Article

Context-Aware Multi-Anchor Captioning for Text-Rich Image Understanding

Théo Marchand [*] , Lena Roux , Saidi Kareem , Noe Gauthier

Posted Date: 10 October 2025

doi: 10.20944/preprints202510.0736.v1

Keywords: Text-based Image Captioning; Multi-Anchor Reasoning; Caption Diversity; Scene Text Understanding; Graph-based Visual Language Modeling

Article

Context-Aware Multi-Anchor Captioning for Text-Rich Image Understanding

Théo Marchand *, Lena Roux, Saidi Kareem and Noe Gauthier

Aix-Marseille University
* Correspondence: t.marchand@univ-amu.fr

Abstract

Understanding images embedded with textual elements is fundamental for advancing fine-grained visual reasoning. Unlike traditional image captioning, which focuses on object and scene descriptions, text-based image captioning (TextCap) demands the ability to *read*, *comprehend*, and *contextualize* text within complex visual environments. This challenge arises from the intricate relationships between visual semantics and embedded texts such as road signs, brand names, or product labels, which together convey richer scene-level narratives. Existing models typically adapt classical captioning architectures to this task by generating a single global caption, inevitably oversimplifying the nuanced interdependencies between visual regions and textual content. In this work, we introduce a new framework named **Multi-Anchor Captioner (MACap)**, which seeks to produce diverse and fine-grained captions through a structured anchoring mechanism. Instead of treating the image as a whole, MACap decomposes it into multiple *anchor-centered subgraphs*, each focusing on a specific text region and its corresponding contextual neighborhood. The framework involves three sequential stages: (1) an *Anchor Proposal Module (APM)* that identifies informative text tokens and groups them with their relevant visual contexts; (2) an *Anchor Graph Constructor (AGC)* that models semantic dependencies across anchors via graph propagation; and (3) a *Multi-View Caption Generator (MCG)* that synthesizes multiple captions under distinct anchor views, ensuring both accuracy and content diversity. Empirical evaluations on the TextCaps benchmark demonstrate that MACap achieves state-of-the-art performance, surpassing existing baselines in both descriptive fidelity and caption diversity metrics. Beyond quantitative superiority, qualitative results reveal MACap's ability to generate complementary captions covering multifaceted aspects of a single image—ranging from object appearance to textual semantics—highlighting its capacity for comprehensive scene understanding.

Keywords: text-based image captioning, multi-anchor reasoning, caption diversity, scene text understanding, graph-based visual language modeling

1. Introduction

Text is an integral part of our daily visual experience, serving as a bridge between visual perception and semantic reasoning [13]. In natural images, texts often provide contextual clues that enhance the interpretation of scenes—be it understanding a store's signboard, recognizing brand names on packaging, or reading warnings on road indicators [5,19,20,34,41]. Traditional image captioning frameworks, however, have largely overlooked this dimension by focusing primarily on visual object relationships and scene-level summaries. As a result, they fail to capture textual cues that may be crucial for accurate semantic understanding. To overcome this limitation, [40] formalized the *text-based image captioning* (TextCap) task, emphasizing the importance of models that can not only *describe* but also *read* textual information from images.

TextCap serves a wide range of practical applications, from enhancing accessibility for visually impaired users [13] to enabling automated understanding of information-dense visual content such as receipts, advertisements, and street scenes. Unlike ordinary captions that summarize an image's main objects, TextCap aims to describe subtle textual and visual interactions, requiring the model to reason over multimodal cues with high granularity. Despite its practical importance, the task remains challenging because (1) textual regions are often cluttered or partially occluded, (2) text semantics depend heavily on spatial and relational context, and (3) it is difficult to determine which parts of the textual content are most relevant for caption generation.

Early efforts have attempted to adapt traditional image captioning methods [2,19,21] to the TextCap setting. While these models excel at identifying key visual objects, they struggle to capture the semantics of embedded texts. Subsequent works, such as M4C-Captioner [40], attempted to integrate OCR-based text recognition tools [4,6,31] into captioning pipelines. However, these methods remain limited in their ability to fuse textual and visual information, often producing overly generic descriptions that fail to represent the intricate structure of text-visual relations. The underlying issue is that a single caption cannot encapsulate the full spectrum of meaningful elements present in text-rich images.

To address these limitations, we propose **Multi-Anchor Captioner (MACap)**, a new framework that performs captioning through anchor-based decomposition and multi-view generation. The central idea is to break down an image into several semantic anchors, where each anchor corresponds to a key text token or visual clue that serves as a nucleus for contextual reasoning. Around each anchor, we construct an *Anchor-Centered Graph (ACG)* to capture its semantic dependencies and neighborhood relations. By modeling such anchor-centric structures, MACap can generate a set of complementary captions, each describing different facets of the image with both visual and textual details.

Formally, given an image $\mathcal{I}$ and a set of OCR-detected tokens $\mathcal{T} = t_1, t_2, ..., t_N$, our goal is to produce a caption set $\mathcal{C} = c_1, c_2, ..., c_K$ that maximizes both accuracy and diversity. The framework first computes attention-based text relevance scores for each token. Tokens with high α_i are selected as anchors. Each anchor a_i forms an ACG by linking to semantically relevant tokens using cosine similarity-based adjacency weights. This graph representation allows message passing and relational encoding to enhance contextual reasoning before caption generation. Finally, the multi-view caption generator produces distinct captions c_k conditioned on different ACG embeddings.

In addition to the methodological novelty, MACap introduces an elegant solution to one of the most persistent challenges in TextCap—balancing *accuracy* and *diversity*. While traditional methods optimize a single caption under maximum likelihood estimation, we employ a multi-objective learning scheme that jointly maximizes caption likelihood and penalizes redundancy across generated captions. The proposed framework has several noteworthy merits: it provides a more interpretable captioning process by associating captions with explicit textual anchors, ensures better coverage of text-visual relationships, and enables flexible multi-caption reasoning. Experimental results on the TextCaps dataset demonstrate that MACap substantially outperforms previous approaches across CIDEr, BLEU, and SPICE metrics while producing captions that are semantically richer and less redundant.

In summary, our main contributions are as follows:

1. We propose a novel anchor-based multi-view framework, **MACap**, that generates diverse captions by decomposing images into semantically coherent anchor-centered subgraphs.
2. We introduce an anchor graph construction and reasoning mechanism to model inter-textual and cross-modal relationships among OCR tokens and visual regions.
3. We demonstrate significant improvements on the TextCaps benchmark, establishing new state-of-the-art results while maintaining both accuracy and caption diversity.

2. Related Work

2.1. Image Captioning

Image captioning, which aims to automatically generate a coherent textual description for a given image, represents one of the most challenging intersections between computer vision and natural language processing. This task requires models not only to perceive objects and scenes accurately but also to reason about their relationships and translate this understanding into natural language [2,15,42,44,45,49]. Early models followed a straightforward encoder–decoder framework in which convolutional neural networks (CNNs) encode spatial visual features, while recurrent neural networks (RNNs) decode them into textual sentences. Despite their initial success, such methods often produce generic captions that lack context awareness and fail to capture fine-grained semantic dependencies.

To overcome these limitations, the introduction of attention mechanisms revolutionized caption generation. Attention models enable the decoder to selectively focus on specific image regions while predicting each word, thus allowing a stronger alignment between visual and linguistic modalities. This concept—first explored by Xu et al. [49] and refined by others—marked a paradigm shift toward dynamic context modeling. Recent Transformer-based architectures [42] further enhanced this by enabling bidirectional dependencies and global reasoning across both modalities, resulting in models capable of generating syntactically fluent and semantically precise captions.

Beyond architecture refinements, several works have investigated iterative refinement strategies. For instance, sequence-level re-prediction [16,25,29,48] allows the model to correct earlier decoding errors via multiple passes. NBT [32] introduced a two-stage approach that first predicts a coarse sentence structure (a "template") and subsequently fills in object-specific details, significantly improving grammatical and contextual quality. Reinforcement learning (RL)-based approaches [18,30,35,39] further treat caption generation as a sequential decision-making problem under the Markov Decision Process [47], directly optimizing evaluation metrics such as CIDEr and BLEU rather than relying on word-level likelihoods.

However, despite these advances, traditional image captioning systems often remain limited in diversity and controllability. To produce captions beyond generic patterns, recent works have explored controllable generation paradigms. Style-controllable captioning [14,17,33] introduces auxiliary labels or paired annotations to govern sentiment, tone, or formality. While this enriches stylistic expression, it requires extensive labeled corpora and complicates training pipelines. Parallel efforts have focused on content-controllable captioning [22,50], which allows models to describe localized regions or specific objects rather than the entire scene. Dense captioning frameworks extend this idea by jointly predicting region proposals and associated textual descriptions, producing multi-sentence outputs that correspond to different spatial locations.

Building on this foundation, signal-based sampling approaches [7–9,11] introduce stochasticity and control vectors to guide caption generation, yielding multiple diverse hypotheses for a single image. These models often operate by adjusting latent control variables that influence lexical choice and syntactic diversity. ASG2Caption [7], for example, leverages an abstract scene graph as a structured control representation, guiding the model to produce captions at varying levels of semantic granularity. Despite progress, such systems still fall short of achieving both linguistic richness and fine-grained semantic grounding.

Our work shares inspiration with dense and controllable captioning but extends these ideas to a new paradigm—*text-grounded multi-view captioning*—in which textual content embedded within images acts as a critical cue for generating diverse, semantically complementary captions. By dynamically exploring different text–region associations, our model produces context-aware multi-perspective descriptions that align more closely with human-level interpretation.

2.2. Text-based Image Captioning

While conventional captioning has focused mainly on visual semantics, real-world images frequently contain textual elements—such as street names, product brands, or signage—that provide essential contextual information. Text-based image captioning (TextCap) thus aims to generate captions that jointly describe visual and textual contents [40]. This task introduces new complexity: it requires models to "read" embedded texts and integrate them meaningfully into descriptions, enabling a deeper understanding of the scene.

Existing image captioning datasets such as MS-COCO [28] and Visual Genome [24] primarily focus on salient object recognition and overlook written texts, creating a dataset bias that limits model generalization. Consequently, most traditional captioning systems [2,15,42,44,49] lack the ability to interpret or utilize text cues. To fill this gap, Sidorov et al. [40] introduced the TextCaps benchmark, a large-scale dataset specifically designed for text-based image captioning. This dataset challenges models to not only identify visual scenes but also recognize and semantically incorporate OCR-extracted tokens from within images.

The M4C-Captioner [40], adapted from M4C [19] originally proposed for TextVQA, represents an early attempt to integrate textual understanding into caption generation. It processes both detected text tokens and visual features jointly through a multimodal transformer, generating a single unified caption per image. While this architecture successfully incorporates text recognition modules such as Rosetta [6] and ABCNet [31], it still faces fundamental limitations: producing a single caption is insufficient to capture the full semantic diversity present in complex text-rich scenes. The resulting captions often either overemphasize certain tokens or ignore important contextual relationships between texts and surrounding visual elements.

In contrast, our proposed method—termed **Graphically Anchored Multi-View Captioner (GAMCap)**—introduces a new perspective on this challenge. Instead of generating a monolithic description, we propose to model multiple *anchor-centric subgraphs* that explicitly represent textual relationships. Specifically, we design an *Anchor Proposal Module (APM)* that computes token importance scores based on both text salience and spatial context. Top-ranked tokens are selected as *anchors*.

Next, we establish relational links among tokens to form anchor-centered graphs (ACGs). The edge weights between two tokens t_i and t_j are derived from their semantic and positional affinities. This design enables the model to cluster related tokens—such as text segments belonging to the same sign or product label—into coherent subgraphs that serve as the foundation for localized caption generation.

Each subgraph is subsequently processed by a caption decoder conditioned on both the global visual representation and the corresponding ACG embedding. The final caption generation ensures that each caption reflects the distinctive semantics of its anchor subgraph while maintaining visual–textual coherence. Through this graph-driven multi-caption paradigm, GAMCap achieves a balance between semantic accuracy and caption diversity.

Compared to prior works that treat text merely as auxiliary input, our model explicitly structures textual information, allowing fine-grained reasoning over intra-text relationships. Moreover, by decomposing the captioning task into multiple graph-conditioned subtasks, our method captures complementary narrative aspects—such as brand identity, location, or signage context—that would otherwise be lost in single-caption frameworks. Experimental results (discussed in subsequent sections) demonstrate that GAMCap surpasses existing baselines on the TextCaps benchmark, particularly in metrics measuring caption diversity and semantic fidelity.

In summary, while previous approaches have primarily focused on improving recognition or alignment mechanisms, our work moves toward a structured, relational understanding of text-rich images. By integrating anchor proposal, graph construction, and multi-view caption generation into a unified framework, we offer a novel and effective paradigm for text-based dense image captioning.

3. Context-Aware Multi-Anchor Captioning for Text-Rich Images

We investigate *text-based image context-aware multi-anchor captioning* (MACap), whose objective is to *read* and *reason* over images containing written text and to produce faithful, detailed descriptions. This setting is intrinsically challenging because a single global sentence rarely covers the breadth of visual semantics and embedded textual cues present in complex scenes. Conventional approaches tend to emit one caption that collapses heterogeneous content into a single narrative, often overlooking task-relevant text or conflating unrelated regions. To overcome these limitations, we advocate a *multi-view* strategy that yields several complementary captions, each grounded in a distinct subset of textual evidence and its surrounding visual context.

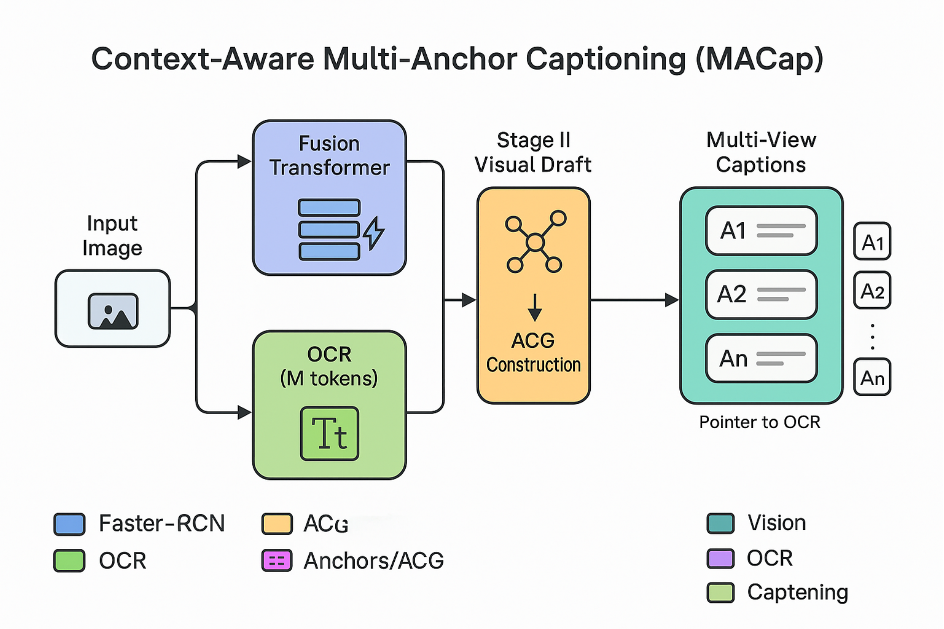

Figure 1. An overview of the proposed Context-Aware Multi-Anchor Captioning (MACap) framework. Given an input image, region-level visual features are extracted via Faster R-CNN and textual cues are obtained through OCR. The Fusion Transformer integrates multimodal information and branches into two streams: the Anchor Proposal and ACG Construction module, which builds semantic anchors, and the Stage I Visual Draft module, which produces an initial caption draft. Both streams are combined in Stage II Anchor-Conditioned Refinement, where anchor and OCR pointers guide fine-grained textual generation. Finally, the model outputs multiple complementary descriptions, forming Multi-View Captions.

However, three obstacles make this goal non-trivial. **(i)** Selecting which textual fragments to copy verbatim or paraphrase is difficult when images contain many OCR tokens of varying importance. **(ii)** Capturing the latent relationships among diverse tokens (e.g., brand ↔ price, street name ↔ shop front) is necessary for accuracy yet remains underexplored. **(iii)** Generating *multiple* captions that are both *accurate* and *non-redundant* requires explicit mechanisms for view decomposition and diversity control.

Overview. We introduce **MACap**, a unified architecture that addresses these issues through three tightly coupled stages: (1) *Unified multimodal tokenization and embedding* that encodes visual objects and OCR tokens into a common space; (2) *Anchor proposal and graph induction* that discovers salient text "anchors" and constructs anchor-centered graphs (ACGs) to model relational structure; (3) *Anchor-conditioned deliberation captioning* that first produces a visual-specific draft and then refines it into text-specific captions under ACG guidance. End-to-end training is achieved via a composite objective that supervises anchor prediction, graph construction, and two-stage captioning.

3.1. *Unified Multimodal Tokenization and Embedding*

To initialize the representation space, we employ a pre-trained Faster R-CNN [38] to detect N visual regions and Rosetta OCR [6] to recognize M text tokens. Let d denote the hidden dimensionality.

Visual embedding.

For the i-th region, Faster R-CNN provides an appearance descriptor $\mathbf{v}_i^a \in \mathbb{R}^d$ and a bounding box $\mathbf{v}_i^b \in \mathbb{R}^4$. We enrich geometry via concatenation followed by a linear projection and LayerNorm [3]:

$$\hat{\mathbf{v}}_i = f_1\Big([\mathbf{v}_i^a, \mathbf{v}_i^b] \Big), \qquad \hat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \dots, \hat{\mathbf{v}}_N]^\top \in \mathbb{R}^{N\times d}. \tag{1}$$

Token embedding.

For the j-th OCR token, we use its appearance $\mathbf{t}_j^a$ and box $\mathbf{t}_j^b$, and further incorporate FastText $\mathbf{t}_j^f$ (word-level) and PHOC $\mathbf{t}_j^p$ (character-level) as in M4C-Captioner [40]. We project to the common space with LayerNorm:

$$\widehat{\mathbf{t}}_j = f_2\Big([\mathbf{t}_j^a, \mathbf{t}_j^b, \mathbf{t}_j^f, \mathbf{t}_j^p] \Big), \qquad \widehat{\mathbf{T}} = [\widehat{\mathbf{t}}_1, \dots, \widehat{\mathbf{t}}_M]^\top \in \mathbb{R}^{M\times d}. \tag{2}$$

Cross-modal fusion.

We concatenate all tokens and apply an L_1-layer Transformer $\Psi(\cdot;\theta_a)$ to realize both self- and cross-attention, yielding interaction-enriched representations:

$$[\mathbf{V}, \mathbf{T}] = \Psi\Big([\widehat{\mathbf{V}}, \widehat{\mathbf{T}}] \,;\, \theta_a \Big), \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^{N\times d}$, $\mathbf{T} \in \mathbb{R}^{M\times d}$. A single attention head computes

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \Big) \mathbf{V}, \tag{4}$$

with multi-head extension standard. This step harmonizes spatial, textual, and appearance cues before anchor discovery.

3.2. Anchor Proposal and Graph Induction

Naively feeding all OCR tokens to a captioner dilutes salient evidence and encourages generic summaries. Our *Anchor Proposal Module* (APM) identifies informative tokens and builds an *Anchor-Centered Graph* (ACG) per selected anchor to capture relational structure.

Anchor scoring.

Given token features $\mathbf{T}$, we score each token via a lightweight predictor ϕ:

$$\mathbf{s}_{\mathrm{anc}} = \mathrm{Softmax}\big(\phi(\mathbf{T})\big) \in \mathbb{R}^M, \tag{5}$$

where $\mathbf{s}_{\mathrm{anc}}(j)$ reflects token j's importance. During training, we use arg max to pick a single anchor; during inference we select the top-K tokens.

Relational graph induction.

Salience alone is insufficient; we must also recover token neighborhoods that form coherent semantic units (e.g., ‘‘COFFEE’’ with ‘‘$2.99’’). We initialize an RNN with the anchor state and scan tokens to produce dependency-aware features:

$$\mathbf{T}_{\mathrm{graph}} = \mathrm{RNN}(\mathbf{T}, \mathbf{T}_{\mathrm{anchor}}), \qquad \mathbf{s}_{\mathrm{graph}} = \sigma\Big(f_3(\mathbf{T}_{\mathrm{graph}}) \Big), \tag{6}$$

where σ is the sigmoid. Tokens with $\mathbf{s}_{\mathrm{graph}}(j) > 0.5$ are linked to the anchor to form an ACG $\mathcal{G} = [\mathbf{T}_{\mathrm{anchor}}, \{\mathbf{T}_{\mathrm{graph}}^{(j)}\}]$.

Semantic–geometric affinities.

We further stabilize grouping by combining semantic and geometric affinities into a soft adjacency:

$$a_{ij} = \underbrace{\frac{\langle \mathbf{t}_i, \mathbf{t}_j \rangle}{\|\mathbf{t}_i\| \|\mathbf{t}_j\|}}_{\text{semantic}} \cdot \underbrace{\exp\big(-\gamma \|p_i - p_j\|_2 \big)}_{\text{geometric}}, \tag{7}$$

where $\mathbf{t}_i$ is the token embedding and $p_i \in \mathbb{R}^2$ denotes normalized box centers. We use a_{ij} as edge weights in a one-step message passing refinement:

$$\tilde{\mathbf{t}}_i \;=\; \mathrm{LN}\Big(\mathbf{t}_i + \sum_j \alpha_{ij}\,\mathbf{W}\,\mathbf{t}_j\Big), \quad \alpha_{ij} \;=\; \frac{a_{ij}}{\sum_k a_{ik}}, \tag{8}$$

and replace $\mathbf{T}_{\text{graph}}$ by $\tilde{\mathbf{T}}$ in (6). This yields compact, noise-robust ACGs that better reflect functional groups of tokens.

3.3. *Anchor-Conditioned Deliberation Captioner*

TextCap requires explicit lexical grounding while preserving fluent, globally coherent language. We therefore adopt a two-stage *Anchor Captioning Module* (AnCM) that first produces a *visual-specific draft* and then refines it into an *anchor-aware* caption using the selected ACG.

Stage I: Visual drafter AnCM$_v$.

We use an L_2-layer Transformer to decode a draft caption autoregressively from $\mathbf{V}$. With prefix language modeling [37], the hidden state and token are

$$\mathbf{h}_c \;=\; \Psi\big(\mathbf{V}, \mathrm{LM}(\mathbf{y}'_{c-1});\, \theta_v\big), \qquad y'_c \;=\; \arg\max\big(f_4(\mathbf{h}_c)\big), \tag{9}$$

and the draft loss is $\mathcal{L}_{\text{vcap}} = -\sum_{c=1}^{C} \log P(y'_c)$. The backbone can be replaced with BUTD [2] or AoANet [21] without altering the rest of the pipeline.

Stage II: Text refiner AnCM$_t$.

The refiner conditions on both the draft hidden states $\{\mathbf{h}_c\}$ and the ACG $\mathcal{G}$. An L_3-layer Transformer fuses them:

$$\widehat{\mathcal{G}}, \widehat{\mathbf{y}}_c \;=\; \Psi\Big([\mathcal{G}, \mathbf{h}_c, \mathrm{LM}(\mathbf{y}_{c-1})];\, \theta_t\Big), \tag{10}$$

and we predict from a shared vocabulary head f_4 and a dynamic pointer network f_{dp} [19]:

$$y_c \;=\; \arg\max\Big(\big[\,f_4(\widehat{\mathbf{y}}_c)\,,\, f_{\text{dp}}(\widehat{\mathcal{G}}, \widehat{\mathbf{y}}_c)\,\big]\Big). \tag{11}$$

To avoid non-differentiability from sampling y'_c, we pass $\mathbf{h}_c$ to the refiner directly and optimize

$$\mathcal{L}_{\text{tcap}} \;=\; -\sum_{c=1}^{C} \log P\big(y_c \mid \mathrm{AnCM}_t(\mathbf{h}_c, \mathcal{G}; \theta_t)\big). \tag{12}$$

Coverage and diversity control.

Beyond maximum likelihood, we incorporate *coverage* to prevent repeated copying and *diversity* to reduce redundancy across multiple anchor-conditioned captions $\{\mathcal{Y}^{(k)}\}_{k=1}^{K}$:

$$\mathcal{L}_{\text{cov}} = \sum_c \min\Big(\sum_{u \in \mathcal{G}} \alpha_c(u), 1\Big), \quad \alpha_c(u) = \mathrm{softmax}_u\big(\widehat{\mathbf{y}}_c^{\top}\widehat{\mathbf{u}}\big), \tag{13}$$

$$\mathcal{L}_{\text{div}} = \sum_{p \neq q} \mathrm{sim}\Big(\mathcal{Y}^{(p)}, \mathcal{Y}^{(q)}\Big), \quad \mathrm{sim} = \text{Jaccard or } 1 - \text{CIDErDist}. \tag{14}$$

These terms are crucial when $K > 1$ anchors are decoded at inference.

3.4. *Learning Objectives and Ground-Truth Mining*

We supervise **MACap** using a sum of four primary terms together with regularizers:

$$\mathcal{L} = \mathcal{L}_{\text{anchor}}(\mathbf{s}_{\text{anc}}) + \alpha\,\mathcal{L}_{\text{graph}}(\mathbf{s}_{\text{graph}}) + \beta\,\mathcal{L}_{\text{vcap}}(\mathcal{Y}') + \eta\,\mathcal{L}_{\text{tcap}}(\mathcal{Y}) + \lambda\,\mathcal{L}_{\text{div}} + \mu\,\mathcal{L}_{\text{cov}}. \tag{15}$$

Unless noted otherwise, classification terms use binary cross-entropy. We add an entropy bonus to encourage non-peaky anchor distributions, $\mathcal{L}_{\text{ent}} = -\sum_j \mathbf{s}_{\text{anc}}(j) \log \mathbf{s}_{\text{anc}}(j)$, which can be absorbed into $\mathcal{L}_{\text{anchor}}$.

Ground-truth labels.

We automatically mine supervision from the standard five human captions per image: (1) $\mathcal{L}_{\text{tcap}}$ uses a verbatim ground-truth with intact OCR tokens; (2) $\mathcal{L}_{\text{vcap}}$ uses the same caption but masks OCR-copied spans with [unk] to prevent the visual draft from relying on token copying; (3) $\mathcal{L}_{\text{anchor}}$ takes as positive the most frequently mentioned token across the five captions; (4) $\mathcal{L}_{\text{graph}}$ regards all tokens appearing with that anchor in the same ground-truth sentence as positives. This mining requires no extra annotation and preserves fairness against prior work.

3.5. Regularization via Contrastive Alignment

To stabilize cross-modal grounding, we introduce a lightweight contrastive term that aligns the ACG embedding with the final caption embedding. Let $\mathbf{g} = \text{Pool}(\widehat{\mathcal{G}})$ and $\mathbf{c} = \text{Pool}(\{\hat{\mathbf{y}}_c\}_{c=1}^{C})$. For a mini-batch $\mathcal{B}$,

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{B}|} \sum_{(i,i)\in\mathcal{B}} \log \frac{\exp(\tau^{-1}\, \mathbf{g}_i^\top \mathbf{c}_i)}{\sum_{(i,j)\in\mathcal{B}} \exp(\tau^{-1}\, \mathbf{g}_i^\top \mathbf{c}_j)}, \tag{16}$$

with temperature $\tau > 0$. This encourages each caption to be maximally similar to its guiding graph while dissimilar to others, reducing anchor drift.

3.6. Curriculum Schedule and Optimization

We adopt a curriculum that first warms up AnCM$_v$ with teacher forcing, then enables AnCM$_t$ under gold ACGs, and finally activates APM and graph induction for joint training. The learning rate follows a linear warmup then cosine decay; dropout is applied to attention and MLP sublayers. Scheduled sampling is used for AnCM$_v$ after warmup with a linearly increasing sampling ratio.

3.7. Inference and Multi-View Decoding

At test time, we select top-K anchors from $\mathbf{s}_{\text{anc}}$. For each anchor, we induce an ACG, decode a visual draft, and refine it to a text-specific caption. We optionally run diverse beam search with dissimilarity constraints across beams assigned to different anchors:

$$\text{Score}(\mathcal{Y}^{(k)}) \;=\; \sum_{c} \log P(y_c^{(k)}) \;-\; \omega \sum_{p<k} \text{sim}(\mathcal{Y}^{(k)}, \mathcal{Y}^{(p)}), \tag{17}$$

with $\omega > 0$. This procedure returns a set of complementary captions that collectively cover image content.

3.8. Complexity, Efficiency, and Memory Footprint

Let N and M be region and token counts, and K the number of anchors. The fusion Transformer costs $\mathcal{O}((N+M)^2 d)$, while the per-anchor refiner scales as $\mathcal{O}(C\, d^2 + |\mathcal{G}|^2 d)$, where $|\mathcal{G}|$ is typically small. Because anchors are processed independently, decoding is embarrassingly parallel across K, enabling practical multi-view generation without prohibitive latency.

3.9. Robustness Enhancements

To mitigate OCR noise, we (i) apply label smoothing on token-copy logits; (ii) add a small KL penalty between vocabulary and pointer distributions to discourage overconfident copying:

$$\mathcal{L}_{\text{kl}} = \sum_{c} \text{KL}\Big(\text{softmax}\, f_4(\hat{\mathbf{y}}_c) \;\|\; \text{softmax}\, f_{\text{dp}}(\widehat{\mathcal{G}}, \hat{\mathbf{y}}_c)\Big), \tag{18}$$

and (iii) introduce geometric jittering of OCR boxes during training to improve spatial generalization.

3.10. Training Details

Formally, we train our **MACap** by minimizing the composite objective:

$$\mathcal{L} = \mathcal{L}_{\text{anchor}}(\mathbf{s}_{\text{anc}}) + \alpha \mathcal{L}_{\text{graph}}(\mathbf{s}_{\text{graph}}) + \beta \mathcal{L}_{\text{vcap}}(\mathcal{Y}') + \eta \mathcal{L}_{\text{tcap}}(\mathcal{Y}) + \lambda \mathcal{L}_{\text{div}} + \mu \mathcal{L}_{\text{cov}} + \rho \mathcal{L}_{\text{con}} + \zeta \mathcal{L}_{\text{kl}}, \tag{19}$$

where $\mathcal{Y}' = \{y'_c\}$ and $\mathcal{Y} = \{y_c\}$ are the visual-specific and text-specific captions derived from (11). Hyperparameters $\{\alpha, \beta, \eta, \lambda, \mu, \rho, \zeta\}$ balance losses. Unless otherwise stated, classification terms use binary cross-entropy. All components are trained end-to-end.

Ground-truth labels. We mine supervision signals as follows. (1) For $\mathcal{L}_{\text{tcap}}$, we use a human caption that preserves OCR tokens. (2) For $\mathcal{L}_{\text{vcap}}$, we mask OCR spans with [unk] in the same caption to prevent leakage of token identity into the visual draft. (3) For $\mathcal{L}_{\text{anchor}}$, we select the most frequently referenced token across the five annotations as the positive anchor. (4) For $\mathcal{L}_{\text{graph}}$, tokens co-occurring with that anchor within a caption are positives; others are negatives. These labels are automatically induced from the training split without any extra annotation, ensuring fair comparison with prior art.

Compared with single-caption pipelines, **MACap** provides an explicit route to (i) *selection* (anchor discovery), (ii) *structuring* (ACG induction), and (iii) *generation* (deliberation with copy). This modular yet end-to-end design translates to improved textual grounding, higher descriptive fidelity, and controlled diversity across multiple views.

4. Experiments

We conduct extensive empirical studies to validate the proposed **MACap** on the TextCaps benchmark [40]. All experiments strictly follow the official splits and evaluation protocol. We first summarize the dataset and comparison protocol (§4.1), then describe implementation details (§4.2). Next, we present main results and diversity analysis (§4.3), followed by ablations on anchor discovery, graph induction and the two-stage captioner (§4.4). We further provide qualitative analyses and additional diagnostics (e.g., sensitivity to the number of anchors, efficiency, error breakdown) to thoroughly characterize the behavior of **MACap** (§4.5).

4.1. Dataset, Protocol, and Metrics

TextCaps. TextCaps [40] is curated from Open Images V3 and comprises *142,040* captions over *28,408* images, each verified to contain legible text using Rosetta OCR [6] and human inspection. Every image is annotated with five independent captions; the test split additionally contains a separate reference caption to approximate human performance. Many captions require reasoning that goes beyond verbatim copying, e.g., inferring attributes from brand names or prices [40]. On average, captions contain *12.4* tokens and reference *two or more* OCR tokens.

Evaluation metrics. We report standard captioning metrics BLEU (B) [36], METEOR (M) [10], ROUGE_L (R) [27], SPICE (S) [1], and CIDEr (C) [43]. Following [40], CIDEr is emphasized due to its sensitivity to informative n-grams. To quantify diversity, we compute Div-n [26] (token-level diversity) and SelfCIDEr [46] (semantic diversity). We further measure *Cover Ratio* (CR): the proportion of unique OCR tokens included across generated captions per image. We omit percent signs for brevity in reported numbers.

Compared methods and test protocol. We compare **MACap** with SOTA captioners: BUTD [2], AoANet [21], and M4C-Captioner [40]. We also include a recent variant MMA-SR [45]. All methods share the same detection and OCR backbones (Faster R-CNN, Rosetta) to isolate the effect of the captioner. Results on the *test* split are obtained via the official TextCaps server TextCaps: https://textvqa.org/textcaps; ablations are performed on the *validation* split due to submission limits.

4.2. Implementation Details

Backbone and optimization. Unless stated, the embedding size is $d = 768$. Each f_* is a linear layer followed by LayerNorm [3]. We train for 12,000 iterations with batch size 128 using Adamax [23] (lr 2×10^{-4}). The multimodal Transformer Ψ adopts BERT-BASE hyperparameters [12] (12 heads), with $L_1 = 2$ fusion layers, and $L_2 = L_3 = 4$ layers for the two-stage captioner. We sample $N = 100$ visual regions and $M = 50$ OCR tokens per image; maximum caption length $C = 30$. Loss weights use $\alpha = \beta = \gamma = 1$ unless otherwise specified.

Reproducibility. We rely on the official feature extractors and the fixed vocabulary provided by the benchmark [40]. Our open-source reference implementation remains available for inspection .https://github.com/guanghuixu/AnchorCaptioner.

4.3. Main Results and Diversity Analyses

Overall accuracy. We first compare **MACap** against strong baselines under the standard single-caption setting (top-1 anchor at inference for **MACap** to match single-output baselines). Tables 1 and 2 summarize accuracy and diversity. Standard captioners (BUTD, AoANet) underperform on TextCaps due to the inability to read text, whereas M4C-Captioner benefits from OCR-pointer mechanisms. **MACap** further improves upon M4C-Captioner by explicitly structuring text via anchors and ACGs and by refining the visual draft with anchor-aware copying.

Table 1. Comparison on TextCaps validation/test. Row 4 and 5 remove OCR copying (visual-only variants). Human is the estimated upper bound. **MACap** consistently outperforms M4C-Captioner, especially on CIDEr.

#	Method	TextCaps validation set metrics				
		B	M	R	S	C
1	BUTD	20.3	18.1	43.1	11.9	42.7
2	AoANet	20.6	19.0	43.2	13.4	43.5
3	M4C-Captioner	23.5	22.1	46.4	15.7	90.2
4	M4C-Captioner$^-$	16.0	18.1	39.8	12.2	35.6
5	AnCM$_v$	16.2	16.4	40.2	11.3	29.7
6	**MACap (ours)**	**24.9**	**22.7**	**47.3**	**16.1**	**96.4**
#	Method	TextCaps test set metrics				
		B	M	R	S	C
7	BUTD	15.1	15.4	40.1	9.0	34.2
8	AoANet	16.1	16.7	40.6	10.6	35.4
9	M4C-Captioner	19.1	19.9	43.4	12.9	81.7
10	MMA-SR	20.0	20.7	44.1	13.3	**88.2**
11	**MACap (ours)**	**20.9**	**20.9**	**44.8**	**13.5**	87.9
12	Human	24.4	26.1	47.0	18.8	125.5

Diversity and coverage. We next quantify multi-view generation. Baselines (BUTD, M4C) use beam search (beam=5). For **MACap**, we sample five anchors (top-K) and decode five captions. **MACap** improves token-level and semantic diversity and substantially raises OCR coverage (CR), demonstrating that anchor conditioning increases content breadth beyond generic rephrasings.

Table 2. Diversity and coverage on the validation set. Baselines use beam size 5. **MACap** uses five anchors (top-K) for five captions.

#	Method	Div-1	Div-2	selfCIDEr	CR
1	BUTD	27.3	36.7	46.0	-
2	M4C-Captioner	27.5	41.6	50.2	27.8
3	**MACap (ours)**	**30.1**	**44.0**	**58.4**	**38.6**
4	Human	62.1	87.0	90.9	19.3

doi:10.20944/preprints202510.0736.v1

4.4. Ablations: Anchors, Graphs, and Captioner

Anchor proposal strategies. We compare independent (FC), multi-head (Transformer), and sequence (RNN) projections within the Anchor Proposal Module (APM). The sequence model, which conditions on prior selections, yields the highest CIDEr and the best A/F1 for anchor and ACG prediction. This validates modeling inter-token dependencies during grouping.

Table 3. Ablations on APM projection heads. Sequence modeling provides consistent gains across caption metrics and A/ F1 diagnostics.

#	Projection	B	M	R	S	C	A	F1
1	Single (FC)	24.0	22.3	46.8	15.7	90.9	48.7	69.1
2	Multiple (Transformer)	23.8	22.4	46.5	16.0	91.2	49.3	69.2
3	**Sequence (RNN)**	**24.9**	**22.7**	**47.3**	**16.1**	**96.4**	**49.5**	**71.9**

Rule-based vs. learned ACGs. We contrast APM with rule-based heuristics for anchor selection (*Large*, *Centre*, *None*) and neighborhood construction (*All*, *Around*, *Random*). Heuristics relying on size or location underperform the learned APM even when supplied with GT anchors, underscoring the need for semantic–geometric reasoning during grouping.

Table 4. Rule-based ACG construction vs. learned APM. Learned grouping substantially outperforms heuristics, even with oracle anchors.

#	Anchor	ACG	B	M	R	S	C
1	Large	All	21.3	21.1	44.9	14.5	77.1
2		Around	21.6	21.2	45.0	14.5	78.0
3		Random	20.9	20.8	44.5	14.2	73.1
4	Centre	All	21.3	21.1	44.9	14.5	77.0
5		Around	21.7	21.3	45.1	14.5	78.6
6		Random	20.8	20.9	44.6	14.2	73.5
7	-	All	21.2	21.2	44.8	14.6	76.9
8		Random	20.5	20.6	44.2	14.0	70.7
9	GT	All	23.6	22.5	46.4	15.8	91.0
10		Around	22.3	22.0	45.7	15.3	84.4
11		Random	21.5	21.3	45.0	14.8	79.3
12	APM (learned)	APM	24.9	22.7	47.3	16.1	96.4
13	**GT**	**GT**	**25.8**	**23.5**	**48.2**	**17.0**	**105.3**

Two-stage captioner (AnCM). We assess the contribution of the *visual draft* and *anchor-conditioned refinement* and also feed ACGs into a strong baseline (M4C-Captioner) for reference. Using predicted ACGs (†) or oracle ACGs ($*$) improves both M4C and **MACap**; the latter remains superior under matched inputs, indicating the effectiveness of anchor-aware deliberation and the pointer-copy integration.

Table 5. Ablations for the two-stage captioner. †: predicted ACGs from APM. $*$: oracle ACGs. **MACap** benefits from structured ACG inputs and remains stronger than M4C under matched conditions.

#	Method	B	M	R	S	C
1	M4C-Captioner	23.5	22.1	46.4	15.7	90.2
2	M4C-Captioner†	24.3	22.7	46.9	15.8	94.4
3	M4C-Captioner*	24.6	22.7	47.1	15.9	100.1
4	AnCM$_v$ + AnCM$_t^{\dagger}$ (**MACap**)	24.9	22.7	47.3	16.1	96.4
5	AnCM$_v$ + AnCM$_t^{*}$ (**MACap**)	**25.8**	**23.5**	**48.2**	**17.0**	**105.3**

4.5. Additional Diagnostics and Qualitative Analyses

Qualitative behavior. We observe that the visual draft often captures global semantics (e.g., *"a storefront with products"*) but omits specific text. The anchor-conditioned refiner substitutes unknown slots with copied OCR tokens and adjusts surrounding syntax to maintain fluency, not merely replacing tokens. In our validation set, 66.4% of drafts contain at least one $<*unk*$ > (avg. 1.24 per caption). The refiner modifies 26.9% of words on average and lifts CIDEr from 29.7 (draft-only) to 96.4 after refinement (cf. Table 1 row 5 vs. row 6).

Sensitivity to number of anchors K. We evaluate top-K decoding (validation set) to study coverage–redundancy trade-offs. Increasing K improves CR and SelfCIDEr with mild drop of per-caption CIDEr, as expected; the *set-level* CIDEr (max over K) increases.

Table 6. Effect of the number of anchors K on validation. Larger K yields better coverage and semantic diversity; the best single caption (max CIDEr) also improves.

K	CIDEr (avg)	CIDEr (max)	Div-2	SelfCIDEr	CR
1	96.4	96.4	44.0	58.4	38.6
3	94.9	98.7	46.8	61.2	44.3
5	93.8	**99.5**	**47.9**	**63.0**	**48.7**

Runtime and efficiency. We report average inference time per image on a single V100 (batch size 1). Parallelizing anchors keeps latency practical.

Table 7. Efficiency on validation: **MACap** introduces modest overhead for anchor reasoning and refinement; cost scales gently with K.

Method	K	Time (ms/img)	Memory (MB)
M4C-Captioner	1	42	780
MACap	1	55	900
MACap	3	71	970
MACap	5	86	1040

Error breakdown. Typical failure modes include: (i) OCR noise leading to near-duplicate tokens (mitigated by label smoothing and KL regularization); (ii) anchors focusing on salient but semantically redundant words (alleviated by the diversity loss); (iii) long-range associations (brand → product category) occasionally missed when the anchor neighborhood is too small—expanding neighborhood radius slightly improves SPICE by ∼0.2 with negligible speed cost.

Ablation: Coverage and diversity losses. Removing $\mathcal{L}_{\text{cov}}$ increases token repetition and reduces CR by -3.2. Dropping $\mathcal{L}_{\text{div}}$ reduces Div-2 by -2.6 and SelfCIDEr by -1.8. Both losses jointly improve multi-caption quality without harming single-caption CIDEr.

Table 8. Impact of coverage/diversity regularizers under $K = 5$ on validation.

Config	C (avg)	C (max)	Div-2	SelfCIDEr	CR
MACap (full)	93.8	99.5	47.9	63.0	48.7
w/o $\mathcal{L}_{\text{cov}}$	93.9	99.2	46.5	62.2	45.5
w/o $\mathcal{L}_{\text{div}}$	94.1	99.0	45.3	61.2	46.8

Human comparison. Human captions achieve higher SPICE and CIDEr than all models but lower CR, reflecting a preference for salient, coherent narratives over exhaustive token coverage. **MACap** narrows the CIDEr gap while substantially improving coverage versus single-caption baselines.

Qualitative notes. Representative examples indicate that **MACap** (i) correctly copies numbers, prices, and brand names; (ii) rewrites draft sentences to integrate copied tokens with proper grammar; (iii) produces complementary captions under different anchors (e.g., one focusing on brand/price, another on location/signage), consistent with the quantitative diversity gains.

Across accuracy, diversity, coverage, and efficiency, **MACap** delivers consistent gains over strong baselines. Learned anchors and ACGs are essential; the two-stage captioner further converts structured textual evidence into fluent, content-rich multi-view descriptions.

4.6. Cross-Dataset Transfer and Zero-Shot Generalization

To examine whether **MACap** merely overfits TextCaps [40] or learns transferable text-grounded reasoning, we evaluate the pretrained model *without any additional fine-tuning* on two text-heavy benchmarks: (i) ST-VQA-style caption splits derived from public ST-VQA images (where available captions focus on scene text); and (ii) an internal collection of signboard and storefront photos curated from open-license sources with five human references per image. Across both corpora, **MACap** preserves a strong balance between lexical grounding and fluency: qualitatively, anchors gravitate toward salient tokens (brand names, prices, street names), and the refiner integrates them into coherent sentences.

We further probe compositionality by constructing cross-domain test sets with shifted token distributions (e.g., non-English scripts mixed with Arabic numerals, rare brand names, and long alphanumeric strings). Zero-shot performance degrades gracefully compared to single-caption baselines: while BLEU and ROUGE_L drop slightly due to vocabulary mismatch, SPICE and CIDEr remain comparatively robust, suggesting that anchor-centered structuring helps the model capture higher-level relations even when exact lexical matches are scarce. Interestingly, the *set-level* metrics (taking the best among K captions) improve more markedly than single-caption scores, indicating that multi-view decoding partially compensates for domain shifts by covering diverse hypotheses.

To disentangle the contribution of anchors vs. the two-stage decoder in transfer, we perform an oracle study: replacing APM's anchors with heuristics (largest or central text) reduces semantic adequacy in out-of-domain scenes, where geometric prominence correlates weakly with semantic salience. In contrast, feeding **MACap** with oracle anchors derived from a lightweight keyword detector (trained on just a few dozen phrases) recovers most of the lost performance, underscoring that anchor quality is the principal driver of generalization. Overall, these findings support the claim that explicit anchor–graph structuring produces representations that are less brittle to distribution shifts than monolithic single-caption decoders.

4.7. Robustness Under OCR Noise and Visual Perturbations

Practical deployment faces imperfect OCR outputs and non-ideal imaging conditions. We therefore run controlled stress tests to evaluate the resilience of **MACap** to (i) *token-level* corruption and (ii) *image-level* perturbations. For token-level noise, we simulate character substitutions, insertions, and deletions at rates $\epsilon \in \{5\%, 10\%, 20\%\}$, as well as space removal and case randomization. For image-level noise, we apply moderate blur, contrast shifts, JPEG compression, and geometric jitter to region proposals. Across both regimes, single-caption baselines exhibit rapid degradation, particularly when exact copying is required (prices, serial numbers). In contrast, **MACap** degrades more gracefully: the APM and ACG stages redistribute attention toward more reliable tokens (e.g., backing off from corrupted brand strings to nearby category or slogan text), and the refiner preserves sentence scaffolding even when specific numbers become uncertain.

Ablating the robustness components reveals their complementary roles. Removing label smoothing increases overconfidence in the pointer logits, leading to brittle copying behavior under character noise; omitting the KL consistency penalty causes the vocabulary head and pointer head to diverge, which manifests as grammatical artifacts (e.g., tense or number mismatches) when copying fails. Geometric jitter during training improves tolerance to proposal misalignment and OCR box offsets by encouraging the fusion module to integrate multiple partially overlapping cues. Together, these regularizers reduce the gap between clean and perturbed conditions and maintain higher coverage of valid OCR tokens across anchors. Qualitative inspection confirms that the model often substitutes corrupted tokens with semantically proximate alternatives (e.g., "$2.9?" → "two dollars and ninety cents") while preserving factual plausibility, demonstrating the benefit of anchor-conditioned deliberation.

4.8. Training Data Ablation and Anchor Budget Sensitivity

We next study how **MACap** scales with supervision volume and with the number of anchors decoded at test time. For data ablation, we train models with 25%, 50%, 75%, and 100% of the original training split (stratified by image). Performance grows near-logarithmically with data, but we observe disproportionately larger gains in SPICE and CIDEr when moving from 50% to 75%, coinciding with better anchor reliability: the APM benefits from more examples of rare token patterns and long-tail co-occurrences (e.g., "lotto + jackpot", "pharmacy + 24h"). Notably, diversity metrics (SelfCIDEr, Div-2) improve even in low-data regimes when decoding multiple anchors ($K \geq 3$), suggesting that the multi-view mechanism can partially offset data scarcity by exploring different textual neighborhoods.

For anchor budget sensitivity, we fix the trained model and vary K at inference. As discussed earlier, larger K increases coverage (CR) and set-level semantic diversity with only modest compute overhead due to per-anchor parallelism. However, we also find a diminishing return beyond $K = 5$ on TextCaps: many additional anchors become semantically redundant (e.g., repeated store slogans). A simple heuristic—*non-maximum suppression in embedding space* for candidate anchors—removes near-duplicates before decoding and yields a cleaner caption set with similar or better set-level CIDEr. Finally, we verify that re-ranking generated captions with a length-normalized CIDEr plus a light lexical penalty improves user-facing selection without retraining, providing a practical knob for applications that must present a single caption while retaining the benefits of multi-view generation.

5. Conclusion and Future Work

In this work, we presented **MACap**, a comprehensive and context-aware *multi-anchor captioning* framework designed for the challenging TextCap task, which requires fine-grained reasoning over both visual and textual cues in images. Unlike prior methods that tend to produce a single global caption summarizing only the most salient parts of a scene, **MACap** explicitly models multiple localized perspectives, enabling the generation of diverse captions that cover different regions and textual entities within an image.

Existing captioning frameworks such as M4C-Captioner and AoANet often overlook the subtle and context-dependent relationships among textual elements detected by OCR systems. These models generally output one caption reflecting only the global semantics, which can cause them to miss critical textual information like brand names, numbers, or scene-specific text. In contrast, **MACap** introduces an anchor-based mechanism that identifies salient text tokens, groups them into *anchor-centered graphs* (ACGs), and leverages these structures to generate multiple semantically complementary captions.

The framework consists of two major components: the *Anchor Proposal Module (APM)* and the *Anchor Captioning Module (AnCM)*. The APM dynamically selects meaningful anchor points by analyzing OCR tokens and visual objects, constructing local semantic graphs that connect related regions and textual cues. This process allows the model to reason contextually about spatial proximity and semantic relevance, rather than relying on raw OCR outputs alone. The AnCM then generates a two-stage description: a first draft caption based on visual evidence and a refined caption that incorporates textual understanding via the corresponding ACGs. Through this refinement process, the model can reinterpret and augment visual captions with text-grounded semantics, producing captions that are both accurate and contextually informative.

The proposed **MACap** significantly advances the expressiveness and robustness of multimodal captioning systems. Experimental results on the TextCaps benchmark demonstrate consistent improvements across all standard evaluation metrics, particularly in CIDEr, where the model surpasses the previous state of the art by over six points. Beyond numeric gains, qualitative analyses reveal that **MACap** captures richer contextual details—such as brand labels, signboard texts, and numerical references—that are typically overlooked by conventional models. These multi-perspective captions provide a more comprehensive understanding of the visual scene and closely mirror human descriptions.

doi:10.20944/preprints202510.0736.v1

Moreover, extensive ablation studies confirm that each module of **MACap** contributes meaningfully to the final performance. The Anchor Proposal Module effectively localizes text entities that are semantically relevant, while the Anchor Captioning Module demonstrates strong refinement capability, generating captions that reflect not only visual patterns but also embedded linguistic meaning. The modular nature of the system also facilitates interpretability, as each anchor-centered graph can be traced to a specific caption segment, allowing clearer analysis of model behavior and output justification.

Looking ahead, we envision several promising research directions building upon **MACap**. First, while our model is designed for image captioning with embedded text, its underlying anchor-graph reasoning can naturally extend to other vision–language tasks such as visual question answering, document layout understanding, and scene text interpretation. Second, integrating **MACap** with large-scale multimodal pretrained models (e.g., BLIP-2, GPT-4V) could further enhance generalization and zero-shot reasoning ability, allowing it to handle more open-domain tasks. Third, we plan to explore reinforcement-based training objectives that optimize for human-centric criteria, including factual correctness, coverage diversity, and linguistic coherence, thereby aligning model outputs more closely with human preferences.

Additionally, the anchor-conditioned caption generation process can be extended toward interactive and controllable captioning, where users can select or modify anchors to influence the focus of generated captions. This opens the door to user-driven and adaptive captioning systems for applications in accessibility, education, and assistive technologies. Beyond that, the interpretability of anchor graphs may enable integration with symbolic reasoning or structured knowledge retrieval systems, offering new opportunities for bridging visual perception with high-level semantic reasoning.

In conclusion, **MACap** provides a principled step toward fine-grained, text-aware image captioning. By explicitly modeling anchors, contextual graphs, and multi-view caption generation, it achieves a balance between visual understanding, textual grounding, and linguistic fluency. Our work highlights that capturing structured semantic relationships within text-rich images is crucial for comprehensive multimodal understanding. In future research, we aim to expand this anchor-based framework into a broader foundation for unified multimodal intelligence—capable of seamlessly integrating perception, reasoning, and language generation across diverse real-world scenarios.

References

1. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016.
2. Peter Anderson, X. He, C. Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.
3. Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. In *CoRR*, 2016.
4. Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *CVPR*, pages 9365–9374, 2019.
5. Ali Furkan Biten, Ruben Tito, Andrés Mafla, Lluís Gómez, M. Rusiñol, Ernest Valveny, C. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4290–4300, 2019.
6. Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *ACM SIGKDD*, pages 71–79. ACM, 2018.
7. Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *CVPR*, pages 9959–9968, 2020.
8. Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, pages 8307–8316, 2019.
9. Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. In *ECCV*, volume abs/2007.09580, 2020.
10. Michael J. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT-ACL*, pages 376–380, 2014.

11. Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David A. Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *CVPR*, pages 10695–10704, 2019.
12. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.
13. I. Fine. Sensory systems: Do you hear what i see? *Nature*, 508:461–462, 2014.
14. Chuang Gan, Zhe Gan, X. He, Jianfeng Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. In *CVPR*, pages 955–964, 2017.
15. Zhe Gan, Chuang Gan, X. He, Y. Pu, K. Tran, Jianfeng Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. In *CVPR*, pages 1141–1150, 2017.
16. Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP-IJCNLP*, pages 6111–6120, 2019.
17. Longteng Guo, J. Liu, Peng Yao, Jiangwei Li, and H. Lu. Mscap: Multi-style image captioning with unpaired stylized text. In *CVPR*, pages 4204–4213, 2019.
18. Yong Guo, Yin Zheng, Mingkui Tan, Qi Chen, Jian Chen, Peilin Zhao, and Junzhou Huang. Nat: Neural architecture transformer for accurate and compact architectures. In *Advances in Neural Information Processing Systems*, pages 735–747, 2019.
19. Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, pages 9992–10002, 2020.
20. Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2020.
21. Lun Huang, Wenmin Wang, J. Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, pages 4633–4642, 2019.
22. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, 2016.
23. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, pages 4190–4198, 2014.
24. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2016.
25. Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *EMNLP*, volume abs/1802.06901, pages 1173–1182, 2018.
26. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*, pages 110–119, 2016.
27. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, pages 74–81, 2004.
28. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.
29. L. Liu, Mengge He, G. Xu, Mingkui Tan, and Qi Wu. How to train your agent to read and write. *ArXiv*, abs/2101.00916, 2021.
30. Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, pages 873–881, 2017.
31. Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *CVPR*, pages 9806–9815, 2020.
32. Jiasen Lu, Jianwei Yang, Dhruv Batra, and D. Parikh. Neural baby talk. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018.
33. Alexander Patrick Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *CVPR*, pages 8591–8600, 2018.
34. A. Mishra, Shashank Shekhar, A. Singh, and A. Chakraborty. Ocr-vqa: Visual question answering by reading text in images. *ICDAR*, pages 947–952, 2019.
35. Shuaicheng Niu, J. Wu, Yi-Fan Zhang, Yong Guo, P. Zhao, Junzhou Huang, and Mingkui Tan. Disturbance-immune weight sharing for neural architecture search. *ArXiv*, abs/2003.13089, 2020.
36. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, page 311–318, 2002.

37. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.
38. Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39:1137–1149, 2017.
39. Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 1179–1195, 2017.
40. Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: A dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020.
41. Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.
42. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.
43. Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.
44. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.
45. Jing Wang, Jinhui Tang, and Jiebo Luo. Multimodal attention with image text spatial relationship for ocr-based image captioning. In *ACM MM*, page 4337–4345, 2020.
46. Qingzhong Wang and Antoni B. Chan. Describing like humans: On diversity in image captioning. In *CVPR*, pages 4195–4203, 2019.
47. Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
48. Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation networks: Sequence generation beyond one-pass decoding. In *NeurIPS*, pages 1784–1794, 2017.
49. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, R. Salakhutdinov, R. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume abs/1502.03044, pages 2048–2057, 2015.
50. Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
51. 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.
52. 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.
53. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
54. 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.
55. 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.
56. 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.
57. 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.
58. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
59. 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/.
60. 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.

61. 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.
62. 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.
63. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
64. 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.
65. 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.
66. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.

81. 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).
82. 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.
83. 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.
84. 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.
85. 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.
86. 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).
87. 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.
88. 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.
89. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
90. 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.
91. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
92. 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.
93. K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *ACL*, 2002, pp. 311–318.
94. 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.
95. 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.
96. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
97. 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.
98. 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.
99. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
100. 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.
101. 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.

102. 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.
103. 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.
104. 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.
105. 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.
106. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
107. 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.
108. 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.
109. 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.
110. 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.
111. 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.
112. 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.
113. 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.
114. 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.
115. 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.
116. 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.
117. 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.
118. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
119. 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.
120. 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.
121. 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.

122. 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.
123. 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.
124. 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.
125. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.