is therefore both scientifically warranted and practically relevant.

2. Research Objectives

Objective: Develop deep learning–based captioning methods that fuse visual features, relational graphs, external knowledge, and AMR to improve descriptive accuracy and linguistic naturalness.

The dissertation specifies the research objectives as follows::

(i) Systematically survey existing approaches and identify gaps in relational and semantic modelling.

(ii)Propose a model that exploits relational graphs to enhance scene-structure understanding.

(iii) Design a multi-source fusion mechanism that integrates external knowledge (e.g., ConceptNet, LLMs) with vision–language representations.

(iv) Leverage AMR/AMR-like representations to strengthen abstract semantic understanding.

(v) Validate on MS COCO and Flickr30K using BLEU, METEOR, ROUGE-L, CIDEr, SPICE, and the proposed SCS.

3. Object and Scope of the Study

Object: Automatic generation of natural-language descriptions within an encoder–decoder paradigm, integrating visual features, relational graphs, AMR/AMR-like semantics, and external knowledge; decoding via LSTM/Transformer with attention.

Scope:

- Data: MS COCO, Flickr30K (five human references per image).
- Models: Encoders for objects/relations, AMR/AMR-like, and CLIP-ViT; Transformer/LSTM decoders with cross-fusion and adaptive attention; GPT-based re-ranking at inference.
- Evaluation: BLEU, METEOR, ROUGE-L, CIDEr, SPICE, and SCS; domain-specific settings (e.g., medical, satellite) are out of scope.

4. Research Methods

- Literature review: Taxonomise prior work (CNN-LSTM, relational-graph methods, Transformer-based models, external-knowledge integration, AMR-based approaches) to locate gaps.
- Model/algorithm design: Linearise AMR; map relation graphs to AMR-like structures; compute node embeddings with GraphSAGE; employ cross-attention/cross-fusion and adaptive attention; modulate token probabilities with ConceptNet-derived priors.
- Experimentation and evaluation: Train/test on MS COCO and Flickr30K; conduct quantitative comparisons with baselines and component-wise ablations; provide qualitative analyses.

5. Principal Contributions

This dissertation advances a coherent sequence of four progressively integrated models:

(i) OD-VR-Cap: Combines object-centric features with a relation graph; LSTM decoder with dual attention.

(ii) RGTranCNet: Replaces the LSTM with a Transformer decoder using cross-attention and integrates ConceptNet, yielding marked gains on MS COCO.

(iii) AMR-GT&RG: Leverages AMR from ground-truth captions and AMR-like structures derived from the relation graph; uses GraphSAGE embeddings and ConceptNet priors; introduces the Semantic Consistency Score (SCS); attains state-of-the-art results on MS COCO and Flickr30K.

(iv) CLIP-AMR-GPT: Unifies CLIP-ViT, relation graphs, and AMR/AMR-like semantics via Cross-Fusion and Adaptive Attention; applies GPT-based re-ranking; achieves superior performance, notably on SPICE, CIDEr, and SCS.

6. Dissertation Structure

Five core chapters, in addition to the Introduction, Conclusion, and References, progressing from background to methods and empirical evaluation:

- Introduction
- Chapter 1. Literature review, gap analysis, and problem framing.
- Chapter 2. OD-VR-Cap (object features + relation graph; LSTM with dual attention).
- Chapter 3. RGTranCNet (Transformer decoder with cross-attention; integration of ConceptNet).
- Chapter 4. AMR-GT&RG (AMR/AMR-like semantics + ConceptNet + attention mechanisms).
- Chapter 5. CLIP-AMR-GPT (Cross-Fusion, Adaptive Attention, and GPT-based re-ranking).
- Conclusion and recommendations.

CHAPTER 1. DEEP LEARNING NETWORK–BASED IMAGE CAPTIONING

1.1. Introduction

Image captioning (IC) links computer vision and natural language processing to produce fluent descriptions that are faithful to an image's content - its salient objects and their relations. As a core multimodal task, it underpins assistive technologies, content-based retrieval, medical imaging, behaviour analysis, and automated moderation. Deep learning has driven IC from CNN–LSTM pipelines to Transformers and, recently, large pretrained vision–language models (VLMs). However, most methods still under-capture fine-grained relations and underuse external semantic knowledge, yielding captions that can lack naturalness or contain semantic errors.

1.2. Approaches to Image Captioning

1.2.1.Traditional Methods

Traditional pre–deep learning IC methods fall into two categories:

- Retrieval-based: Match the input image to a labelled image–caption database by visual similarity and return the caption(s) of the nearest neighbour(s), sometimes with light adaptation.
- Template-based: Detect semantic elements (objects, actions, context) and fill hand-crafted or learned syntactic templates to compose a sentence.

These approaches are easy to deploy but generalise poorly: performance hinges on database coverage (retrieval) or template rigidity (generation), limiting adaptation to novel images and linguistic constructions.

1.2.2.Deep Learning–Based Methods

The advent of CNNs and RNNs/LSTMs opened a new era for IC. Early models (Show and Tell; Show, Attend, and Tell) achieved end-to-end captioning with visual attention but still relied mainly on local-region cues and under-modelled object relations.

(i) CNN–LSTM: Works such as NeuralTalk2, NIC, and Soft-Attention apply visual attention within an encoder–decoder setup. Effective for region-level grounding, they nonetheless lack structured relational reasoning.

(ii) Graph-based: Scene Graph Captioning, GCN-LSTM, and OD-VR-Cap encode the object–relation structure via Relation/Scene Graphs, improving scene layout and inter-entity semantics, yet language decoding remains a bottleneck.

(iii) Transformer-based: Transformers strengthen long-range context and token interactions (e.g., X-Transformer, COS-Net, RGTranCNet), but vanilla variants still need external semantic priors to curb "language hallucination".

(iv) External knowledge integration: Integrating ConceptNet, WordNet, or domain ontologies improves semantic fidelity; for instance, RGTranCNet conditions generation on ConceptNet cues to produce richer, context-appropriate text.

(v) AMR-based (Abstract Meaning Representation): AMR represents sentence semantics as a concept–relation graph. Recent models combine AMR from ground-truth captions with AMR-like graphs induced from relation graphs, reinforcing abstract semantic understanding and yielding more natural, coherent, and consistent captions.

1.3. Research Gaps and Directions

1.3.1. Research Gaps

A synthesis of the above lines of work reveals several persistent limitations:

- Insufficient multi-source knowledge integration: visual features, relational structure, external semantics, and AMR are rarely unified within a single framework.

▪ Lack of a unifying mechanism across representational layers: current models do not coherently align global/local vision features with structured semantic formalisms.

▪ Evaluation shortcomings: prevalent metrics (BLEU, CIDEr, SPICE) emphasise n-gram overlap and only partially capture semantic consistency and coherence.

▪ Limited generalisation and rare-object handling: robustness degrades in novel contexts or with infrequent entities.

1.3.2. Research Directions

This dissertation pursues a staged, incremental program of models that inherit and extend one another:

(i) OD-VR-Cap: fuse object features with a relational graph via a dual-attention LSTM to enhance scene-structure understanding.

(ii) RGTranCNet: replace the LSTM with a Transformer decoder and inject ConceptNet knowledge to modulate token-generation probabilities.

(iii) AMR-GT&RG: combine AMR derived from ground-truth captions with AMR-like graphs induced from relational structure; obtain graph embeddings via GraphSAGE.

(iv) CLIP-AMR-GPT: unify CLIP-ViT features, relational-graph embeddings, and AMR/AMR-like semantics through a Cross-Fusion layer with Adaptive Attention, followed by GPT-based re-ranking.

1.4. Experimental Protocol and Evaluation

1.4.1. Datasets

This dissertation employs two widely used benchmarks:

▪ MS COCO: over 120,000 images, each annotated with five human-written captions.

▪ Flickr30K: 31,000 images with linguistically diverse captions spanning everyday contexts.

All experiments follow the Karpathy splits for preprocessing and partitioning to ensure fair and reproducible comparisons with prior work.

1.4.2. Evaluation Metrics

Models are assessed with a comprehensive suite of metrics: BLEU-1/2/3/4, METEOR, ROUGE-L, CIDEr, SPICE, and the newly proposed Semantic Consistency Score (SCS), which quantifies semantic alignment between generated captions and reference annotations beyond surface n-gram overlap.

1.5. Chapter Summary

Chapter 1 surveyed the principal research streams in image captioning, analysed their strengths and limitations, and identified the key scientific gaps that remain. Building on this diagnosis, the dissertation charts a development roadmap for a family of models that hierarchically integrate multiple sources of knowledge - visual features, relational structure, external common sense, and AMR-based semantic representations - to enhance caption quality, semantic fidelity, and generalisation in automatic image captioning.

CHAPTER 2. AN IMAGE CAPTIONING MODEL BASED ON OBJECT RELATIONSHIP GRAPH REPRESENTATION

2.1. Introduction

Most captioning methods rely on global or local features and thus omit explicit modelling of object relations, weakening contextual grounding and limiting relational semantics. This chapter presents OD-VR-Cap, which encodes object–relation structure as a Relational Graph (R-Graph) and fuses it with visual cues via an LSTM decoder with dual attention. The pipeline comprises: (i) ODwGCN for object detection; (ii) relation prediction; (iii) graph encoding; (iv) dual-attention fusion; and (v) LSTM-based sentence generation.

Core contributions: (1) ODwGCN boosts detection by exploiting label co-occurrence priors; (2) the relation predictor leverages knowledge from an entity graph; (3) the R-Graph is normalised to R-Graph* and embedded with GraphSAGE for structure-aware representations; (4) dual (visual + graph) attention guides the LSTM to generate fluent, semantically rich captions that capture inter-object relations.

2.2. Proposed Method

2.2.1.Image Encoder

(i) ODwGCN - Object detection with a GCN prior: ODwGCN exploits label co-occurrence to recalibrate the confidence matrix of a base detector (SSD/Faster R-CNN/YOLOX). The pipeline has two stages: (1) learn a label–correlation graph and node embeddings via GraphSAGE; (2) adjust per-class posterior probabilities using learned weighting vectors derived from these embeddings.

(ii) Triplet classification <subject, predicate, object>: Inter-object relations are modelled as a multi-class classification task over N_{R+1} relation categories (including a "$none$" class). The module consumes features from

the subject–object region pair and their union region, together with relational priors - i.e., embeddings of entity and predicate nodes learned from an Entity Graph (E-Graph) using GraphSAGE;

(iii) Construction and representation of R-Graph / R-Graph*: From predicted region pairs and their relations, we build a directed Relational Graph (R-Graph) whose nodes correspond to object regions. To better interface with the language model, the graph is normalized into R-Graph*: nodes are expanded to include both object labels and relation labels, and edges are routed through predicate nodes to "canonicalize" the structure. Nodes of R-Graph* are embedded in an unsupervised manner with GraphSAGE to produce structure-aware representations Z^*, which are later consumed by the graph-attention branch during decoding.

2.2.2. Language Decoder

- Dual-attention mechanism: At each decoding step ttt, we compute visual attention over object-region features and, in parallel, graph attention over the node embeddings of R-Graph*, yielding two context vectors $c_t^{(v)}$ and $c_t^{(g)}$. Feeding both contexts to the decoder enforces grounding in salient image regions while preserving the relational semantics encoded by the graph.
- LSTM decoder: The LSTM consumes the concatenated input $\left[x_t, c_t^{(v)}, c_t^{(g)}\right]$, updates its hidden state, and produces a next-word distribution via a softmax layer. Training optimises the cross-entropy loss with teacher forcing.

2.3. Experiments

2.3.1. Datasets and experimental setup

- Relational knowledge: Visual Genome is used to learn predicate/entity relations (with preprocessing and consolidation to frequent classes).

▪ Image captioning: MS COCO with the Karpathy split is adopted for train/val/test; each image has five human-written references.

▪ Environment & architecture: Standard backbone detectors serve as the visual front-end; GraphSAGE is applied to both graphs; the captioning head is an LSTM decoder equipped with dual attention (visual + graph).

2.3.2.Evaluation metrics

▪ Object detection: mAP, mAP@0.5, mAP@0.75.

▪ Relationship prediction: Recall@50/100.

▪ Image captioning: BLEU-1/4, METEOR, ROUGE-L, CIDEr, SPICE.

2.3.3.Results

▪ Object detection (ODwGCN): ODwGCN consistently increases mAP across multiple detector backbones, validating the benefit of modelling label co-occurrence priors to recalibrate class confidences.

▪ Visual relationship prediction (VRP+RK): The relation classifier augmented with relational knowledge (RK) and union-box features attains higher Recall@50/100 than baselines under the same settings.

▪ Image captioning (OD-VR-Cap): The complete model surpasses CNN–LSTM attention baselines and several scene-graph–based methods on BLEU-4, METEOR, and CIDEr. Qualitative cases provide more accurate descriptions of actions and relations.

Attribution of gains: Transitioning from flat-region features to a graph-structured representation enables context propagation among objects; the dual-attention (visual + graph) guides the decoder to select region- and relational-level evidence at each step, thereby reducing linguistic hallucination and improving semantic consistency.

2.4. Chapter Summary

This chapter introduced OD-VR-Cap, the foundational model in the dissertation's method series. The approach calibrates object detection with a

GCN-based prior, constructs and embeds a visual relationship graph, and decodes captions using an LSTM with dual attention (visual and graph). Experiments on MS COCO and on visual relationship prediction demonstrate both the effectiveness and scalability of this structure-aware, knowledge-driven paradigm. OD-VR-Cap serves as a springboard for subsequent extensions - replacing the decoder with a Transformer, integrating ConceptNet, leveraging AMR/AMR-like representations, and employing Cross-Fusion with GPT re-ranking - with the overarching goal of enhancing semantic depth and linguistic naturalness in image captioning.

CHAPTER 3. IMAGE CAPTIONING WITH TRANSFORMER AND CONCEPTNET KNOWLEDGE

3.1. Introduction

Motivated by the limitations of LSTM decoders - particularly their difficulty in modelling long-range dependencies - and by the underutilization of external semantic knowledge, this chapter introduces RGTranCNet. The model replaces the LSTM with a Transformer Decoder to capture distant contextual dependencies more effectively, and integrates ConceptNet knowledge to enhance semantic expressiveness and generalization to rare or unseen objects. The core architecture performs multi-source fusion - combining region-level visual features, the relationship graph from Chapter 2, and ConceptNet knowledge - via cross-attention within the Transformer Decoder, thereby producing captions that are coherent, natural, and semantically rich.

3.2. Proposed Image Captioning Method

The framework follows an encoder–decoder design with three parts: (1) Image Encoder - learns visual representations (region features, relational cues) to ground generation; (2) Semantic Extractor - retrieves commonsense about detected objects/relations from ConceptNet to complement visual evidence; (3) Transformer Decoder - ingests both visual features and ConceptNet cues, using cross-attention to align tokens with these sources and improve lexical choice, syntactic fluency, and semantic fidelity.

3.2.1.Image Encoder

We retain the upstream pipeline: **ODwGCN** improves detection mAP; **VRP+RK** predicts pairwise relations to construct the **R-Graph**; the graph is then encoded into node embeddings Z via message passing. For each image I, the encoder outputs:

- F_I: region-level object features;

▪ Z_I: relation-graph embeddings after semantic propagation.

Both matrices are provided to the decoder.

3.2.2. Object-Centric Semantic Knowledge Extractor

ConceptNet is modelled as a knowledge graph $K = (V, E, W)$. For each detected object label, we query the top-k related concepts (e.g., *IsA, PartOf, UsedFor*, …) along with their confidence weights, forming a set OOO (related concepts with weights). During generation, O is used to adjust the token probability distribution, increasing the likelihood of contextually appropriate, knowledge-consistent words.

3.2.3. Language Decoder

▪ Single-step Multi-Head Cross-Attention fusion: The query is derived from the partially generated caption, while the keys/values come from F_I (visual features) and Z_I (relation graph embeddings). The two attention outputs are linearly combined with learnable gating coefficients to form a unified contextual representation for decoding.

▪ ConceptNet-guided logit adjustment: After the prediction layer, the logits of vocabulary items appearing in the retrieved concept set O are augmented by βw (with w the concept relevance weight), thereby biasing the decoder toward tokens that are semantically consistent with the visual context and commonsense knowledge.

3.3. Experiments and Results

3.3.1. Datasets and Experimental Setup

We evaluate on MS COCO using the Karpathy split. Relation graphs and their embeddings are inherited from Chapter 2. The captioning module employs a Transformer Decoder with six blocks, eight heads, and d_model=512, optimised with Adam. ConceptNet is queried by detected object labels to retrieve semantically related concepts, which are injected into the decoder during both training and inference.

3.3.2.Results and Discussion

Ablation (w/ vs. w/o ConceptNet): Without external knowledge, RGTran attains BLEU-1 77.5 / BLEU-4 34.9 / METEOR 28.3 / ROUGE-L 55.3 / CIDEr 98.4 / SPICE 18.7. Incorporating ConceptNet, RGTranCNet improves across the board to 79.8 / 36.3 / 35.6 / 57.2 / 107.8 / 20.5, respectively. The gains - +7.3 METEOR, +9.4 CIDEr, and +1.8 SPICE - indicate stronger semantic adequacy and closer alignment with human references.

Comparison with OD-VR-Cap (LSTM + dual attention): RGTran and RGTranCNet surpass OD-VR-Cap on all metrics; even without ConceptNet, RGTran delivers +13.3 CIDEr, owing to the Transformer decoder and unified cross-attention. Incorporating ConceptNet further boosts generalisation to rare/novel objects and strengthens relational reasoning.

Interpretation: Gains arise from: (i) self/cross-attention for long-range dependencies and global context; (ii) single-step fusion of region features and relation-graph embeddings, which coordinates signals more effectively than separate attentions; and (iii) ConceptNet-guided logit adjustments that curb language hallucination and improve semantic naturalness and diversity.

3.4. Chapter Summary

RGTranCNet demonstrates that replacing the LSTM decoder with a Transformer and integrating ConceptNet yields an effective and scalable approach to image captioning. The model surpasses OD-VR-Cap and multiple baselines across BLEU, METEOR, ROUGE-L, CIDEr, and SPICE, while also improving the semantic fidelity of generated sentences. This model lays the groundwork for Chapter 4, where the dissertation further deepens semantic modelling by incorporating AMR/AMR-like representations into the pipeline.

CHAPTER 4. INCORPORATING AMR INTO TRANSFORMER FOR IMAGE CAPTIONING

4.1. Introduction

Recent image captioning models - particularly Transformer-based approaches - have achieved substantial gains, yet they still fall short in capturing abstract semantics and more profound relational logic. Building on RGTranCNet (Chapter 3), this chapter proposes AMR-GT&RG, which integrates Abstract Meaning Representation (AMR) into the Transformer decoder and unifies three layers of knowledge: (1) Visual (object-level region features); (2) Relational structure (a relation graph converted to an AMR-like graph), and (3) Abstract semantics (AMR parsed from reference captions, i.e., AMR-GT), augmented with external knowledge (ConceptNet).

This unified framework enables the model to understand, reason, and express at the levels of concepts–actions–relations, yielding captions that are more natural, semantically consistent, and better generalised.

4.2. Proposed Image Captioning Method

4.2.1.Architecture

(i) Image Encoder: Detect object regions to obtain F_I; construct a relation graph and compute GNN-based embeddings to obtain $E_{RG,I}$.

(ii) Abstract Semantic Extractor.

- AMR-GT: Convert reference captions → AMR →PENMAN → BERT to yield $E^{GT}_{AMR,I}$ (used only during training).
- AMR-like (AMR-RG): Map the image relation graph to an AMR-like graph via concept/relation rules; embed the graph with GraphSAGE to obtain $E^{RG}_{AMR,I}$.
- Semantic Knowledge Extractor: Query ConceptNet to collect a weighted set of related concepts C_I for probability adjustment.

(iii) Transformer Language Decoder

▪ Masked Multi-Head Self-Attention conditioned on AMR: Inject $E_{AMR,I}^{GT}$ into the self-attention branch to infuse abstract semantics at each decoding step.

▪ Cross-Modal Attention fusing three sources: Attend jointly over F_I; $E_{RG,I}$; $E_{AMR,I}^{GT}$ (at inference, E$E_{AMR,I}^{GT}$ is omitted).

▪ ConceptNet-based logit adjustment: Add βw to the logits of vocabulary items present in C_I (with relevance weight w), biasing the decoder toward contextually and semantically appropriate words.

4.2.2. Training and Inference Algorithms

▪ TrainTransformerDecoder: Auto-regressive training with (i) self-attention augmented by AMR-like signals, (ii) multi-source cross-attention, and (iii) ConceptNet-based logit adjustment; parameters are optimised using cross-entropy.

▪ GenerateCaptionAMR: At inference, the model uses $\{F_I, E_{RG,I}, E_{AMR,I}^{RG}, C_I\}$ only, and decodes with beam search or greedy strategy.

4.2.3. New Semantic Metric- SCS

$$SCS = cosine\left(SBERT(S_{gen}), SBERT(S_{ref})\right) \in [0,1]$$

Purpose: quantify semantic consistency beyond n-gram overlap, addressing limitations of BLEU/METEOR/ROUGE.

4.3. Experiments and Results

4.3.1. Experimental Setup

▪ Datasets: MS COCO and Flickr30K (Karpathy split); Visual Genome is used to support relational supervision.

▪ Encoder: For COCO, ODwGCN is employed; for Flickr30K, Faster R-CNN features are combined with GCN/GraphSAGE to build and embed relation graphs.

▪ Decoder: Transformer decoder with 6 layers and 8 heads; dmodel=768; Adam optimizer, learning rate = 0.00004; batch size 32; experiments run on NVIDIA T4/Colab.

▪ AMR Processing: Ground-truth captions are parsed by NeuralAMR → linearised in PENMAN → embedded with BERT (AMR-GT). Image relation graphs are converted to AMR-like structures and embedded via GraphSAGE (AMR-RG).

▪ ConceptNet Integration: For each detected object, retrieve the top-k most relevant concepts/relations to form the auxiliary knowledge set.

4.3.2.Results and Discussion

MS COCO: AMR-GT&RG attains B@1 = 81.2, B@4 = 39.5, METEOR= 37.2, ROUGE-L = 59.9, CIDEr = 136.7, SPICE = 25.1, and SCS = 89.1. Compared with recent SOTA systems (e.g., COS-Net, X-Transformer, SGAE, MLA-LRN), the model excels on metrics emphasising semantic adequacy and coherence (METEOR, ROUGE-L, CIDEr, SPICE, SCS). Notably, relative to RGTranCNet (Chapter 3), AMR-GT&RG improves CIDEr by +28.9, SCS by +6.8, and METEOR by +1.6, evidencing the clear benefit of the AMR layer for strengthening semantic consistency and alignment with human references.

Flickr30K: On Flickr30K, AMR-GT&RG achieves B@1 = 79.1, B@4 = 36.4, METEOR = 35.6, ROUGE-L = 56.7, CIDEr = 94.5, SPICE = 22.7, and SCS = 87.2. The model again leads on semantics-focused metrics (METEOR, CIDEr, SPICE, SCS), confirming robust cross-domain generalisation and stable caption quality across Flickr30K's diverse linguistic styles.

Ablation on MS COCO: Component analysis shows both AMR sources are beneficial, but their combination is optimal: using AMR-GT only yields CIDEr = 132.5, SPICE = 24.6, SCS = 87.3; AMR-RG only yields CIDEr =

129.8, SPICE = 23.9, SCS = 84.5; the AMR-GT&RG combination achieves the highest scores across all metrics. This confirms their complementarity: language-derived AMR contributes abstract, general knowledge, while image-derived AMR-like enforces fidelity to visual content and inter-object relations.

4.4. Chapter Summary

The proposed AMR-GT&RG model integrates AMR representations into a Transformer decoder via a multi-source fusion mechanism that unifies visual features, relational graphs, AMR embeddings, and external knowledge from ConceptNet, thereby propagating consistent semantics throughout the entire pipeline. Consequently, the generated captions exhibit greater semantic depth and textual coherence, with marked gains on meaning-oriented metrics - CIDEr, SPICE, and SCS - across both MS COCO and Flickr30K. This approach also reduces reliance on superficial n-gram overlap and improves generalisation to rare entities/relations.

Limitations remain: the model currently depends on AMR derived from ground-truth captions during training and incurs nontrivial costs for AMR parsing/embedding. Future directions include leveraging pretrained vision–language models (e.g., CLIP, BLIP, BLIP-2) to reduce reliance on gold captions; broadening external knowledge sources (e.g., Wikidata, WordNet, LLMs) to enhance coverage of concept–relation pairs; and extending to multilingual and domain-specific settings to increase practical applicability.

CHAPTER 5. AN IMAGE CAPTIONING MODEL BASED ON MULTIMODAL SEMANTIC FUSION AND GPT RE-RANKING

5.1. Introduction

This chapter presents CLIP-AMR-GPT, addressing three gaps: (i) the lack of a flexible fusion mechanism for heterogeneous knowledge (visual cues, relational structure, and abstract AMR semantics); (ii) limited semantic depth and unstable linguistic coherence; and (iii) no high-level language quality control. The model follows an encoder - decoder design that unifies pretrained CLIP vision–language features, relational-graph/AMR-like embeddings, and AMR parsed from ground-truth captions. Two complements are introduced: Adaptive Attention, which dynamically gates AMR-like signals during decoding, and GPT-based re-ranking, which selects the final caption to maximise naturalness and coherence.

5.2. Proposed Image Captioning Method

5.2.1. Architectural Overview

CLIP-AMR-GPT adopts a knowledge-unified encoder–decoder. The encoder yields (a) CLIP-ViT features - one global [CLS] plus 256 patch tokens; (b) R-Graph embeddings via GraphSAGE; and (c) AMR-GT embeddings derived by NeuralAMR → PENMAN → BERT. The Transformer decoder employs Cross-Fusion Attention to integrate {CLIP, R-Graph, AMR-GT} and Adaptive Attention to inject AMR-like signals (embedded from the R-Graph) at the right timesteps. After beam search, **GPT re-ranking** selects the caption with the highest GPTScore.

5.2.2. Image Encoder

Images are processed by CLIP ViT-L/14 to obtain both a global vector and patch-level representations, enabling the model to capture holistic context and fine-grained details. In parallel, an object detector paired with relation inference (trained on Visual Genome) constructs an R-Graph;

bidirectional GraphSAGE then learns node/relation embeddings $E_{RG,I}$, supplying semantic structure in terms of roles–entities–relations.

5.2.3.Abstract Semantic Extraction

Ground-truth captions are parsed into AMR and embedded to yield $E_{AMR,I}^{GT}$, which serve as stable semantic priors during training. From the R-Graph, we derive an AMR-like graph (concepts/relations aligned to AMR conventions) and embed it with GraphSAGE to obtain EAMR, $E_{AMR,I}^{RG}$, a vision-grounded semantic source used at inference. The two AMR sources are complementary: AMR-GT steers conceptual abstraction, while AMR-like injects image-consistent structure token-by-token.

5.2.4.Cross-Fusion Attention

Instead of summation or averaging, the model preserves all tokens/nodes from the three modalities. Each modality is projected to its own Key–Value space and then concatenated into K_{multi}, V_{multi}. Decoder queries learn to selectively weight information from {CLIP, R-Graph, AMR-GT} conditioned on the current context, avoiding detail loss and enabling cross-modal compensation.

5.2.5.Adaptive Attention cho AMR-like

To prevent uniformly forcing AMR-like influence, the decoder computes a time-dependent gate $g_t = \sigma(W_g \cdot h_{t-1} + b_g)$. Each node e_i in $E_{AMR,I}^{RG}$ is re-weighted by $\mathrm{g_t}$ before entering the masked multi-head attention. When generating verbs, the model emphasises relation/action nodes; when generating nouns, it emphasises entity/attribute nodes. This mechanism reduces noise and improves coherence and grounding.

5.2.6.GPT Re-ranking

The decoder produces k candidate captions via beam search. Each candidate is scored using GPTScore (length-normalised log probability under GPT-2 medium), and the highest-scoring caption is selected. This acts

as a language filter, improving fluency, grammar, and discourse connectivity without altering the primary training objective.

5.2.7. Training and Deployment

Training uses cross-entropy with teacher forcing and the Adam optimiser; the decoder has N=6 blocks, eight heads, and d=512. Beam size is 5; re-ranking uses GPT-2 medium. The system runs reliably on an NVIDIA T4. It is readily extensible - backbones (CLIP/LLMs) can be swapped, external knowledge (Wikidata/WordNet/LLMs) can be incorporated, and multilingual/domain-specific adaptation is supported.

5.3. Experiments and Results

5.3.1. Experimental Setup

We evaluate on MS COCO and Flickr30K using the same splits and preprocessing as in Chapter 4 (Karpathy protocol). Metrics include BLEU-1/4, METEOR, ROUGE-L, CIDEr, SPICE, and the proposed Semantic Consistency Score (SCS). Implementation is in PyTorch 2.0 with Python 3.9; decoding uses beam size = 5 with GPT-2 medium for re-ranking.

5.3.2. Results and Discussion

MS COCO (Karpathy split): CLIP-AMR-GPT attains B@1 82.9, B@4 41.4, M 38.5, R 61.8, C 144.2, S 26.7, SCS 91.7, topping most columns and surpassing AMR-GT&RG (C 136.7; S 25.1; M 37.2) and CLIP-Captioner (C 139.4; S 23.9; M 30.0). Relative to AMR-GT&RG, CIDEr improves by +7.5, SPICE by +1.6, and SCS by +2.6, confirming the benefit of multi-source fusion, adaptive regulation, and re-ranking.

Flickr30K: The model achieves B@1 80.5, B@4 38.2, M 36.9, R 58.2, C 102.8, S 24.0, SCS 89.4, again outperforming AMR-GT&RG (B@4 36.4; M 35.6; C 94.5; S 22.7). Stable gains on a smaller, linguistically diverse corpus indicate strong generalisation.

CLIP-AMR-GPT excels by tightly coupling semantic grounding with linguistic fluency. CLIP supplies rich visual concepts (notably for rare objects/relations); AMR/AMR-like injects role–action–entity structure, curbing missing or misattributed relations; Adaptive Attention schedules source contributions to prevent modality bias; and GPT re-ranking polishes grammar and naturalness. This multi-source synergy drives consistent gains on CIDEr/SPICE/METEOR (semantic quality) and BLEU/ROUGE (n-gram/structure).

5.4. Conclusion

This chapter presented CLIP-AMR-GPT, a multi-source semantic unification model that pairs Cross-Fusion and Adaptive Attention with GPT-based re-ranking. It outperforms state-of-the-art baselines on MS COCO and Flickr30K across key metrics (notably CIDEr/SPICE/METEOR/SCS). Ablations confirm each component's essential contribution.

Limitations: (i) Reliance on AMR parsed from reference captions during training; (ii) cost of constructing/embedding AMR and R-Graphs; (iii) slight inference latency from re-ranking.

Future directions: (1) Replace or augment the backbone with stronger vision–language LLMs (e.g., newer CLIP, BLIP/BLIP-2, OFA, LLaVA) and exploit large knowledge graphs (Wikidata, WordNet) and LLMs as soft knowledge; (2) extend to multilingual and multi-domain settings (healthcare, cartography, industrial) via continued domain pretraining; (3) cut AMR overhead with faster/approximate parsers or by learning latent AMR from image–text pairs; (4) explore joint semantic–visual re-ranking (CLIPScore + GPTScore) and enforce consistency between predicted AMR and generated captions.

CONCLUSION

This section summarises the dissertation, from its motivation, objectives, scope, and methodology to the review of existing approaches and the proposal of four progressively developed image captioning models. It also highlights the main results, limitations, open issues, and future research directions.

1. Summary of research and key results

The dissertation addresses image captioning through four models: OD-VR-Cap, RGTranCNet, AMR-GT&RG, and CLIP-AMR-GPT. These models progressively exploit relationship graphs, ConceptNet knowledge, AMR/AMR-like representations, CLIP visual–language features, and GPT-based re-ranking to enhance semantic representation and caption quality. They are evaluated on MS COCO and Flickr30K using BLEU, METEOR, ROUGE-L, CIDEr, SPICE, and SCS.

The results show consistent improvements across model generations. On MS COCO, CIDEr increases from 85.1 for OD-VR-Cap to 107.8 for RGTranCNet, 136.7 for AMR-GT&RG, and 144.2 for CLIP-AMR-GPT, with SCS reaching 91.7. On Flickr30K, CLIP-AMR-GPT achieves B@1 = 80.5, B@4 = 38.2, METEOR = 36.9, ROUGE-L = 58.2, CIDEr = 102.8, SPICE = 24.0, and SCS = 89.4. These results confirm the effectiveness of multi-source knowledge integration and the generalization capability of the proposed model.

2. Limitations

Despite these results, several limitations remain. Most experiments are reported from a single main training run, without multiple independent runs for computing mean scores, standard deviations, or statistical significance. The evaluation is limited to MS COCO and Flickr30K, without extension to NoCaps, out-of-domain, zero-shot, or Vietnamese image captioning settings.

In addition, AMR representation quality depends on AMR parsing tools, while models integrating multiple knowledge sources require higher computational costs than simpler encoder–decoder architectures.

3. Open challenges

Future research may focus on enhancing deep semantic reasoning, including intentions, causal relations, and complex contexts; improving robustness to rare objects, out-of-domain data, and zero-shot scenarios through CLIP, LLMs, and external knowledge; standardizing semantic evaluation metrics such as SCS with both automatic and human evaluation; extending image captioning to Vietnamese and other low-resource settings; and improving interpretability, computational efficiency, and real-world deployment in assistive technologies, education, digital archiving, multimodal retrieval, and intelligent visual assistants.

4. Final remarks

In summary, the dissertation establishes a coherent research trajectory from visual relationship modelling and external knowledge integration to AMR-based semantic representation and controlled multimodal fusion with GPT-based re-ranking. The proposed approaches improve captioning accuracy, semantic depth, and linguistic naturalness, while clarifying the role of each component in the overall architecture. These contributions provide a foundation for developing more semantically rich, generalizable image captioning models for practical applications in both national and international contexts.

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

[1] **Nguyen Van Thinh**, Tran Van Lang, Van The Thanh (2024). *OD-VR-Cap: Image captioning based on detecting and predicting relationships between objects*. Journal of Computer Science and Cybernetics, 40(4), 326-345, DOI: 10.15625/1813-9663/20929.

[2] **Nguyen Van Thinh**, Tran Van Lang, Van The Thanh, Tran Huu Quoc Thu, Le Thi Vinh Thanh (2024). *ViT-Trans-AMR: Enhancing image captioning efficiency with AMR Semantic Graphs and Transformer Networks*. Proceeding of The 17th National Conference on Fundamental and Applied IT Research, FAIR'2024, Ha Noi, Aug 08-09,2024. ISBN: 978-604-357-304-6, Natural Science and Technology Publishing House, DOI:10.15625/vap.2024.0289, pp.878-888..

[3] **Nguyen Van Thinh**, Tran Van Lang, Tran Huu Quoc Thu, Nguyen Thi Ngoc Hoa (2024). *Enhancing the efficiency of image annotation using Transformer networks and the ConceptNet Knowledge Base*. Proceedings of the National Workshop on Selected Issues in Information and Communication Technology, Nha Trang, 11-12/10/2024, ISBN: 978-604-67-3029-3, Science and Technics Publishing House. pp.404-409.

[4] **N. V. Thinh**, T. V. Lang, and V. T. Thanh (2026). *RGTranCNet: Effective Image Captioning Model using Cross-Attention and Semantic Knowledge.* Vietnam Journal of Science and Technology, vol. 64, no. 1, 2026, DOI: https://doi.org/10.15625/2525-2518/22381 (**Scopus-Q4**).

[5] **N. V. Thinh**, T. V. Lang and V. T. Thanh (2025). *Integrating Abstract Meaning Representation to Enhance Transformer-Based Image Captioning*, in IEEE Access, vol. 13, pp. 112528-112551, 2025, DOI: 10.1109/ACCESS.2025.3584128 (**SCIE-Q1**).

[6] **N. V. Thinh**, T. V. Lang, and N. M. Hai (2025). *CLIP-AMR-GPT: Enhancing Image Captioning via Cross-Modal Semantics Fusion and GPT-Based Re-Ranking*. Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2025. Lecture Notes in Computer Science, vol 16354. Springer, Singapore. https://doi.org/10.1007/978-981-95-4960-3_17 (**Scopus, Q2**).