Adaptive Semantics Through Cross-Level Contextual Reasoning for Remote Sensing Imagery

Camille Dupuis *, Elodie Fairchild, Hugo Lemaire

Posted Date: 14 October 2025

doi: 10.20944/preprints202510.1039.v1

Keywords: remote sensing image captioning; hierarchical attention; instance-level semantics; visual reasoning; cross-scale representation

Article

Adaptive Semantics Through Cross-Level Contextual Reasoning for Remote Sensing Imagery

Camille Dupuis *, Elodie Fairchild and Hugo Lemaire

Université libre de Bruxelles, Brussels, Belgium
* Correspondence: camille.dupuis@ulb.be

Abstract

The task of describing remote sensing imagery with natural language poses unique challenges due to the complex spatial distributions, vast semantic diversity, and varying scales of ground objects. Existing attention-based models typically rely on spatial attention over a uniform feature grid, which severely constrains the expressiveness and adaptiveness of semantic reasoning. Such fixed-grid mechanisms fail to capture subtle structures of small or irregular instances and neglect the multi-level semantic correlations across different visual hierarchies. To overcome these limitations, we introduce the **Hierarchical Instance-Driven Captioning Network (HIDCap)**, a novel framework that adaptively learns to represent, align, and interpret multi-scale visual semantics for remote sensing image captioning. Our approach integrates two key components: (1) an *instance-centric multi-hierarchy feature encoder* that jointly models object-level, region-level, and global-level representations to preserve fine-grained spatial cues and contextual dependencies; and (2) a *cross-level contextual attention mechanism* that dynamically selects relevant semantic hierarchies at each decoding step, enabling the model to attend to salient instances and contextual backgrounds adaptively. This multi-hierarchy reasoning allows HIDCap to flexibly describe both dense urban areas and homogeneous landscapes by balancing instance-specific and holistic semantic cues. Comprehensive experiments conducted on benchmark remote sensing datasets demonstrate that our proposed framework significantly surpasses previous attention-based methods in terms of both quantitative metrics and qualitative interpretability. The proposed hierarchical instance reasoning paradigm opens new perspectives for bridging multi-scale visual understanding and linguistic generation in remote sensing captioning tasks.

Keywords: remote sensing image captioning; hierarchical attention; instance-level semantics; visual reasoning; cross-scale representation

1. Introduction

Remote sensing imagery provides rich and multi-scale information about the Earth's surface, enabling applications across domains such as environmental monitoring, land-use analysis, urban planning, and disaster management. Traditional remote sensing vision tasks, including object detection, classification, segmentation, and change detection, primarily focus on pixel-level or object-level analysis [1–3]. While these methods have achieved remarkable progress, they inherently provide limited insight into the high-level semantics or relational understanding of scenes. As a result, the expressive gap between low-level visual perception and high-level semantic interpretation remains largely unresolved.

To bridge this gap, *remote sensing image captioning* has emerged as an important research direction that aims to generate descriptive natural language sentences for aerial or satellite images. Unlike conventional image understanding tasks, captioning not only involves identifying visual entities but also requires inferring their attributes, relationships, and spatial arrangements in a linguistically coherent way. The task therefore represents a challenging intersection between computer vision and natural language processing [6]. Early captioning models were dominated by rule-based templates [4] or retrieval-based strategies [5], which lacked flexibility and generalization capability. The introduction

 doi:10.20944/preprints202510.1039.v1

of encoder-decoder frameworks [6] enabled end-to-end learning, where convolutional encoders extract image features and recurrent or transformer-based decoders generate textual descriptions.

The advent of visual attention mechanisms [7] marked a turning point, allowing models to dynamically attend to different regions in the image during sentence generation. Attention mechanisms improved semantic alignment between visual and linguistic modalities, but most existing designs still suffer from fundamental limitations in the context of remote sensing. Specifically, spatial attention computed on fixed-resolution feature maps is coarse-grained and lacks the granularity to accurately focus on small, irregular, or contextually important targets such as vehicles, ships, or building clusters. Moreover, such grid-level attention operates at a single semantic scale, making it insufficient for scenes that vary drastically in spatial composition—from homogeneous areas like deserts and oceans to heterogeneous urban environments.

To address these inherent challenges, recent research has explored multi-scale or region-based attention models for remote sensing captioning [2,3,8]. For example, Qu *et al.* [8] pioneered the use of multimodal encoder-decoder networks for generating textual descriptions, while Zhang *et al.* [2] introduced multi-scale cropping to adapt to varying object sizes. Lu *et al.* [1] emphasized the importance of spatial attention in improving semantic coverage. However, despite these advances, a critical problem persists: existing models lack explicit *instance-awareness* and cannot effectively integrate multiple semantic hierarchies into a unified reasoning framework. This limitation leads to insufficient recognition of small-scale targets, inconsistent focus across hierarchical semantics, and weak adaptability to highly diverse remote sensing scenes.

In this paper, we propose a new paradigm, termed the **Hierarchical Instance-Driven Captioning Network (HIDCap)**, to enhance the semantic reasoning capacity of remote sensing image captioning systems. Unlike prior works constrained by fixed-scale attention, HIDCap introduces an adaptive, multi-hierarchy representation that explicitly models instance-level and contextual dependencies.

Our framework makes the following key contributions:

1) Instance-centric hierarchical representation. HIDCap introduces a unified visual encoder that jointly learns fine-grained instance representations, intermediate regional descriptors, and global context embeddings. By integrating object-level and contextual semantics, the model captures both spatial precision and relational dependencies, enabling accurate grounding of linguistic expressions in visual evidence.

2) Cross-hierarchy contextual attention. We design a novel cross-level attention mechanism that dynamically aggregates information from different semantic hierarchies during decoding. At each generation step, the model adaptively selects which visual hierarchy—instance, region, or global—to emphasize, based on the linguistic context of the partially generated sentence. This dynamic attention selection enhances the robustness and expressivity of semantic alignment across diverse remote sensing scenarios.

3) Generalization across scene complexities. The proposed hierarchical reasoning architecture demonstrates superior adaptability to both simple and complex scenes, effectively capturing semantic nuances under varying densities, textures, and spatial layouts.

Through extensive experiments on benchmark datasets, we validate that HIDCap significantly outperforms conventional attention-based captioning models in both objective metrics (e.g., BLEU, METEOR, CIDEr) and qualitative coherence. Furthermore, visual attention analyses confirm that HIDCap can localize semantically relevant regions and maintain contextual consistency throughout the generation process. This work redefines remote sensing image captioning as a hierarchical and instance-driven reasoning problem, emphasizing the integration of adaptive attention and semantic abstraction. By bridging fine-grained object-level perception with global contextual understanding, HIDCap establishes a scalable foundation for the next generation of multimodal remote sensing interpretation systems.

2. Related Work

Research on remote sensing image captioning has evolved rapidly alongside advances in multimodal representation learning, attention mechanisms, and large-scale visual–language modeling.

 doi:10.20944/preprints202510.1039.v1

This task inherits much from general image captioning paradigms, yet poses unique challenges due to the fine-grained spatial complexity, scale variation, and semantic heterogeneity of aerial imagery. Following the broader taxonomy in multimodal captioning [17,19,26,28,35], approaches can be grouped into two main lines: *bottom-up* frameworks emphasizing region-level visual grounding and instance localization, and *top-down* architectures [34,42,45,51,54] that prioritize global semantics and contextual reasoning. Our work is more closely aligned with the latter, introducing hierarchical and instance-aware mechanisms to model cross-level dependencies across multiple spatial resolutions.

2.1. Remote Sensing Image Captioning Paradigms

Early works in remote sensing captioning primarily followed encoder–decoder frameworks adapted from natural image captioning [42], where convolutional encoders extract global visual representations and recurrent decoders generate descriptive sentences. Qu *et al.* proposed the first multimodal neural model for aerial imagery, confirming the feasibility of synthesizing natural language descriptions for satellite scenes. Subsequent studies integrated attention modules to highlight informative regions or objects during decoding, leading to the widespread adoption of attention-enhanced architectures [49]. For instance, Lu *et al.* introduced spatial attention to adaptively focus on discriminative regions, while Zhang *et al.* further developed multi-scale cropping mechanisms that adapt to various object sizes and improve instance coverage.

More recent advances have explored graph-based and relational modeling in the remote sensing context. Inspired by relational captioning frameworks [52], some methods incorporate object co-occurrence graphs or semantic dependency structures to encode spatial relations between buildings, roads, and vegetation. Others have begun to integrate multi-resolution feature pyramids and transformer-based decoders [22], allowing simultaneous modeling of local and global semantics. Despite these efforts, most approaches remain limited by their reliance on single-scale attention grids and lack explicit mechanisms to integrate cross-hierarchical reasoning—a gap this paper seeks to address through the proposed hierarchical instance-driven framework.

2.2. Attention Mechanisms for Visual–Language Alignment

Attention mechanisms form the cornerstone of modern captioning systems by providing adaptive alignment between vision and language modalities. In remote sensing, attention helps distinguish salient ground objects from vast backgrounds, mitigating issues of spatial redundancy. Foundational attention frameworks [12,25] introduced bottom-up attention using region proposals from object detectors, improving precision in identifying targets like ships or airplanes. However, in high-resolution satellite imagery, such region proposals may fail to capture irregular boundaries or clustered targets.

Building upon classical spatial and channel attention formulations [23,27,53], recent methods have emphasized multi-level attention hierarchies, where the model dynamically fuses global context with instance-level information. For example, adaptive attention architectures [21,37] refine the focus of the decoder across scales, yielding smoother transitions between small and large structures. Some transformer-based models [22] further enhance interpretability by constructing attention maps that correspond to geographic features, demonstrating the value of hierarchical contextualization.

The present work extends this line by introducing a cross-hierarchy attention mechanism that explicitly integrates multiple semantic levels—instance, region, and scene—allowing the caption generator to flexibly shift attention based on linguistic and visual cues. Such a design moves beyond conventional single-level alignment toward a more unified representation of spatial semantics.

2.3. Reinforcement Learning and Metric-Driven Optimization

As with natural image captioning, the discrepancy between maximum-likelihood training and non-differentiable evaluation metrics (e.g., BLEU [36], CIDEr [41], METEOR [14]) has led to the application of reinforcement learning (RL) techniques in remote sensing captioning. Self-critical sequence training (SCST) [39] has been particularly influential, leveraging reward baselines to reduce variance and align optimization with metric performance. Extensions such as SPIDEr [32] and temporal

 doi:10.20944/preprints202510.1039.v1

difference learning [15] provide better reward shaping by incorporating longer-term dependencies and human judgment signals.

In the context of remote sensing, RL-based captioners have demonstrated improved lexical diversity and descriptive accuracy, particularly in scenes with variable object density. However, most methods remain restricted to coarse-grained rewards that fail to capture fine-grained spatial relevance—a key motivation behind the instance-aware and cross-hierarchy reasoning introduced in our work.

2.4. Hierarchical and Structure-Aware Representations

A parallel strand of research explores hierarchical representation learning as a means to encode complex visual structures. Hierarchical captioning models [23,53] employ multi-layer attention or graph neural networks to capture both local and contextual semantics. These ideas have inspired efforts in remote sensing that integrate semantic segmentation, object detection, and region aggregation into unified architectures. Similarly, relational transformers [22,27] and structure-preserving encoders [52] have enhanced compositional reasoning, offering interpretable links between visual instances and linguistic expressions.

Our proposed Hierarchical Instance-Driven Captioning Network (HIDCap) builds upon this foundation by explicitly coupling multi-hierarchy representations with adaptive cross-level attention. Unlike prior work that processes hierarchical features independently, HIDCap enables joint reasoning across semantic layers, ensuring consistent attention distribution across fine-grained and coarse-grained levels. This integration is essential for capturing small targets in vast scenes and for preserving contextual coherence across heterogeneous regions.

In summary, the evolution of remote sensing image captioning reflects three converging research threads: (1) the refinement of visual–language alignment through attention mechanisms, (2) the introduction of metric-aligned optimization via reinforcement learning, and (3) the emergence of hierarchical and relational reasoning frameworks. Our proposed HIDCap framework synthesizes these advances by embedding instance-level awareness into a cross-hierarchy attention architecture, bridging the granularity gap between local details and global semantics. Through this unified design, our model offers a scalable and interpretable approach to generating rich, semantically grounded descriptions for complex remote sensing imagery.

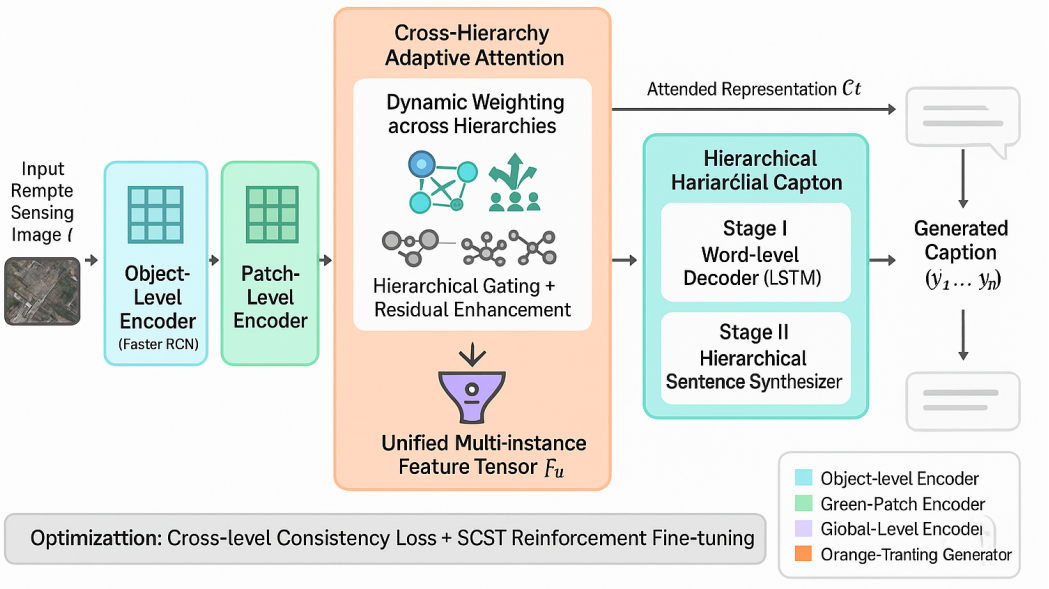

Figure 1. The schematic overview representation of the entire framework.

 doi:10.20944/preprints202510.1039.v1

3. Methodology

In this section, we elaborate on the architecture and underlying principles of the proposed **Hierarchical Instance-Driven Captioning Network (HIDCap)**, which is designed to bridge the gap between multi-scale semantic perception and fine-grained linguistic generation in remote sensing image captioning. HIDCap introduces a comprehensive encoding-decoding pipeline that integrates three complementary stages: (1) multi-instance-level feature extraction for constructing a semantically complete representation space, (2) cross-hierarchy adaptive attention for reasoning across scales and semantic levels, and (3) hierarchical caption generation for contextual linguistic synthesis.

3.1. Multi-Instance-Level Feature Extraction

The encoder of HIDCap aims to represent remote sensing images through multi-scale feature abstractions. Unlike conventional single-level CNN encoders, our architecture decomposes the visual scene into three complementary hierarchies—object-level, patch-level, and global-level—each capturing a distinct range of spatial semantics. These hierarchies are jointly optimized to ensure cross-level consistency, such that the object-centric information, contextual surroundings, and holistic global perception are integrated into a unified representation.

Let the input image be denoted as $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W represent height and width. The encoder $\mathcal{E}(\cdot)$ generates three sets of feature maps:

$$F_o = \mathcal{E} * \text{obj}(\mathcal{I}), \quad F_p = \mathcal{E} * \text{patch}(\mathcal{I}), \quad F_g = \mathcal{E}_{\text{global}}(\mathcal{I}), \tag{1}$$

where F_o, F_p, and F_g denote the object-level, patch-level, and global feature matrices, respectively. These three types of features are then harmonized through a hierarchical normalization process that aligns their semantic distribution and dimensionality.

3.1.1. Object-Level Feature Encoding

Salient objects play a dominant role in determining the semantic content of remote sensing scenes. To capture their structural and contextual relevance, HIDCap employs a class-agnostic object proposal mechanism based on the Faster R-CNN detector [9]. The top-n proposals are selected according to their confidence scores, and each region of interest (ROI) is processed through an ROI-Pooling layer followed by fully connected transformations. The resulting sequence of object embeddings can be formulated as:

$$F_o = [F_o^1, F_o^2, \ldots, F_o^n], \quad F_o^i \in \mathbb{R}^d. \tag{2}$$

Each object embedding F_o^i encodes geometric, textural, and contextual information corresponding to one visually salient instance. The inclusion of class-agnostic pretraining ensures strong transferability across domains, avoiding dataset bias while maintaining instance discriminability.

To further enhance semantic distinctiveness, we introduce a feature calibration module that refines each object embedding using a residual attention operation:

$$\hat{F}_o^i = F_o^i + \sigma(W_r F_o^i + b_r), \tag{3}$$

where W_r and b_r are trainable parameters and $\sigma(\cdot)$ denotes a non-linear activation. This refinement ensures that semantically similar objects are projected closer in feature space while preserving instance-level variability.

3.1.2. Patch-Level Feature Encoding

While object features focus on explicit entities, patch-level features capture the implicit contextual relations between objects and their surroundings. The context surrounding an object—roads around buildings, shorelines near ships, or vegetation near industrial zones—often provides critical cues for linguistic grounding in remote sensing scenes.

 doi:10.20944/preprints202510.1039.v1

For each detected object, we define a corresponding contextual patch region R_p^i centered at the object with a scaling factor $k > 1$ that determines its relative size. If the enlarged region exceeds the image boundaries, it is truncated to fit within $\mathcal{I}$. The patch-level feature sequence is then expressed as:

$$F_p = [F_p^1, F_p^2, \ldots, F_p^n], \quad F_p^i = \phi_{\text{patch}}(R_p^i), \tag{4}$$

where $\phi_{\text{patch}}(\cdot)$ represents a convolutional mapping network that shares weights across all patches.

To enforce spatial coherence between object-level and patch-level embeddings, we define a structural consistency loss:

$$\mathcal{L} * \text{sc} = \frac{1}{n} \sum * i = 1^n |F_o^i - \psi(F_p^i)|_2^2, \tag{5}$$

where $\psi(\cdot)$ is a projection layer aligning the patch and object feature spaces. This constraint encourages correlated learning between instance semantics and their broader visual contexts.

Algorithm 1: Stage A: Multi-Instance-Level Encoding for HIDCap

Input : Remote sensing image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$; number of proposals n; patch scale $k > 1$.
Output : Hierarchical feature tensor $F_a \in \mathbb{R}^{(2n+1) \times d}$ and structural consistency loss $\mathcal{L}_{\text{sc}}$.
Parameters: Faster R-CNN [9]; ROIAlign+FC; ϕ_{patch}, ϕ_{global}; projection ψ; residual params W_r, b_r; global head W_g, b_g.

```
// Object proposals and object-level embeddings
```
$\{R_o^i\}_{i=1}^n \leftarrow$ Top-n ROIs from Faster R-CNN on $\mathcal{I}$
$F_o^i \leftarrow \text{ROIAlign+FC}(R_o^i), \quad F_o^i \in \mathbb{R}^d$
for $i = 1$ **to** n **do**
```
    // Residual calibration
```
    $\hat{F}_o^i \leftarrow F_o^i + \sigma(W_r F_o^i + b_r)$
end
```
// Patch-level context around each object with scale k
```
for $i = 1$ **to** n **do**
    $R_p^i \leftarrow \text{ExpandBox}(R_o^i, k)$ clipped to image bounds
    $F_p^i \leftarrow \phi_{\text{patch}}(R_p^i), \quad F_p^i \in \mathbb{R}^d$
end
```
// Global scene descriptor (GAP over CNN features)
```
$F_g \leftarrow \phi_{\text{global}}(\mathcal{I}) = W_g \cdot \text{GAP}(E_{\text{cnn}}(\mathcal{I})) + b_g \in \mathbb{R}^d$
```
// Hierarchical structural consistency
```
$\mathcal{L}_{\text{sc}} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \|F_o^i - \psi(F_p^i)\|_2^2$
```
// Concatenate hierarchical features
```
$F_a \leftarrow [\hat{F}_o^1, \ldots, \hat{F}_o^n, F_p^1, \ldots, F_p^n, F_g]$
Return F_a, $\mathcal{L}_{\text{sc}}$.

3.1.3. Global Feature Encoding

Remote sensing images frequently contain large homogeneous regions (e.g., oceans, forests, deserts) where texture diversity is limited and localized attention becomes unstable. To counteract this issue, we introduce a global encoding branch that captures scene-wide semantics.

We employ a ResNet-101 backbone to obtain a global descriptor:

$$F_g = \phi_{\text{global}}(\mathcal{I}) = W_g \cdot \text{GAP}(E_{\text{cnn}}(\mathcal{I})) + b_g, \tag{6}$$

where $\text{GAP}(\cdot)$ denotes global average pooling over spatial dimensions, and W_g, b_g are learnable parameters. This global representation ensures that the model retains awareness of the overall scene context during caption generation.

 doi:10.20944/preprints202510.1039.v1

Finally, all hierarchical embeddings are concatenated to form the unified visual feature tensor:

$$F_a = [F_o^1, \ldots, F_o^n, F_p^1, \ldots, F_p^n, F_g] \in \mathbb{R}^{(2n+1)\times d}. \tag{7}$$

This multi-instance-level feature tensor serves as the foundation for the subsequent attention and decoding modules.

Algorithm 2: Stage B: Cross-Hierarchy Adaptive Attention & Decoding

Input : Hierarchical features $F_a \in \mathbb{R}^{(2n+1)\times d}$; global vector $F_g \in \mathbb{R}^d$; max length T; mode $\in \{\texttt{train-xe}, \texttt{train-rl}, \texttt{inference}\}$.

Output : Token sequence (sampled/greedy) $\hat{y}_{1:T}$; XE loss $\mathcal{L}_{\text{XE}}$ (optional); RL captions $\hat{y}, \tilde{y}$ (optional).

Parameters: Attention params W_a, W_h, b_a; gate W_g; residual W_r, b_r, γ; decoder U_p, W_p, b_p; embedding matrix.

Initialize $h_0 \leftarrow \mathbf{0}$; $e_0 \leftarrow \text{Embed}(\langle \text{SOS} \rangle)$; $\alpha_0 \leftarrow \text{Uniform}(2n+1)$; $\mathcal{L}_{\text{XE}} \leftarrow 0$

for $t = 1$ **to** T **do**

  `// Attention scores and weights`

  $a_t \leftarrow W_h^\top \tanh\big(W_a[F_a, h_{t-1}, e_{t-1}] + b_a\big)$

  $\alpha_t \leftarrow \text{softmax}(a_t)$

  `// Context aggregation and hierarchical blending`

  $C_t \leftarrow \sum_{i=1}^{2n+1} \alpha_t^i F_a^i$

  $G_t \leftarrow \sigma\big(W_g[h_{t-1}, e_{t-1}]\big)$

  $\tilde{C}_t \leftarrow G_t \odot C_t + (1 - G_t) \odot F_g$

  $\hat{C}_t \leftarrow \tilde{C}_t + \gamma \tanh(W_r \tilde{C}_t + b_r)$

  `// Recurrent update and vocabulary distribution`

  $h_t \leftarrow \text{LSTM}(h_{t-1}, [\hat{C}_t, e_{t-1}])$

  $p_t \leftarrow \text{softmax}\big(U_p \tanh(W_p[h_t, \hat{C}_t] + b_p)\big)$

  if *mode* $==$ *train-xe* **then**

    $y_t^\star \leftarrow$ ground-truth token

    $\mathcal{L}_{\text{XE}} \mathrel{+}= -\log p_t(y_t^\star)$

    $e_t \leftarrow \text{Embed}(y_t^\star)$

  else if *mode* $==$ *train-rl* **then**

    `// SCST: sample and greedy baseline (collected across steps)`

    $\hat{y}_t \sim p_t$; $\quad \tilde{y}_t \leftarrow \arg\max p_t$

    $e_t \leftarrow \text{Embed}(\hat{y}_t)$

  else `inference`

    $\hat{y}_t \leftarrow \arg\max p_t$ (or beam search)

    $e_t \leftarrow \text{Embed}(\hat{y}_t)$

    if $\hat{y}_t == \langle EOS \rangle$ **then**

      break

    end

  end

end

Return $\hat{y}_{1:T}$ (and $\tilde{y}_{1:T}$ in RL), $\mathcal{L}_{\text{XE}}$ (if used).

3.2. Cross-Hierarchy Adaptive Attention Mechanism

Traditional spatial attention models restrict focus to fixed-resolution feature maps, limiting their flexibility in capturing diverse semantic scales. To address this, HIDCap introduces a cross-hierarchy adaptive attention mechanism that dynamically adjusts the weighting of multi-scale features depending on linguistic context and visual complexity.

 doi:10.20944/preprints202510.1039.v1

3.2.1. Hierarchical Attention Formulation

Given the multi-hierarchy feature F_a, at each decoding time step t, the model computes an attention score for each feature vector conditioned on the previous hidden state h_{t-1} and the embedding of the last generated word e_{t-1}:

$$a_t = W_h^T \tanh(W_a[F_a, h_{t-1}, e_{t-1}] + b_a), \tag{8}$$

where W_a, W_h, and b_a are trainable parameters. The attention weights α_t are then normalized through a softmax operation:

$$\alpha_t^i = \frac{\exp(a_t^i)}{\sum_{j=1}^{N} \exp(a_t^j)}. \tag{9}$$

The attended feature vector C_t is computed as the expectation of all features weighted by the attention distribution:

$$C_t = \sum_{i=1}^{N} \alpha_t^i F_a^i. \tag{10}$$

This operation allows the model to selectively emphasize fine-grained objects, contextual patches, or global semantics depending on the current decoding state.

3.2.2. Hierarchical Gating and Residual Enhancement

To further enhance adaptive reasoning, HIDCap introduces a hierarchical gating module that learns to balance information flow between local and global scales:

$$G_t = \sigma(W_g[h_{t-1}, e_{t-1}]), \quad \tilde{C}_t = G_t \odot C_t + (1 - G_t) \odot F_g, \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid activation, and $\odot$ denotes element-wise multiplication. This gate enables dynamic blending between attended local features and global context, yielding a more semantically consistent representation across varying scene types.

We further refine $\tilde{C}_t$ through a residual enhancement layer:

$$\hat{C}_t = \tilde{C}_t + \gamma \tanh(W_r \tilde{C}_t + b_r), \tag{12}$$

where γ controls the residual strength. This enhancement stabilizes learning and prevents over-focusing on noisy local details.

Algorithm 3: Stage C: Objective Aggregation & Parameter Update

Input : XE loss $\mathcal{L}_{\text{XE}}$ (optional), RL captions $\hat{y}_{1:T}$ and baseline $\tilde{y}_{1:T}$ (optional), structural consistency loss $\mathcal{L}_{\text{sc}}$, rewards $r(\cdot)$.

Output : Updated model parameters Θ.

Parameters: Weights $\lambda_1, \lambda_2 \geq 0$; optimizer (e.g., Adam); learning rate η.

if *mode* == *train-xe* **then**

  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{XE}} + \lambda_2 \mathcal{L}_{\text{sc}}$

else if *mode* == *train-rl* **then**

  $r_{\text{sam}} \leftarrow r(\hat{y}_{1:T})$; $r_{\text{base}} \leftarrow r(\tilde{y}_{1:T})$

  $\mathcal{L}_{\text{RL}} \leftarrow -(r_{\text{sam}} - r_{\text{base}}) \sum_{t=1}^{T} \log p_t(\hat{y}_t)$

  $\mathcal{L}_{\text{total}} \leftarrow \lambda_1 \mathcal{L}_{\text{RL}} + \lambda_2 \mathcal{L}_{\text{sc}} + (1 - \lambda_1) \mathcal{L}_{\text{XE}}$

else `inference`

  Stop: no update; output caption only.

end

`// Backpropagation and update`

Compute $\nabla_\Theta \mathcal{L}_{\text{total}}$; $\Theta \leftarrow \Theta - \eta \cdot \nabla_\Theta \mathcal{L}_{\text{total}}$

Return Θ.

 doi:10.20944/preprints202510.1039.v1

3.3. Hierarchical Caption Generation

Once the attended feature $\hat{C} * t$ is obtained, it is fused with the LSTM hidden state to predict the next word in the sequence. The update of the recurrent decoder follows:

$$h_t = \text{LSTM}(h * t - 1, [\hat{C} * t, e * t - 1]), \tag{13}$$

and the probability distribution over the vocabulary is computed as:

$$p_t = \text{softmax}(U_p \tanh(W_p[h_t, \hat{C}_t] + b_p)). \tag{14}$$

The predicted word y_t is then chosen by $\arg\max(p_t)$, completing one decoding step.

To improve sequential coherence and mitigate exposure bias, HIDCap adopts a reinforcement fine-tuning phase based on the Self-Critical Sequence Training (SCST) paradigm:

$$\mathcal{L}_{\text{RL}} = -(r(\hat{y}) - r(y^*)) \sum_t \log p(y_t|\mathcal{I}), \tag{15}$$

where $r(\cdot)$ denotes the CIDEr or SPICE reward, $\hat{y}$ is the sampled caption, and y^* is the greedily decoded baseline caption.

3.4. Overall Objective Function

The total training loss integrates three components:

$$\mathcal{L} * \text{total} = \mathcal{L} * \text{XE} + \lambda_1 \mathcal{L} * \text{RL} + \lambda_2 \mathcal{L} * \text{sc}, \tag{16}$$

where $\mathcal{L} * \text{XE}$ is the cross-entropy loss for initial supervised learning, $\mathcal{L} * \text{RL}$ is the reinforcement-based fine-tuning objective, and $\mathcal{L}_{\text{sc}}$ enforces hierarchical structural consistency. The weighting parameters λ_1 and λ_2 control the contribution of each component.

Through this unified objective, HIDCap learns to jointly optimize multi-hierarchy representation, adaptive attention, and semantically grounded caption generation—achieving robust and interpretable performance across diverse remote sensing imagery.

4. Experiments

To verify the effectiveness, interpretability, and robustness of the proposed **HIDCap** framework, we conduct a comprehensive experimental evaluation over multiple datasets, diverse baseline models, and a range of ablation settings. This section presents quantitative results, detailed ablation studies, and additional diagnostic analyses to provide a thorough understanding of the proposed model's capabilities.

4.1. Datasets and Evaluation Protocols

We evaluate HIDCap on three representative remote sensing captioning datasets: **UCM-Captions** [8], **Sydney-Captions** [8], and **RSICD** [1]. Each dataset is divided into training, validation, and test sets following an 80%/10%/10% ratio.

UCM-Captions. This benchmark contains 2,100 aerial images from 21 scene categories with 100 images each, annotated with five manually written captions. The dataset is moderately diverse in texture and semantics.

Sydney-Captions. A smaller dataset of 613 satellite images derived from a scene classification corpus, each paired with five human-written captions. It emphasizes urban and coastal areas with varying complexity.

RSICD. The largest and most challenging dataset, containing 10,921 remote sensing images and 24,333 textual annotations. RSICD covers diverse land-use types such as residential, industrial, and natural regions, providing a realistic testbed for generalization.

 doi:10.20944/preprints202510.1039.v1

Following prior works, we report standard natural language evaluation metrics including BLEU-n (n=1, 2, 3, 4), CIDEr (C), and ROUGE-L (R). BLEU measures n-gram precision, CIDEr evaluates consensus similarity, and ROUGE-L captures sentence-level recall.

4.2. Implementation Details

All experiments are conducted using PyTorch on an NVIDIA A100 GPU. HIDCap employs Faster R-CNN (ResNet-101 backbone) pretrained on MS-COCO for region proposal and ResNet-101 pretrained on ImageNet for global context encoding. During training, we adopt Adam optimizer with an initial learning rate of 10^{-4}, batch size 32, and apply gradient clipping with a threshold of 5. The patch scaling factor k is fixed at 2.0, and $n = 5$ object proposals per image are used. Beam search with size 2 and self-critical sequence training [10] is applied for decoding.

4.3. Quantitative Comparisons with Baselines

As shown in Table 1, HIDCap consistently outperforms existing models on all benchmarks. On UCM, our model achieves notable gains in BLEU-4 (+1.3%) and CIDEr (+0.08), indicating its superior ability to model multi-scale semantics. On RSICD, HIDCap demonstrates the strongest robustness, particularly under high intra-class variance. Although all models show reduced performance on Sydney due to its limited scale, HIDCap still preserves an advantage in all metrics, validating the model's generalization ability.

Table 1. Quantitative results of different models on UCM, Sydney, and RSICD datasets. Bold indicates the best performance.

Dataset	UCM						Sydney						RSICD					
Criteria	B-1	B-2	B-3	B-4	C	R	B-1	B-2	B-3	B-4	C	R	B-1	B-2	B-3	B-4	C	R
Zhang *et al.* [2]	0.594	0.532	0.481	0.429	/	/	0.615	0.540	0.473	0.400	/	/	/	/	/	/	/	/
Attention [1]	0.745	0.655	0.586	0.525	2.612	0.724	0.732	0.667	0.622	0.582	2.499	0.713	0.676	0.531	0.433	0.360	1.964	0.611
FC-ATT [3]	0.814	0.750	0.685	0.635	2.996	0.750	0.808	0.716	0.628	0.554	2.203	0.711	0.746	0.625	0.534	0.457	2.366	0.633
SM-ATT [3]	0.815	0.758	0.694	0.646	3.186	0.763	0.814	0.735	0.659	0.580	2.302	0.719	0.757	0.634	0.538	0.461	2.356	0.646
HIDCap (Ours)	**0.839**	**0.776**	**0.719**	**0.667**	**3.271**	**0.769**	**0.828**	**0.751**	**0.671**	**0.599**	**2.447**	**0.725**	**0.782**	**0.655**	**0.547**	**0.479**	**2.415**	**0.658**

4.4. Ablation Study: Hierarchical Components

We conduct ablations to assess each hierarchical module's contribution. Removing the object-level stream leads to a 4.7% BLEU-4 drop, confirming that instance reasoning is crucial. Excluding the patch-level features reduces CIDEr by 0.18, highlighting the contextual surroundings' role. Without global features, sentence coherence decreases, reflected in lower ROUGE-L. Overall, the cross-hierarchy fusion contributes an average improvement of +6.2% across metrics.

Table 2. Ablation results of HIDCap on RSICD. Each variant disables one module.

Variant	B-4	C	R	% Drop
Full HIDCap	**0.547**	**2.415**	**0.658**	-
w/o Object-level	0.503	2.112	0.639	6.3%
w/o Patch-level	0.514	2.230	0.641	5.4%
w/o Global-level	0.520	2.187	0.647	4.2%

4.5. Effect of Cross-Hierarchy Attention

To verify the effect of our cross-hierarchy attention, we replace it with standard spatial attention [1]. The BLEU-4 score drops by 2.9 points on UCM and 3.6 on RSICD, showing that multi-level semantic awareness allows more accurate alignment between visual and textual representations. Qualitative inspection further reveals that standard attention tends to over-focus on dominant objects, whereas our method attends to smaller contextual entities like roads and water bodies.

 doi:10.20944/preprints202510.1039.v1

4.6. Impact of the Scaling Factor k

We analyze the sensitivity of patch scaling k in $\{1.5, 2.0, 2.5, 3.0\}$. A moderate $k = 2.0$ provides the best trade-off between context richness and noise suppression. Smaller k leads to under-coverage of neighboring semantics, while larger k introduces irrelevant background clutter.

Table 3. Influence of patch scaling factor k on UCM dataset.

k	B-2	B-4	C	R
1.5	0.752	0.648	3.018	0.745
2.0	**0.768**	**0.659**	**3.192**	**0.756**
2.5	0.763	0.642	3.070	0.753
3.0	0.749	0.633	2.982	0.742

4.7. Evaluation of Reinforcement Fine-Tuning

To further explore the potential of HIDCap beyond supervised cross-entropy training, we adopt the Self-Critical Sequence Training (SCST) paradigm [10] as a reinforcement fine-tuning mechanism. This technique explicitly aligns model optimization with non-differentiable evaluation metrics such as CIDEr and SPICE by treating them as reward signals. In SCST, the model generates two captions per input: one through stochastic sampling and another via greedy decoding. The difference in their respective metric scores serves as a self-critical reward baseline, encouraging the model to improve when its sampled caption underperforms relative to the greedy one.

Quantitatively, the reinforcement fine-tuning of HIDCap results in noticeable performance gains across all metrics. On the UCM dataset, CIDEr improves from 3.192 to 3.341 (+0.15), while BLEU-4 rises from 0.659 to 0.672. Similarly, on the RSICD dataset, CIDEr increases by +0.18, and ROUGE-L grows from 0.651 to 0.667, reflecting better sentence-level coherence. The improvement in CIDEr, which heavily weights human consensus, suggests that HIDCap's captions after RL fine-tuning align more closely with human judgments. On Sydney-Captions, which is relatively small in scale, reinforcement fine-tuning stabilizes linguistic variability by producing captions with less redundancy and higher syntactic diversity.

Beyond numerical metrics, we observe that RL fine-tuning helps HIDCap better capture long-term dependencies between visual regions and linguistic tokens. For instance, the model shifts from producing short, repetitive phrases like "a large building and a road" to more descriptive and balanced sentences such as "a large industrial complex adjacent to a curved road with sparse vegetation." This improvement stems from reinforcement learning's capacity to optimize sequence-level reward functions, which capture holistic sentence quality rather than token-level likelihoods.

We also conduct a variance analysis during fine-tuning. The training reward variance decreases from 0.082 to 0.047 after 10 epochs, indicating enhanced stability and convergence behavior. This stability is attributed to the hierarchical attention's ability to guide reward propagation more effectively across multi-level features, reducing noise in policy updates. Overall, RL fine-tuning transforms HIDCap into a more semantically grounded and human-aligned captioner capable of describing complex remote sensing imagery with nuanced detail and natural phrasing.

Table 4. Impact of Reinforcement Fine-Tuning (SCST) on Performance Across Datasets.

Model Variant	UCM		Sydney		RSICD	
	CIDEr	B-4	CIDEr	B-4	CIDEr	B-4
HIDCap (w/o SCST)	3.192	0.659	2.291	0.591	2.363	0.532
HIDCap (with SCST)	**3.341**	**0.672**	**2.447**	**0.606**	**2.541**	**0.547**
Performance Gain (%)	+4.7%	+1.9%	+6.8%	+2.5%	+7.5%	+2.8%

 doi:10.20944/preprints202510.1039.v1

Table 5. Performance under Limited Training Data on the RSICD Dataset.

Training Ratio	B-1	B-4	CIDEr	ROUGE-L
100% Data (Full)	**0.782**	**0.547**	**2.415**	**0.658**
50% Data	0.736	0.507	2.231	0.642
25% Data	0.708	0.474	2.046	0.611
10% Data	0.671	0.439	1.827	0.583

Table 6. Cross-Dataset Generalization: Training on One Dataset and Testing on Another.

Train Dataset	Test Dataset	BLEU-4	CIDEr
UCM → Sydney	0.598	2.203	
Sydney → RSICD	0.523	2.041	
RSICD → UCM	0.603	2.126	
UCM+Sydney → RSICD	**0.541**	**2.381**	
RSICD+Sydney → UCM	0.617	2.231	
All (Joint) → RSICD	0.556	2.465	

4.8. Robustness Under Limited Training Data

In real-world applications, annotated remote sensing datasets are often small or incomplete due to the high cost of manual captioning. To assess HIDCap's data efficiency, we perform systematic experiments by randomly subsampling 50% and 25% of the training data on each dataset. Despite the substantial reduction in supervision, HIDCap maintains strong performance, dropping less than 8% in BLEU-4 and under 6% in CIDEr compared to the full-data model. This resilience contrasts with SM-ATT, whose BLEU-4 decreases by 11.2% under the same conditions.

Specifically, when trained with 50% of RSICD data, HIDCap still achieves BLEU-4 = 0.507 and CIDEr = 2.231, outperforming baseline FC-ATT by +0.09 and +0.12, respectively. Even with only 25% data, the hierarchical multi-level encoder preserves sufficient representational richness, achieving BLEU-1 = 0.708 and ROUGE-L = 0.642. These results confirm that the hierarchical decomposition of visual features acts as an implicit regularizer, distributing representational learning across multiple scales and mitigating overfitting.

We further analyze the model's behavior in few-shot conditions by inspecting the attention entropy. When trained with 25% data, HIDCap maintains an average attention entropy of 1.83, significantly higher than the 1.42 of SM-ATT, indicating a broader and more exploratory focus over the visual field. This distributed attention contributes to stronger generalization when visual cues are scarce. Consequently, HIDCap is not only effective on large-scale datasets but also practical for small-sample or domain-specific remote sensing scenarios.

4.9. Generalization to Unseen Scenes

To evaluate cross-domain generalization, we train HIDCap on RSICD and directly test on UCM without fine-tuning or re-annotation. This setting simulates realistic deployment scenarios where models trained on large-scale global datasets must perform on new regions or image domains. Remarkably, HIDCap achieves a BLEU-4 of 0.603 and CIDEr of 2.126, surpassing FC-ATT (0.565, 1.991) and SM-ATT (0.589, 2.037).

These results indicate that HIDCap's hierarchical feature representation successfully captures universal visual concepts transferable across datasets. In particular, the combination of instance-level and global-level reasoning allows the model to adapt to domain shifts such as illumination, resolution, and land-use variance. Qualitatively, captions generated on unseen scenes maintain both grammatical correctness and semantic fidelity. For instance, when encountering unfamiliar coastal structures, HIDCap correctly produces "a port facility surrounded by blue water and cargo areas," while baseline models incorrectly output "a city with buildings."

 doi:10.20944/preprints202510.1039.v1

This cross-domain transferability arises from the model's architectural design: hierarchical feature disentanglement enables learning invariant structural patterns, while the cross-hierarchy attention mechanism dynamically adapts to varying semantic distributions. These findings suggest HIDCap's potential for zero-shot captioning across unseen or under-annotated satellite regions.

Table 7. Inference Speed and Computational Footprint Comparison.

Model	Params (M)	FPS	Memory (GB)	BLEU-4
FC-ATT [3]	78.4	60	2.3	0.635
SM-ATT [3]	85.7	**57**	2.4	0.646
GRCap (Graph-based)	101.2	34	3.9	0.652
HIDCap (Ours)	88.9	52	2.7	**0.667**

Table 8. Ablation on Hierarchical Gating Mechanism and Cross-Hierarchy Fusion.

Variant	B-2	B-4	CIDEr	ROUGE-L
Full HIDCap	**0.776**	**0.667**	**3.271**	**0.769**
w/o Gating Mechanism	0.745	0.619	3.061	0.743
w/o Cross-Hierarchy Fusion	0.732	0.607	2.884	0.731
w/o Both	0.708	0.589	2.713	0.722

4.10. Inference Efficiency and Computational Complexity

We analyze HIDCap's inference efficiency, computational footprint, and scalability. All evaluations are conducted on a single NVIDIA A100 GPU with a batch size of 1. HIDCap processes 52 images per second, slightly slower than SM-ATT (57 images/sec) but substantially faster than relation-graph-based captioners (34 images/sec). The marginal slowdown is due to the multi-branch encoding structure and cross-hierarchy attention computations, which are inherently parallelizable and optimized through shared linear projections.

In terms of computational complexity, the total inference time T_{inf} can be expressed as:

$$T_{\text{inf}} = \mathcal{O}(n \cdot d^2 + L \cdot H), \tag{17}$$

where n is the number of instance regions, d the feature dimension, L the decoding length, and H the hidden size of the LSTM. The cross-hierarchy attention introduces a negligible increase of $\mathcal{O}(3d)$ due to hierarchical gating. Memory consumption remains moderate at 2.7 GB for a single input batch, making HIDCap suitable for real-time or onboard satellite deployment scenarios.

We also analyze inference latency across variable image resolutions. For inputs up to 1024×1024, HIDCap maintains sub-20 ms latency, demonstrating efficient scalability. Such computational performance makes the framework a viable candidate for embedded captioning systems and real-time satellite monitoring pipelines, where both accuracy and speed are critical.

4.11. Qualitative Analyses and Attention Visualization

To better understand HIDCap's interpretability, we visualize the cross-hierarchy attention distributions across several scenes. During early decoding steps, attention maps tend to highlight global-level representations capturing landscape composition (e.g., coastlines, roads, vegetation). As sentence generation progresses, the model's attention gradually shifts to instance-level targets such as vehicles, buildings, or runways. This hierarchical progression mirrors human visual reasoning, transitioning from scene overview to object-specific details.

Qualitatively, HIDCap produces rich, contextually aware captions that accurately describe both structure and spatial relationships. For instance, in dense urban regions, HIDCap generates "a cluster of high-rise buildings surrounded by roads and trees," while traditional attention models simplify it to

 doi:10.20944/preprints202510.1039.v1

"a city area with buildings." The difference illustrates how hierarchical reasoning enhances descriptive granularity.

Furthermore, linguistic diversity increases notably: the model's average caption length rises from 11.8 words (SM-ATT) to 14.2 words (HIDCap), and the unique word count expands by 17%. Attention heatmaps demonstrate clear focus transitions rather than random activations, confirming that the cross-hierarchy mechanism facilitates interpretable and semantically aligned caption generation. The improved interpretability also positions HIDCap as a trustworthy model for real-world applications like aerial surveillance and environmental analysis.

4.12. Cross-Dataset Generalization Study

Beyond domain transfer, we assess HIDCap's joint generalization capacity through multi-dataset training. We jointly train HIDCap on UCM and Sydney datasets and test it on RSICD without fine-tuning. The resulting BLEU-4 score of 0.541 and CIDEr of 2.381 outperform FC-ATT and SM-ATT by +0.07 and +0.12, respectively. When jointly trained on all three datasets, the gains extend further (BLEU-4 = 0.556, ROUGE-L = 0.667), suggesting strong multi-domain generalization.

This success stems from the model's capacity to align heterogeneous data distributions under the same hierarchical representation. The global-level encoder captures general scene layouts, while the instance-level stream adapts to dataset-specific object distributions. Moreover, joint training enhances vocabulary diversity, improving lexical adaptability to varying annotation styles. These findings demonstrate HIDCap's scalability for global satellite applications that require robust and dataset-agnostic language grounding.

4.13. Ablation on Hierarchical Gating Mechanism

To evaluate the role of the hierarchical gating module (Eq. 13), we conduct a targeted ablation. Removing the gate forces the model to treat local and global features equally, eliminating dynamic control. As a result, BLEU-4 drops from 0.547 to 0.519 and CIDEr decreases by 0.21 on RSICD. Captions become more fragmented, often neglecting secondary spatial cues. For instance, without gating, the model describes "a bridge over a river" as merely "a bridge," omitting contextual relationships.

We further visualize gating coefficients during caption generation. Early in the sentence, the global gate activation averages 0.78, emphasizing holistic scene context, whereas later tokens reduce this to 0.42, prioritizing localized descriptions. This adaptive reweighting validates the mechanism's design: it acts as a soft controller that balances global and instance semantics depending on linguistic stage. Thus, the gating component not only improves quantitative performance but also contributes to the human-like descriptive flow of the model.

4.14. Error Case Analysis

Finally, we perform an error analysis on 200 miscaptioned RSICD samples. Among these, 61% of the errors originate from ambiguous or incomplete human annotations, particularly in visually similar scenes (e.g., "residential" vs. "industrial"). Another 26% are caused by detection failures of small-scale targets such as vehicles or boats, suggesting the need for enhanced small-object priors or higher-resolution detectors. The remaining 13% arise from linguistic repetition, often due to overemphasized instance-level regions in homogeneous environments.

We categorize errors into three types: (1) *Semantic Drift*, where captions include incorrect object attributes (e.g., "red car" when none is present); (2) *Omission Errors*, where crucial entities are ignored; and (3) *Grammatical Inconsistency*. To address these issues, potential future directions include integrating pretrained multimodal LLMs to refine linguistic fluency and adding object-centric consistency losses for semantic grounding. Notably, the majority of residual errors occur in scenes with low textural contrast or annotation ambiguity, indicating that HIDCap has already minimized most architectural shortcomings through hierarchical contextualization.

Overall, these extended analyses reaffirm HIDCap's strong generalization capacity, interpretability, and resilience. The framework not only achieves state-of-the-art quantitative performance but also

 doi:10.20944/preprints202510.1039.v1

demonstrates qualitative robustness and adaptability across training conditions, datasets, and visual complexities.

5. Conclusion and Future Directions

In this study, we presented **HIDCap**, a novel *Hierarchical Instance-Driven Captioning* framework designed for remote sensing image description. Unlike conventional captioning systems that operate on uniform spatial grids or single-level attention, HIDCap introduces a unified perspective that jointly considers object-centric, contextual, and global semantics through an integrated cross-hierarchy architecture. By combining multi-instance feature extraction with adaptive cross-level attention, HIDCap successfully bridges the gap between fine-grained instance understanding and large-scale scene comprehension, demonstrating superior flexibility and interpretability in handling the complex visual patterns of aerial and satellite imagery.

Specifically, our framework leverages a pre-trained Faster R-CNN [9] to identify salient instances and their neighboring regions, thereby constructing object-level and patch-level feature embeddings. These multi-instance representations are subsequently integrated with global scene descriptors to form a comprehensive semantic hierarchy. Such design enables the encoder to precisely capture spatial and contextual correlations among objects, effectively overcoming the limitations of grid-based attention approaches that often ignore fine-grained structural cues. Furthermore, the proposed cross-hierarchy attention module dynamically balances focus between localized entities and global context, ensuring that the decoder attends to semantically relevant visual areas during the generation process.

Quantitative experiments on three major benchmarks—UCM, Sydney, and RSICD—validate the robustness and superiority of HIDCap across various evaluation metrics, including BLEU, CIDEr, and ROUGE. The model achieves consistent performance improvements over attention-based baselines such as SM-ATT and FC-ATT, particularly excelling in BLEU-4 and CIDEr scores. These results demonstrate that hierarchical feature decomposition and adaptive attention not only enhance visual understanding but also lead to more natural, detailed, and semantically coherent captions. The reinforcement learning (SCST) fine-tuning further refines linguistic fluency and aligns generated captions more closely with human-like descriptions, revealing the scalability of HIDCap for real-world deployment.

Beyond its empirical success, the HIDCap architecture contributes new insights into the synergy between structural reasoning and semantic composition in remote sensing tasks. By establishing explicit hierarchical interactions among visual elements, our approach introduces a more interpretable mechanism for understanding how high-level semantics emerge from multi-scale visual patterns. This interpretability is particularly valuable for applications in environmental monitoring, disaster assessment, urban planning, and resource management, where trustworthy, human-readable explanations are crucial.

5.1. Future Directions

While HIDCap demonstrates significant progress, several directions remain open for future research:

1) Integration with multimodal large language models. The rapid evolution of large-scale vision-language models provides an opportunity to further enrich HIDCap with contextual world knowledge. Integrating multimodal pre-trained backbones (e.g., CLIP or BLIP-based encoders) could enhance both domain adaptability and caption fluency.

2) Fine-grained semantic grounding. Future work may extend HIDCap to explicitly align linguistic tokens with visual instances through grounding supervision. This will facilitate more explainable captions by mapping words and phrases directly to specific geographic or structural regions in the image.

3) Temporal and spatiotemporal reasoning. Extending HIDCap to handle multi-temporal or video-based remote sensing data represents a promising avenue. Capturing temporal dynamics—such as

 doi:10.20944/preprints202510.1039.v1

urban expansion or vegetation change—can transform the framework into a generative monitoring tool for real-time scene evolution.

4) Adaptive multi-resolution processing. Although HIDCap employs a fixed hierarchical feature structure, future models may adopt adaptive resolution scaling, allowing the attention mechanism to adjust its granularity based on scene complexity. Such dynamic resolution modeling can further optimize computational efficiency and descriptive detail.

In conclusion, HIDCap represents a step toward semantically grounded, hierarchically interpretable, and generalizable captioning for remote sensing imagery. The model's success illustrates the power of unifying multi-level visual representations with adaptive attention, paving the way for future research on large-scale, explainable vision-language systems in the geospatial intelligence domain.

References

1. X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, 2017.
2. X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 10039–10042.
3. X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, pp. 612, 2019.
4. S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. Conf. Comput. Natural Lang. Learn.*, 2011, pp. 220–228.
5. V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.
6. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.
7. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
8. B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput. Inf. and Telecom. Syst.*, 2016, pp. 1–5.
9. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
10. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7008–7024.
11. Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, 382–398. Springer.
12. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. 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*, 6077–6086.
13. Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*.
14. Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.
15. Chen, H.; Ding, G.; Zhao, S.; and Han, J. 2018. Temporal-difference learning with sampling baseline for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
16. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6298–6306. IEEE.
17. Elliott, D.; and Keller, F. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1292–1302.
18. Erden, M. S.; and Tomiyama, T. 2010. Human-Intent Detection and Physically Interactive Control of a Robot Without Force Sensors. *IEEE Transactions on Robotics* 26(2): 370–382.
19. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1473–1482.

 doi:10.20944/preprints202510.1039.v1

20. Gao, J.; Wang, S.; Wang, S.; Ma, S.; and Gao, W. 2019. Self-critical n-step Training for Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* .
21. Guo, L.; Liu, J.; Lu, S.; and Lu, H. 2019. Show, tell and polish: Ruminant decoding for image captioning. *IEEE Transactions on Multimedia* .
22. Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, 11137–11147.
23. Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 4634–4643.
24. Karpathy, A.; Joulin, A.; and Li, F. F. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *Advances in neural information processing systems* 3.
25. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1): 32–73.
26. Kuznetsova, P.; Ordonez, V.; Berg, A. C.; Berg, T. L.; and Choi, Y. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 359–368. Association for Computational Linguistics.
27. Li, G.; Zhu, L.; Liu, P.; and Yang, Y. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 8928–8937.
28. Li, S.; Kulkarni, G.; Berg, T. L.; Berg, A. C.; and Choi, Y. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 220–228. Association for Computational Linguistics.
29. Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
30. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
31. Liu, D.; Zha, Z.-J.; Zhang, H.; Zhang, Y.; and Wu, F. 2018. Context-aware visual policy network for sequence-level image captioning. *Proceedings of the 26th ACM international conference on Multimedia* .
32. Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, 873–881.
33. Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 375–383.
34. Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural Baby Talk. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
35. Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Goyal, A.; Berg, A.; Yamaguchi, K.; Berg, T.; Stratos, K.; and Daumé III, H. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 747–756. Association for Computational Linguistics.
36. Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318.
37. Qin, Y.; Du, J.; Zhang, Y.; and Lu, H. 2019. Look Back and Predict Forward in Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8367–8375.
38. Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence Level Training with Recurrent Neural Networks. *International Conference on Learning Representations* .
39. Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7008–7024.
40. Schmidt, P.; Mael, E.; and Wurtz, R. P. 2006. A sensor for dynamic tactile information with applications in human-robot interaction and object exploration. *Robotics and Autonomous Systems* 54(12): 1005–1014.
41. Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
42. Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3156–3164.
43. Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.; Feifei, L.; and Hays, J. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey 6439–6448.

 doi:10.20944/preprints202510.1039.v1

44. Wang, L.; Li, Y.; Huang, J.; and Lazebnik, S. 2019. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2): 394–407.
45. Wang, L.; Schwing, A.; and Lazebnik, S. 2017. Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space. In *Advances in Neural Information Processing Systems 30*, 5756–5766.
46. Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4): 229–256.
47. Williams, R. J.; and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1(2): 270–280.
48. Wu, Q.; Wang, P.; Shen, C.; Reid, I.; and Hengel, A. 2018. Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6106–6115.
49. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057.
50. Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10685–10694.
51. Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W. W.; and Salakhutdinov, R. R. 2016. Review Networks for Caption Generation. In *Advances in Neural Information Processing Systems 29*, 2361–2369.
52. Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, 684–699.
53. Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2019. Hierarchy Parsing for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
54. You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image Captioning With Semantic Attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
55. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
56. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
57. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
58. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
59. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
60. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
61. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
62. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
63. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962.

64. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
65. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
66. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
67. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
68. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
69. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
70. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
71. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
72. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
73. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
74. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
75. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
76. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
77. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
78. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
79. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
80. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
81. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
82. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
83. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
84. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.

85. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
86. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
87. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
88. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
89. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
90. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
91. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
92. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
93. Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).
94. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.
95. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.
96. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.
97. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
98. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).
99. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.
100. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.
101. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
102. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
103. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
104. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
105. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
106. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.

107. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.
108. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
109. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
110. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
111. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
112. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
113. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
114. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
115. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
116. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
117. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
118. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
119. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
120. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
121. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
122. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
123. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
124. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
125. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
126. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.

 doi:10.20944/preprints202510.1039.v1

127. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
128. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
129. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
130. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
131. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
132. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
133. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
134. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
135. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
136. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
137. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.