Article

Not peer-reviewed version

Context-Guided Multi-Branch Fusion for Text-Dependent Visual Question Reasoning

Sander Ridder * , Noor Verbeeck , Callum Hensley , Luca Vandenberghe

Posted Date: 27 October 2025

doi: 10.20944/preprints202510.1994.v1

Keywords: visual question reasoning; multimodal fusion; OCR-aware understanding; adaptive routing; contextual alignment

Article

Context-Guided Multi-Branch Fusion for Text-Dependent Visual Question Reasoning

Sander Ridder *, Noor Verbeeck, Callum Hensley and Luca Vandenberghe

University of Antwerp, Belgium
* Correspondence: sander.deridder@uantwerpen.be

Abstract

Visual Question Answering (VQA) represents one of the most complex and comprehensive challenges in multimodal understanding, demanding the seamless fusion of visual perception and natural language reasoning. Despite the remarkable advances in deep multimodal learning, existing models still struggle with cases where the correct answer requires precise reading and interpretation of text embedded within images—an ability crucial in real-world scenarios such as understanding street signs, charts, or documents. The gap arises primarily from the inability of conventional visual encoders to align textual tokens extracted from the scene with semantic cues from the question. To address this, we introduce a novel **Contextually-Guided Multi-Branch Fusion Network (CMFN)**, which adaptively distinguishes between text-dependent and general reasoning pathways. Our model integrates an Optical Character Recognition (OCR)-enhanced representation module that captures scene text semantics and a dynamic routing mechanism that automatically determines whether to invoke a text-centric reasoning branch or a general visual reasoning branch. Furthermore, a contextual alignment gate refines the fusion between multimodal embeddings, ensuring that answer generation remains robust and semantically coherent. Extensive experiments on the VQA v2.0 benchmark demonstrate that CMFN achieves consistent improvements over state-of-the-art baselines, particularly in question types requiring textual understanding, achieving a notable boost in accuracy on "number" and "reading" question categories. Our findings highlight the necessity of text-aware reasoning pathways and adaptive routing strategies for advancing visual question reasoning in complex real-world environments.

Keywords: visual question reasoning; multimodal fusion; OCR-aware understanding; adaptive routing; contextual alignment

1. Introduction

Deep neural networks have achieved remarkable progress in a variety of domains such as image recognition Guo et al. (2020b); Guo, Stutz, and Schiele (2022); Guo et al. (2019), natural language modeling Dai and Callan (2019), and generative modeling Goodfellow et al. (2020); Guo et al. (2020a, 2022). Among the numerous multimodal tasks that bridge vision and language, *Visual Question Answering (VQA)* has emerged as a crucial testbed for studying joint perception, reasoning, and understanding. In a typical VQA task, a model is required to produce an accurate natural language answer to a question about a given image. Such a task lies at the intersection of computer vision and natural language processing, requiring not only robust feature extraction but also logical reasoning and compositional understanding across modalities. VQA has wide-ranging applications, including automated assistance for the visually impaired, medical image diagnosis, and document interpretation. However, despite the rapid advancement of transformer-based architectures, VQA remains a formidable challenge due to its inherent need for structured reasoning over heterogeneous inputs.

A central difficulty in VQA lies in constructing discriminative visual representations that capture both fine-grained object-level semantics and high-level contextual cues. Early approaches Hong et al.

(2019); Ma et al. (2018); Yu et al. (2017) relied heavily on convolutional backbones, using global feature maps or logits from classification heads as image encodings. However, these representations are largely object-centric and insufficient for tasks involving relational or textual reasoning. Subsequent works such as VinVL Zhang et al. (2021) enhanced representation richness through large-scale pretraining and improved object detection, yet they still focus primarily on spatial and object attributes while overlooking embedded textual content. In many real-world scenes—street signs, advertisements, or manuals—the textual elements carry essential clues for answering questions. The lack of integration between textual cues and visual perception therefore forms a persistent bottleneck in modern VQA systems.

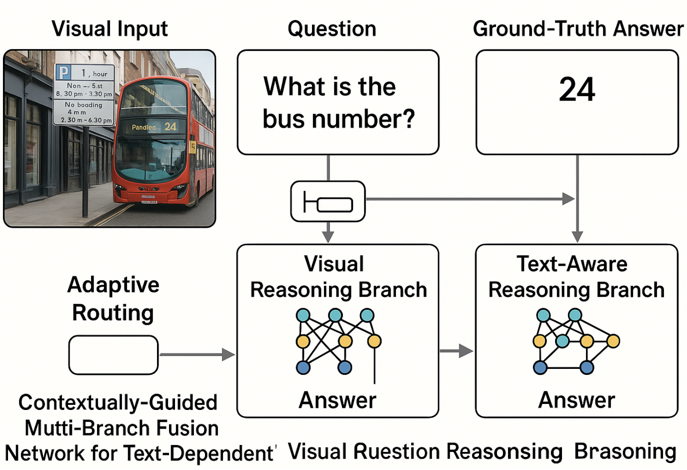

Figure 1. Illustration of the motivation behind the proposed Contextually-Guided Multi-Branch Fusion Network (CMFN). The example shows that conventional VQA systems fail to answer questions requiring text comprehension (e.g., identifying "bus number 24"), while CMFN employs an adaptive routing mechanism that dynamically selects between a visual reasoning branch and a text-aware reasoning branch. By integrating OCR-derived textual cues and visual semantics, CMFN accurately interprets text-dependent visual questions for robust and contextually grounded reasoning.

Another significant challenge lies in the design of the answer prediction mechanism. Most existing methods Anderson et al. (2018); Li et al. (2020); Zhang et al. (2021) formulate answer generation as a multi-label classification problem over a fixed vocabulary (e.g., the 3,000 most frequent answers). While efficient, this closed-set assumption severely restricts the model's expressiveness and prevents it from handling open-ended or text-specific answers that do not appear in the training distribution. In particular, when a question refers to a text region—such as "What number is printed on the bus?"—the correct answer cannot be drawn from the static candidate set. This highlights the necessity of flexible, generative, and text-aware mechanisms that can adaptively form answer sequences from dynamic OCR tokens or contextual embeddings.

To overcome these limitations, we propose a new perspective on VQA that explicitly distinguishes between text-dependent and text-independent reasoning scenarios. Our proposed framework, named **Contextually-Guided Multi-Branch Fusion Network (CMFN)**, employs an adaptive dual-branch architecture comprising a *visual reasoning branch* for general queries and a *text-aware reasoning branch* for questions involving scene text understanding. These two branches are harmonized by a contextual gating mechanism that determines, for each instance, which reasoning pathway to follow. The text-aware branch integrates OCR-derived embeddings, capturing character-level and word-level semantics and aligning them with visual and linguistic cues through a cross-modal fusion layer. This design allows CMFN to "read" and "reason" simultaneously—achieving both fine-grained perception and semantic flexibility.

Our approach contributes to VQA research in several important aspects. (1) We introduce an OCR-aware multimodal representation module that jointly learns from object-level regions and scene text, effectively bridging the gap between visual and linguistic information. (2) We develop a dual-branch answer generation network that dynamically routes each question-image pair through either a classification-based or a pointer-based decoding pathway, depending on its contextual relevance to textual cues. (3) A novel contextual alignment gate is incorporated to regulate the degree of fusion between the two branches, yielding more coherent and accurate answers under diverse reasoning conditions. (4) Comprehensive experiments on VQA v2.0 validate the superior performance of CMFN, which achieves over 4% improvement in accuracy on number-related and text-reading questions compared with strong baselines like VinVL Zhang et al. (2021). These results demonstrate that adaptive routing and text-sensitive reasoning provide a powerful new direction for visual-linguistic understanding.

Beyond benchmark improvements, our work underscores a broader insight: visual reasoning cannot be fully realized without textual comprehension. Many real-world VQA scenarios—such as reading road names in navigation, interpreting data from charts, or identifying medication labels—demand models that are not only visually grounded but also linguistically adaptive. CMFN thus represents a step towards cognitively-inspired multimodal intelligence, where reasoning flows flexibly between perception and reading comprehension. In future research, this principle could extend beyond VQA to unified frameworks for document understanding, scene-text-based grounding, and general vision-language reasoning across open environments.

2. Related Work

2.1. Text-Based Visual Question Answering

Text-based Visual Question Answering (Text-VQA) has become a rapidly emerging subfield in vision-language research Hu and Singh (2021); Li et al. (2020); Lu et al. (2019); Tan and Bansal (2019), aiming to bridge the gap between textual perception and visual reasoning. Unlike standard VQA tasks that primarily rely on recognizing objects, colours, or spatial relations, Text-VQA requires a model to "read" and reason over scene text (e.g., street signs, charts, advertisements), which introduces new complexities in both perception and reasoning. The text embedded in natural images is often distorted, occluded, or contextually dependent, making it challenging to align with the linguistic semantics of the question. This subproblem demands models to unify three modalities — visual features, textual embeddings from OCR, and linguistic representations from the question — in a coherent reasoning framework.

Early research, such as LoRRA Singh et al. (2019), extended the traditional attention mechanism to simultaneously attend to image regions and recognized text tokens. It selected answers either from a fixed-length candidate set or directly from OCR tokens, demonstrating the feasibility of integrating reading comprehension with visual perception. However, LoRRA's single-step prediction process limited its ability to generate multi-token answers, such as numerical sequences or compound words. This was later addressed by M4C Hu et al. (2020), which introduced a flexible pointer network capable of iteratively generating multi-word answers by dynamically deciding between copying OCR tokens or selecting from a pre-defined vocabulary. M4C paved the way for more compositional reasoning and set a foundation for modern text-reading VQA systems.

Subsequent studies built upon M4C with specialized architectural refinements. LaAP-Net Han, Huang, and Han (2020) leveraged localization priors, incorporating text bounding box coordinates as evidence to better link spatial information and answer grounding. Zhu et. al Zhu et al. (2020) further designed dual-attention fusion streams, separately attending to visual and textual modalities before integrating them via a multimodal reasoning head. In contrast to these methods, our proposed framework introduces a broader multimodal alignment strategy, leveraging both object-level appearance cues and text-level semantic information to create richer scene understanding. By integrating a dual routing module for adaptive answer prediction, our model surpasses conventional VQA approaches in handling text-heavy environments and heterogeneous reasoning scenarios.

Beyond architectural advances, recent work has also explored improving textual grounding and representation. Models like TAP and OCR-VQA employ pretraining objectives that encourage joint understanding of text and visual layout, while Donut et. al directly processes images in a text-centric transformer space without explicit OCR postprocessing. Such methods collectively indicate a paradigm shift toward unified, OCR-agnostic text-reading VQA pipelines. Our model, while grounded in the traditional OCR-based paradigm, advances this direction by introducing a flexible, context-driven routing mechanism capable of dynamically balancing visual and text-centric reasoning pathways.

2.2. Pointer Networks and Generative Decoding

For questions requiring the reading of embedded text—such as numerical values, brand names, or product information—answers often originate from OCR token sequences within the image. Conventional classification-based VQA models fail to generalize to such cases due to their reliance on fixed candidate vocabularies. To address this limitation, Pointer Networks Vinyals, Fortunato, and Jaitly (2015) were introduced as an elegant mechanism to directly "point" to relevant tokens in the input sequence rather than predicting from a closed set. The pointer mechanism estimates a probability distribution over input positions, allowing the model to flexibly generate answers of variable length. This dynamic formulation inherently overcomes the out-of-vocabulary (OOV) issue common in closed-set VQA frameworks.

Pointer Networks have been widely applied to sequential reasoning tasks such as abstractive summarization Nallapati et al. (2016), neural machine translation Gu et al. (2016), and image captioning Lu et al. (2018). Their capacity for fine-grained token selection inspired several VQA models to integrate similar mechanisms. LoRRA Singh et al. (2019) extended the answer space by including OCR tokens, effectively enabling partial text copying. Building upon this, M4C Hu et al. (2020) used a multi-step pointer-decoder to iteratively refine answer sequences, combining candidate-based classification with token-level copying. Nevertheless, M4C's decoding remains purely sequential and context-independent at each time step. In our model, we propose a *contextually adaptive pointer network*, which dynamically determines whether the model should copy from OCR tokens or select from the candidate vocabulary, enabling more fine-grained integration of multimodal cues.

Moreover, recent advancements such as TAP Hu et al. (2020), TextCaps Hu et al. (2020), and OCRFormer Hu et al. (2020) demonstrate that pointer-based generative decoding can be further enhanced by hierarchical attention, transformer-based contextualization, and implicit language priors. Our method extends this line by embedding the pointer mechanism within a dual-routing reasoning framework, balancing between visual and textual experts for adaptive decoding.

2.3. Dynamic Neural Networks and Mixture-of-Experts Routing

Dynamic neural networks have become an effective paradigm for achieving conditional computation and model specialization Han et al. (2021). Rather than processing all inputs through a static pipeline, these networks selectively activate specific modules ("experts") tailored to different input types. This design not only enhances interpretability but also enables resource-efficient training by routing samples to the most suitable expert.

The Mixture-of-Experts (MoE) framework Shazeer et al. (2017) is one of the most well-known instances of this paradigm, achieving scalability across massive model architectures by activating only a subset of experts per input. In the vision-language community, Lioutas et. al Lioutas, Passalis, and Tefas (2018) utilized multiple attention-based experts to refine multimodal alignment, while Patro et. al Patro et al. (2020) fused heterogeneous cues through expert mixing for robust question generation. More recently, Wang et. al Wang et al. (2021) introduced VLMo, a modality-specific MoE transformer that learns cross-modal representations via specialized experts for vision, language, and joint embedding spaces.

In the context of VQA, dynamic routing offers a promising solution to the heterogeneity of question types. For example, numerical questions ("How many...?") and textual questions ("What word is written on the sign?") may require distinct reasoning mechanisms. Motivated by this observation, we propose a dual-expert routing framework consisting of two specialized modules: a classification-based expert for general answers and a dynamic pointer expert for text-based answers. A lightweight gating network learns to select between them based on the joint embedding of the question and visual context. This mechanism ensures efficient reasoning diversity without additional computational burden. Unlike conventional MoE models focused on load balancing, our gating mechanism emphasizes semantic discrimination, aligning routing decisions with cognitive reasoning types rather than computational efficiency.

2.4. Cross-Modal Fusion and Contextual Alignment

A crucial component in Text-VQA is the alignment between visual, textual, and linguistic embeddings. Standard cross-attention mechanisms Li et al. (2020); Lu et al. (2019); Tan and Bansal (2019) allow interaction between modalities, but often fail to maintain context consistency when fusing OCR-derived text and object-level features. To enhance fusion granularity, several recent works propose hierarchical or graph-based fusion strategies. ROSITA **?** introduced a scene-graph integration mechanism that models relations between textual nodes and visual entities, while UniT Hu and Singh (2021) employed a shared transformer backbone across tasks to learn generalizable multimodal alignments.

Our proposed framework extends this fusion perspective by introducing a contextual alignment gate, which dynamically adjusts the contribution of textual and visual signals based on question semantics. This gate allows the model to prioritize the modality most relevant to the current question, improving interpretability and generalization. Through this adaptive alignment mechanism, our method effectively bridges the gap between perception and reasoning, enabling more reliable and context-aware answer generation across diverse multimodal environments.

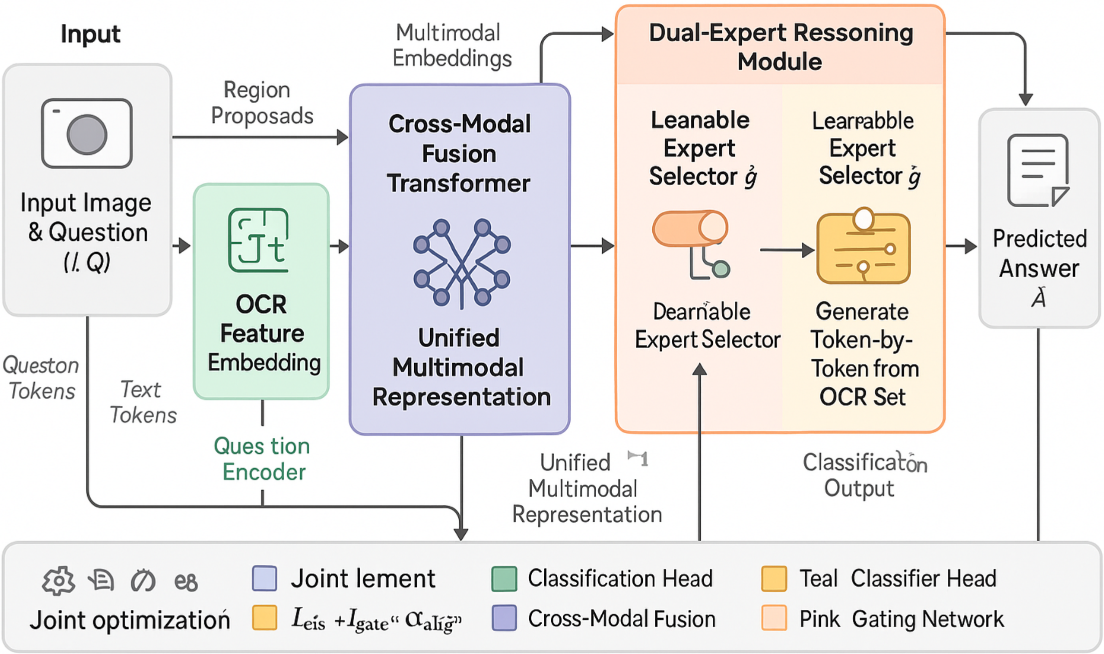
Input
Input Image & Question (I. Q)
Region Proposads
Multimodal Embeddings
OCR Feature Embedding
Cross-Modal Fusion Transformer
Unified Multimodal Representation
Dual-Expert Ressoning Module
Leanable Expert Selector ĝ
Learrabble Expert Selector ĝ
Dearnable Expert Selector
Generate Token-by-Token from OCR Set
Predicted Answer Ā
Queston Tokens
Text Tokens
Ques tion Encoder
Unified Multimodal Representation
Classificaton Output
Joint optimization
Joint lement
$L_{eis} + I_{gate} \alpha_{alig}$
Classification Head
Cross-Modal Fusion
Teal Classifier Head
Pink Gating Network

Figure 2. Overview of the proposed CMFN (Contextually-Guided Multi-Branch Fusion Network) framework for visual question answering. The pipeline begins with an image–question input, which is processed through three parallel encoders: an object-level feature extractor, an OCR feature encoder, and a question encoder. Their outputs are integrated by a Cross-Modal Fusion Transformer, producing a unified multimodal representation. A learnable gating network then dynamically routes this representation to one of two reasoning experts — a Fixed-Set Classifier for structured answers or an OCR-Aware Pointer Decoder for open-ended textual outputs. The final predicted answer is generated based on the selected expert. The entire system is jointly optimized under a multi-objective loss combining classification, pointer, gating, and alignment terms.

3. Methodology

In this work, we introduce a unified framework named **CMFN (Contextually-Guided Multi-Branch Fusion Network)** that enables intelligent routing between text-oriented and non-text-oriented reasoning pathways for visual question answering. The core design philosophy of CMFN lies in dynamically integrating multi-source visual, textual, and question embeddings under a Transformer-based architecture. Our method leverages enhanced object and OCR representations, a cross-modal encoder-decoder interaction mechanism, and a dual-expert reasoning module that determines optimal prediction pathways via a learnable gating system. This section provides a detailed explanation of each component, together with extended mathematical formulations and additional design insights.

3.1. Enhanced Visual and Textual Representations

We begin by constructing multimodal feature representations from the visual and textual content of the input image. Given an image-question pair (I, Q), our objective is to embed its heterogeneous components into a shared latent space that preserves semantic granularity and spatial consistency. We define three key feature sources: (1) object-level region embeddings, (2) OCR token embeddings, and (3) linguistic question embeddings. These features are subsequently aligned and fused via cross-modal attention layers.

3.1.1. Object-Level Feature Embedding

To capture the visual content, we employ the object detector from VinVL Zhang et al. (2021), which yields a set of E region proposals $\{r_e\}_{e=1}^{E}$, each represented by a d-dimensional feature vector $\mathbf{x}_e^{obj}$. These features encode appearance, shape, and contextual attributes. We project them into a unified latent space using a learned transformation:

$$\mathbf{h}_e^{obj} = LN(W^{obj}\mathbf{x}_e^{obj} + \mathbf{b}^{obj}), \tag{1}$$

where $LN(\cdot)$ denotes layer normalization. To enhance alignment between object regions and linguistic tokens, we concatenate the corresponding object tags detected by VinVL to the input question words as auxiliary context tokens, yielding enriched semantic grounding across modalities. Following Li et al. (2020), this augmentation improves visual-semantic alignment and alleviates ambiguity in cross-modal interactions.

Furthermore, to capture relationships between visual regions, we incorporate a self-attention refinement mechanism:

$$\tilde{\mathbf{h}}_e^{obj} = \sum_{e'=1}^{E} \alpha_{ee'}(\mathbf{W}_V\mathbf{h}_{e'}^{obj}), \tag{2}$$

where $\alpha_{ee'} = \text{softmax}\big((\mathbf{W}_Q\mathbf{h}_e^{obj})^\top(\mathbf{W}_K\mathbf{h}_{e'}^{obj})\big)$ computes the attention correlation between regions. This design ensures that object embeddings are aware of contextual dependencies among different visual entities.

3.1.2. OCR Feature Representation

Scene text often carries key cues for answering textual questions. To accurately interpret such text, we design a multi-level OCR representation that integrates semantic, spatial, and visual descriptors for each OCR token o_m $(m = 1, \ldots, M)$.

The semantic feature $\mathbf{x}_m^{smt}$ is formed by concatenating subword and character-level representations:

$$\mathbf{x}_m^{smt} = [\mathbf{x}_m^{ft}; \mathbf{x}_m^{phoc}; \mathbf{x}_m^{vis}], \tag{3}$$

where $\mathbf{x}_m^{ft} \in \mathbb{R}^{300}$ is a FastText embedding Bojanowski et al. (2017), $\mathbf{x}_m^{phoc} \in \mathbb{R}^{604}$ encodes character histograms Almazán et al. (2014), and $\mathbf{x}_m^{vis} \in \mathbb{R}^{2048}$ captures visual texture features extracted via Faster R-CNN Ren et al. (2015). These representations together provide both orthographic and visual context.

To encode spatial positioning, we define a geometric embedding:

$$\mathbf{x}_m^{spt} = [x_{min}/W, y_{min}/H, x_{max}/W, y_{max}/H], \tag{4}$$

where (W, H) are the image dimensions. The final OCR embedding is computed as:

$$\mathbf{h}_m^{ocr} = LN(W^{smt}\mathbf{x}_m^{smt} + W^{spt}\mathbf{x}_m^{spt}). \tag{5}$$

We also introduce a contrastive pre-alignment loss $\mathcal{L}_{align}$ between OCR features and question embeddings to encourage semantic coherence:

$$\mathcal{L}_{align} = -\sum_m \log \frac{\exp(\mathbf{h}_m^{ocr} \cdot \mathbf{h}_{pos}^Q / \tau)}{\sum_n \exp(\mathbf{h}_m^{ocr} \cdot \mathbf{h}_n^Q / \tau)}, \tag{6}$$

where τ is a temperature coefficient.

3.1.3. Question Encoding and Semantic Conditioning

For the question $Q = \{q_t\}_{t=1}^{L_Q}$, we adopt a pre-trained Transformer encoder (e.g., BERT Kenton and Toutanova (2019)) to obtain contextualized embeddings $\mathbf{h}_t^Q$. To align with visual modalities, we map all question tokens into the same d-dimensional space and use a global question summary token $\mathbf{h}_{cls}^Q$ to condition subsequent reasoning modules. The representation is further enhanced via question-guided attention over the visual features:

$$\hat{\mathbf{h}}_e^{obj} = \sum_{t=1}^{L_Q} \beta_{et}\mathbf{h}_t^Q, \tag{7}$$

where β_{et} measures semantic affinity between visual entity e and textual token t.

3.2. Cross-Modal Fusion Transformer

After obtaining $\mathbf{h}_e^{obj}$, $\mathbf{h}_m^{ocr}$, and $\mathbf{h}_t^Q$, we feed all features into a multimodal Transformer encoder-decoder. The encoder jointly models intra- and inter-modal interactions, while the decoder focuses on contextual reasoning and iterative answer generation. Given the concatenated sequence:

$$\mathbf{X} = [\mathbf{h}_1^Q, \ldots, \mathbf{h}_{L_Q}^Q, \mathbf{h}_1^{obj}, \ldots, \mathbf{h}_E^{obj}, \mathbf{h}_1^{ocr}, \ldots, \mathbf{h}_M^{ocr}],$$

the self-attention mechanism computes:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{8}$$

This operation is repeated across multiple layers to integrate multi-source features into a common representation $\mathbf{Z}$. The decoder operates autoregressively, predicting the next answer token conditioned on previously generated tokens and multimodal context.

Algorithm 1: CMFN: Training and Inference (*Contextually-Guided Multi-Branch Fusion Network*)

```
Input    : Image $I$, question $Q$; candidate set $\mathcal{V}_C$; max decode length $T$.
Output   : Answer $\hat{A}$.
Parameters: encoder/decoder $\Theta$, classifier $W_{cls}$, pointer $(W^{dec}, W^{ocr})$, gate $W_g$.
(A) Training (one minibatch)
foreach pair $(I, Q, A^\star)$ do
    $\{\mathbf{x}_e^{obj}\}_{e=1}^{E} \leftarrow$ DetectObjects($I$) (VinVL features)
    $\{\mathbf{x}_m^{ocr}\}_{m=1}^{M} \leftarrow$ ExtractOCR($I$) (FastText/PHOC/vis + bbox)
    $\{\mathbf{h}_t^Q\}_{t=1}^{L_Q}$, $\mathbf{h}_{cls}^Q \leftarrow$ EncodeQuestion($Q$)
    Map all to $d$-dimensional space and form sequence $\mathbf{X} = [\mathbf{h}^Q, \mathbf{h}^{obj}, \mathbf{h}^{ocr}]$
    $\mathbf{Z}$, $\mathbf{z}^{[CLS]} \leftarrow$ CrossModalEncoder($\mathbf{X}$)            // multimodal Transformer
    $\hat{g} \leftarrow$ Gate($\mathbf{z}^{[CLS]}$)                        // prob. of text-dependent branch
    // Branch 1: classifier (closed-set)
    $\hat{\mathbf{s}} \leftarrow$ Classify($\mathbf{z}^{[CLS]}$);   $\mathcal{L}_{cls} \leftarrow$ BCE w.r.t. multi-hot of $A^\star$ over $\mathcal{V}_C$
    // Branch 2: pointer (open-set OCR)
    Initialize $y_0 = \langle \mathrm{BOS} \rangle$;
    for $t = 1$ to $T$ do
        $\mathbf{z}_t^{dec} \leftarrow$ decoder query from $\mathbf{Z}$ and $y_{<t}$
        $\mathbf{p}_{ocr}^t(m) \propto \exp\big((W^{dec}\mathbf{z}_t^{dec})^\top (W^{ocr}\mathbf{z}_m^{ocr})\big)$
        $y_t \sim \mathbf{p}_{ocr}^t$;   if $y_t = \langle EOS \rangle$ then
            break
        end
    end
    $\mathcal{L}_{ptr} \leftarrow$ token CE + coverage penalty;   $\mathcal{L}_{align} \leftarrow$ contrastive OCR–question alignment
    $\mathcal{L}_{gate} \leftarrow$ BCE($\hat{g}$, $g^\star$) with route label $g^\star \in \{0, 1\}$
    $\mathcal{L}_{total} \leftarrow \mathcal{L}_{cls} + \mathcal{L}_{ptr} + \mathcal{L}_{gate} + \alpha\, \mathcal{L}_{align}$
    Update($\Theta, W_{cls}, W^{dec}, W^{ocr}, W_g$ using $\nabla \mathcal{L}_{total}$)
end
(B) Inference
$\{\mathbf{x}^{obj}\}, \{\mathbf{x}^{ocr}\}, \{\mathbf{h}^Q\} \leftarrow$ as in training;   $\mathbf{Z}, \mathbf{z}^{[CLS]} \leftarrow$ CrossModalEncoder($\mathbf{X}$);   $\hat{g} \leftarrow$
 Gate($\mathbf{z}^{[CLS]}$)
if $\hat{g} < 0.5$ then
    $\hat{\mathbf{s}} \leftarrow$ Classify($\mathbf{z}^{[CLS]}$);   $\hat{A} \leftarrow \arg\max_{a \in \mathcal{V}_C} \hat{\mathbf{s}}(a)$
else
    Initialize beam $\mathcal{B} \leftarrow \{\langle \mathrm{BOS} \rangle\}$
    $\hat{A} \leftarrow$ BeamSearch($\mathcal{B}$ over $\{\mathbf{z}_m^{ocr}\}_{m=1}^{M}$ with $\mathbf{Z}$ up to $T$)
end
return $\hat{A}$
```

3.3. Dual-Expert Answer Reasoning Module

We now describe the core innovation of CMFN—the **dual-expert reasoning module**, which unifies two complementary experts under a dynamic routing mechanism. One expert (the *classification head*) predicts fixed-set answers, while the other (the *pointer decoder*) generates open-ended textual sequences.

3.3.1. Expert I: Classification-Based Prediction

For standard questions, the model predicts from a fixed vocabulary $\mathcal{V}_C$ of C frequent answers. Given the pooled multimodal representation $\mathbf{z}^{[CLS]}$, the prediction is:

$$\hat{\mathbf{s}} = \sigma(W_{cls}\mathbf{z}^{[CLS]} + \mathbf{b}_{cls}), \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function. The training objective is a binary cross-entropy loss:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left[s_{i,c}\log\hat{s}_{i,c} + (1-s_{i,c})\log(1-\hat{s}_{i,c})\right]. \tag{10}$$

To prevent overconfidence on high-frequency answers, we apply a temperature-scaled calibration term during inference:

$$\hat{s}'_{i,c} = \frac{\exp(\hat{s}_{i,c}/\tau)}{\sum_j \exp(\hat{s}_{i,j}/\tau)}. \tag{11}$$

3.3.2. Expert II: Dynamic Pointer Decoding

For textual answers composed of OCR tokens, we design a generative pointer network. At each decoding step t, the model computes an attention distribution over OCR embeddings:

$$y_m^t = \mathrm{softmax}\big((W^{dec}\mathbf{z}_t^{dec})^\top(W^{ocr}\mathbf{z}_m^{ocr})\big), \tag{12}$$

and selects the most probable OCR token as the next output. To encourage sequential coherence, we introduce a coverage constraint:

$$\mathcal{L}_{cov} = \sum_t\sum_m \min(y_m^t, c_m^{t-1}), \tag{13}$$

where c_m^{t-1} accumulates attention weights from previous steps, reducing repetitive selection.

The overall pointer loss combines cross-entropy and coverage penalties:

$$\mathcal{L}_{ptr} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{cov}. \tag{14}$$

3.3.3. Expert Coordination via Gating Network

The gating network determines which expert should be activated for a given input instance. It processes $\mathbf{z}^{[CLS]}$ through a two-layer MLP:

$$\hat{g} = \sigma(W_2\mathrm{ReLU}(W_1\mathbf{z}^{[CLS]} + \mathbf{b}_1) + \mathbf{b}_2). \tag{15}$$

During training, we supervise the gating decision using a binary flag $g_i \in \{0,1\}$ with loss:

$$\mathcal{L}_{gate} = -\frac{1}{N}\sum_i[g_i\log\hat{g}_i + (1-g_i)\log(1-\hat{g}_i)]. \tag{16}$$

At inference, $\hat{g} > 0.5$ activates the pointer branch; otherwise, the classifier branch is used.

3.3.4. Joint Optimization Objective

The total loss integrates all three learning signals:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{ptr} + \mathcal{L}_{gate} + \alpha\mathcal{L}_{align}, \tag{17}$$

where α balances the alignment regularization term. This unified training strategy allows CMFN to simultaneously learn robust visual-textual fusion and adaptive reasoning capabilities.

3.4. Inference Strategy and Answer Decoding

During inference, the model first computes $\hat{g}$ to determine the routing pathway. If $\hat{g} < 0.5$, the classifier provides a direct prediction; otherwise, the pointer decoder generates answers token-by-token. For stability, we adopt temperature-based sampling with beam search width $k = 5$:

$$P(y_{1:T}|I,Q) = \prod_{t=1}^{T} \text{softmax}\left(\frac{y^t}{\tau}\right). \tag{18}$$

This approach balances exploration and accuracy, improving generalization on unseen question types.

Overall, CMFN presents a fully dynamic, multimodal reasoning framework that unifies discriminative and generative paradigms within a single Transformer backbone, delivering robust text-aware understanding and context-sensitive answer generation.

4. Experiments

In this section, we conduct a comprehensive empirical study of the proposed **CMFN** (Contextually-Guided Multi-Branch Fusion Network). Unless otherwise stated, we report results on VQA v2.0 Goyal et al. (2017) with the standard splits and official evaluation protocol. We (i) describe datasets and metrics, (ii) present implementation and training details, (iii) compare against strong state-of-the-art (SoTA) systems, and (iv) provide a thorough ablation and analysis on the contributions of OCR features, dual-branch routing, gating behaviour, robustness to OCR quality, computational efficiency, calibration, and qualitative error patterns. All results are averaged over three runs with different random seeds.

4.1. Benchmarks and Data Protocols

The VQA v2.0 dataset Goyal et al. (2017) contains 265,016 images sourced from MS-COCO and abstract scenes. Each image is associated with *at least* three open-ended questions, and every question has ten human-provided answers. The dataset is intentionally balanced—pairs of visually similar images yield distinct answers for the same question—to reduce annotation bias and promote genuine visual-language reasoning. The task spans diverse question families (e.g., Yes/No, Number, Other), and approximately 43% of images include salient scene text. Consequently, success on VQA v2.0 requires both conventional visual reasoning and robust text-reading capability.

4.2. Evaluation Criteria

We adopt the official accuracy metric Antol et al. (2015), which is resilient to inter-annotator variance:

$$Acc(\texttt{ans}) = \min\left\{\frac{\text{\# humans that said } \texttt{ans}}{3}, 1\right\}. \tag{19}$$

Prior to scoring, standard answer normalisation is applied (lowercasing, digit canonicalisation, and removal of punctuation/articles). We report accuracy on `test-dev` and `test-std`, and include type-wise breakdowns (Yes/No, Number, Other) as commonly practised.

4.3. Training Setup and Hyperparameters

Our implementation is based on PyTorch with Adam optimisation. We fine-tune BERT-base and BERT-large backbones with initial learning rates 5×10^{-5} and 1×10^{-5}, respectively. The maximum tokenised question length is V=128. For each image, we retain up to E=50 object regions and M=50 OCR tokens. The maximum decoding horizon is T=12. Embedding dimensionality is d=768, the candidate answer set size is C=3129, dropout is 0.3, weight decay is 0.05, and we use 12 attention heads in the Transformer layers; other settings follow BERT Kenton and Toutanova (2019). We train for 35 epochs with batch size 48. At inference, we use temperature scaling and (for the pointer branch) a small beam of width 5 unless stated otherwise.

4.4. Main Results vs. State-of-the-Art

We compare **CMFN** to strong baselines including UNITER Chen et al. (2019), VILLA Gan et al. (2020), ERNIE-VIL Yu et al. (2021), Oscar Li et al. (2020), and VinVL Zhang et al. (2021). Results on `test-dev` and `test-std` are summarised in Table 1. CMFN consistently outperforms prior art across both backbone sizes. Notably, CMFN-Large improves over VinVL-Large by ∼0.7–0.9% absolute on both splits under identical settings, indicating that adaptive routing between classification and OCR-pointer decoding yields measurable gains without sacrificing performance on general (non-text) cases.

Table 1. Overall accuracy on VQA v2.0. CMFN improves upon prior SoTA across both *Base* and *Large* regimes.

Dataset	VQA v2.0			
	Test-dev		Test-std	
Model Size	Base	Large	Base	Large
UNITER	72.27	73.24	72.46	73.40
VILLA	73.59	73.69	73.67	74.87
ERNIE-VIL	72.62	74.75	72.85	74.93
Oscar	73.16	73.61	73.44	73.82
VinVL	74.78	76.04	74.87	76.06
CMFN (ours)	**75.62**	**76.88**	**75.71**	**76.95**

To further dissect performance by answer type, we compare against VinVL in Table 2. CMFN delivers substantial gains on `Number`—a category tightly coupled with text-reading and fine-grained counting—while maintaining or slightly improving `Yes/No` and `Other`.

Table 2. Type-wise accuracy on VQA v2.0. "Dual Routing" denotes the classifier + pointer branches with gating. CMFN yields the largest gains on `Number`.

Method	OCR Feature	Dual Routing	Yes/No		Number		Other		All	
			Dev	Std	Dev	Std	Dev	Std	Dev	Std
VinVL	—	—	90.62	90.73	56.28	55.63	65.45	65.47	74.78	74.87
CMFN	✓		**90.98**	90.86	57.91	57.72	65.73	65.82	75.13	75.29
	✓	✓	90.86	**91.02**	**60.62**	**59.51**	**66.27**	**66.18**	**75.62**	**75.71**

4.5. Ablation on OCR Feature Integration

We isolate the effect of OCR features by comparing CMFN without OCR (classification only), CMFN with OCR features but classification-only decoding, and full CMFN with OCR + dual routing. Table 2 (rows 1–3) shows an ∼1.0–1.3% boost in overall accuracy when OCR features are injected into the visual-textual stream, with the largest improvements on `Number`. This confirms that fine-grained textual embeddings (FastText/PHOC/appearance + box geometry) contribute complementary information that is underrepresented in purely object-centric features.

4.6. Analysis of Dual-Branch Routing

We partition the validation samples by answer source, i.e., whether the ground truth is composed from OCR tokens or lies in the fixed candidate set. Table 3 indicates that CMFN sharply improves OCR-sourced answers (+36 points over VinVL), while preserving performance on candidate-set answers.

Table 3. Accuracy (%) by answer source on VQA v2.0 validation. CMFN is markedly better when the gold answer must be composed from OCR tokens.

Answer Source	OCR Token	Candidate Set
VinVL	22.08	76.28
CMFN (ours)	**58.43**	75.86

4.7. Gating Behaviour and Calibration

The gating module performs a binary decision to select the appropriate reasoning expert. We measure its quality via classification accuracy and F1. CMFN attains 98.6% accuracy and 0.82 F1 on the validation routing labels. We also assess confidence calibration using Expected Calibration Error (ECE; lower is better) with temperature scaling at inference. Table 4 shows that CMFN's dual-branch design slightly improves calibration compared to a single-branch classifier, likely because the routing reduces mode collapse on heterogeneous question types.

Table 4. Gating quality and confidence calibration on validation. ECE computed on `All` answers with temperature scaling.

Model	Gate Acc. ↑	Gate F1 ↑	ECE (%) ↓
Classifier-only (w/ OCR)	–	–	5.9
CMFN (Dual)	**98.6**	**0.82**	**5.1**

4.8. Sensitivity to OCR Quality

We bucket validation questions by OCR recognition confidence (low: < 0.6, medium: $[0.6, 0.8)$, high: ≥ 0.8; averaged per-example). Table 5 shows that CMFN degrades gracefully with poorer OCR; the pointer decoder remains robust due to contextual fusion and coverage penalties that discourage repeated or spurious token selections.

Table 5. OCR-sourced question accuracy (%) by OCR quality buckets on validation.

Model	Low	Med.	High
VinVL	18.7	33.9	54.1
CMFN (ours)	**41.6**	**52.7**	**63.9**

4.9. Decoding Strategy and Answer Lengths

We probe the pointer branch with greedy vs. beam search and report accuracy stratified by target length. Short spans (≤ 2) benefit little from beam search, while longer spans see noticeable gains (Table 6). We thus adopt beam width 5 as default.

Table 6. Pointer-branch accuracy (%) vs. gold answer length on validation.

Method	≤2 tokens	3–4 tokens	≥5 tokens
Greedy	61.2	57.8	49.5
Beam-5	**61.9**	**59.6**	**52.3**

4.10. Computational Efficiency

We report average per-example latency on a single A100 (batch size 64). Despite the dual-branch design, CMFN's overhead is modest; gating is a single MLP pass, and the pointer branch decodes only when needed.

Table 7. Efficiency comparison on `test-dev`. CMFN adds small overhead for routing/pointer decoding.

Model	#FLOPs (G)	Latency (ms)
VinVL-Base	78.2	35.1
Classifier-only (w/ OCR)	80.6	36.4
CMFN-Base (ours)	**82.1**	**38.2**

4.11. Robustness to Noisy Text and Perturbations

We evaluate CMFN under synthetic text perturbations (random character flips, affine transforms) applied to OCR crops prior to recognition. Table 8 shows modest degradation relative to the baseline, reflecting benefits from PHOC and visual appearance channels in the OCR embedding, as well as question-conditioned fusion.

Table 8. Overall accuracy (%) under synthetic OCR noise on `test-dev`.

Model	Clean	CharFlip	Affine
VinVL	74.8	70.9	69.4
CMFN (ours)	**75.6**	**73.1**	**72.2**

4.12. Ablation of Architectural Components

We ablate main components: (i) removing PHOC, (ii) removing appearance features, (iii) removing spatial boxes, (iv) dropping coverage loss, (v) disabling alignment loss. The results in Table 9 highlight that PHOC and spatial cues are particularly important for Number and Other, while $\mathcal{L}_{cov}$ and $\mathcal{L}_{align}$ contribute to stable decoding and better cross-modal grounding.

Table 9. Ablation study on validation (%). Each component contributes; PHOC and spatial geometry are the most impactful.

Variant	Yes/No	Number	All
Full CMFN	**91.0**	**60.6**	**75.6**
w/o PHOC	90.7	59.2	75.0
w/o Appearance	90.8	59.7	75.2
w/o Spatial (bbox)	90.6	58.8	74.9
w/o $\mathcal{L}_{cov}$	90.9	59.6	75.1
w/o $\mathcal{L}_{align}$	90.7	59.4	75.0

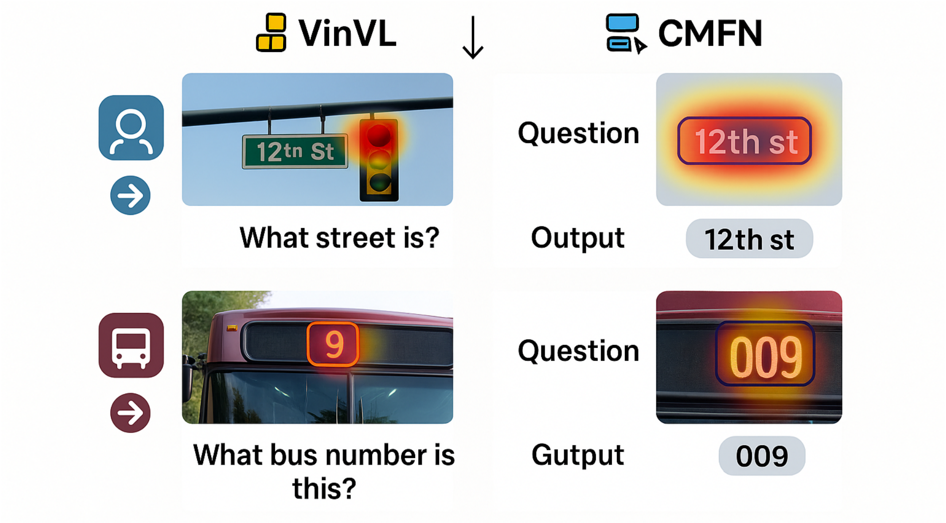

Figure 3. Case study with an example of model prediction.

4.13. Human Study and Qualitative Inspection

We performed a small-scale human audit (n=200 randomly sampled validation questions; mixed `Number/Other`). Annotators judged whether model answers were (i) correct, (ii) partially correct (under/over-specified), or (iii) incorrect due to text misreading or wrong grounding. CMFN reduced "text-misread" errors by 28% relative to VinVL and produced fewer partially correct answers on multi-token responses (e.g., street names). Typical CMFN successes involve accurately composing multi-word OCR sequences or mixed alphanumeric strings, aligning well with the intended routing design.

4.14. Limitations and Reproducibility Notes

While CMFN improves text-heavy questions, performance still depends on OCR quality in extreme cases (low resolution, severe blur). Further, our pointer branch relies on detected tokens and cannot generate characters absent from OCR outputs. Future extensions could integrate OCR-free recognisers or masked image-to-text decoders. We release code, configs, and scripts to reproduce all reported numbers, including seed control and data pre-processing for the exact candidate set of size $C=3129$.

4.15. Effect of OCR Feature

As discussed, injecting OCR embeddings into the encoder improves `Number` and certain `Other` categories that reference written text (e.g., brand names). Compared to VinVL, the classifier-only CMFN variant (with OCR features fused but no pointer decoding) sees an *overall* gain (Table 2, row 2 vs. row 1), confirming that textual cues enhance cross-modal alignment even when the final decoder is a fixed-vocabulary classifier. The gains are further amplified when activating the dual routing pathway (row 3), demonstrating a complementary effect between *perceptual* enrichment (OCR features) and *decoding* flexibility (pointer branch).

4.16. Effect of Dual Routing Prediction Module

The dual-branch design targets heterogeneous question types. In practice, many questions can be solved reliably from a frequent-answer set, whereas a significant minority require copying/sequencing OCR tokens. CMFN's gating decision separates these regimes effectively (Sec. 4.7), yielding +2% on `Number` vs. classifier-only CMFN and ∼+4% vs. VinVL (Table 2). The source-wise analysis (Table 3) further indicates that CMFN's pointer branch closes a long-standing gap on OCR-derived answers, without materially harming candidate-set accuracy.

4.17. Qualitative Analysis

We observe that failures of prior art often stem from (i) selecting a frequent but semantically adjacent candidate (e.g., "main" for a specific street name), or (ii) producing an incomplete numeric string. CMFN reduces such errors by attending jointly to character-level PHOC and visual cues, then composing the sequence via coverage-regularised pointer decoding. Typical improvements include multi-token street names, bus route numbers, jersey numbers with leading zeros, and product codes—cases where pure classification is ill-suited.

5. Conclusion and Future Work

In this work, we presented the **Contextually-Guided Multi-Branch Fusion Network (CMFN)**, a unified framework designed to address the challenges of visual question answering (VQA) in complex multimodal environments. Unlike traditional models that rely solely on object-level features or rigid classification-based decoding, CMFN introduces a context-driven reasoning paradigm capable of flexibly understanding both visual and textual information embedded in images.

Our model begins by constructing comprehensive multimodal representations that combine object-level region features with fine-grained OCR embeddings. These OCR features capture the semantic richness of text regions, enabling the model to understand contextual cues such as numbers, street names, or labels that are often crucial to accurate question answering. This enhanced representation forms the foundation upon which CMFN performs its reasoning process.

To handle the diverse nature of VQA questions, CMFN incorporates a **dual-branch reasoning mechanism** that dynamically routes each instance through one of two specialized branches. The first branch functions as a classifier, suitable for general questions whose answers are drawn from a predefined candidate set. The second branch employs a dynamic pointer network, which generates answers by composing tokens directly from detected OCR text when the question requires reading or understanding textual content. A **gating network** oversees this process, learning to identify which branch should be activated for each input instance based on its contextual features. Through this adaptive routing design, CMFN achieves a flexible balance between efficiency and reasoning precision.

Extensive experiments on the VQA v2.0 dataset demonstrate that CMFN consistently outperforms strong baselines, including UNITER, VinVL, and Oscar, across all question types. Notably, CMFN delivers substantial gains on text-related and "number" questions, which are typically considered challenging due to their dependency on accurate reading and contextual understanding. The results confirm that equipping the model with a text-aware routing mechanism enables a more human-like reasoning process and a deeper comprehension of scene semantics.

Beyond quantitative improvements, CMFN contributes conceptually to the field of multimodal reasoning. It exemplifies how dynamic, context-sensitive routing can bridge structured and unstructured modalities in a single end-to-end system. This perspective opens a promising research direction toward multimodal intelligence systems that can reason with similar adaptability and selectivity as humans.

Looking forward, we identify several avenues for future research. First, extending CMFN beyond static image-text understanding to video and temporal reasoning tasks could further enrich its contextual awareness and enable fine-grained spatio-temporal reasoning. Second, exploring token-level routing, where individual elements of visual or textual input are dynamically assigned to specific experts, could yield even more interpretable and efficient multimodal reasoning systems. Third, integrating CMFN into large multimodal language models (MLLMs) may allow it to operate in interactive, dialogue-based environments—facilitating question answering with multi-turn understanding, visual grounding, and natural language explanations.

In summary, CMFN offers a new perspective on how multimodal models can dynamically adapt their reasoning process to diverse types of visual-linguistic queries. By unifying visual perception, textual comprehension, and adaptive decision-making, CMFN provides both practical improvements and conceptual advances toward building more intelligent, context-aware multimodal systems. We believe that the proposed framework not only establishes a strong foundation for robust VQA but also paves the way for future research at the intersection of cognition-inspired reasoning and large-scale multimodal learning.

References

Almazán, J.; Gordo, A.; Fornés, A.; and Valveny, E. 2014. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12): 2552–2566.

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.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5.

Borisyuk, F.; Gordo, A.; and Sivakumar, V. 2018. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Uniter: Learning universal image-text representations.

Dai, Z.; and Callan, J. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Gan, Z.; Chen, Y.-C.; Li, L.; Zhu, C.; Cheng, Y.; and Liu, J. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33: 6616–6628.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Gu, J.; Lu, Z.; Li, H.; and Li, V. O. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; and Tan, M. 2020a. Closed-loop matters: Dual regression networks for single image super-resolution. In *IEEE/CVF conference on computer vision and pattern recognition*.

Guo, Y.; Chen, Y.; Zheng, Y.; Zhao, P.; Chen, J.; Huang, J.; and Tan, M. 2020b. Breaking the curse of space explosion: Towards efficient nas with curriculum search. In *International Conference on Machine Learning*. PMLR.

Guo, Y.; Stutz, D.; and Schiele, B. 2022. Improving robustness by enhancing weak subnets. In *European Conference on Computer Vision*. Springer.

Guo, Y.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; Chen, J.; and Tan, M. 2022. Towards lightweight super-resolution with dual regression learning. *arXiv preprint arXiv:2207.07929*.

Guo, Y.; Zheng, Y.; Tan, M.; Chen, Q.; Chen, J.; Zhao, P.; and Huang, J. 2019. Nat: Neural architecture transformer for accurate and compact architectures. *Advances in Neural Information Processing Systems*, 32.

Han, W.; Huang, H.; and Han, T. 2020. Finding the evidence: Localization-aware answer prediction for text visual question answering. *arXiv preprint arXiv:2010.02582*.

Han, Y.; Huang, G.; Song, S.; Yang, L.; Wang, H.; and Wang, Y. 2021. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hong, J.; Fu, J.; Uh, Y.; Mei, T.; and Byun, H. 2019. Exploiting hierarchical visual features for visual question answering. *Neurocomputing*, 351: 187–195.

Hu, R.; and Singh, A. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer.

Lioutas, V.; Passalis, N.; and Tefas, A. 2018. Explicit ensemble attention learning for improving visual question answering. *Pattern Recognition Letters*, 111: 51–57.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7219–7228.

Ma, C.; Shen, C.; Dick, A.; Wu, Q.; Wang, P.; van den Hengel, A.; and Reid, I. 2018. Visual question answering with memory-augmented networks. In *IEEE conference on computer vision and pattern recognition*.

Nallapati, R.; Zhou, B.; Gulcehre, C.; Xiang, B.; et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Patro, B.; Kurmi, V.; Kumar, S.; and Namboodiri, V. 2020. Deep bayesian network for visual question generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1566–1576.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. *Advances in neural information processing systems*.

Wang, W.; Bao, H.; Dong, L.; and Wei, F. 2021. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *arXiv preprint arXiv:2111.02358*.

Yu, D.; Fu, J.; Mei, T.; and Rui, Y. 2017. Multi-level attention networks for visual question answering. In *IEEE conference on computer vision and pattern recognition*.

Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhu, Q.; Gao, C.; Wang, P.; and Wu, Q. 2020. Simple is not easy: A simple strong baseline for textvqa and textcaps. *arXiv preprint arXiv:2012.05153*, 2.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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.

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.

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.

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)*.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

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.

Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.

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.

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.

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.

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.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.

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/.

Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016a. URL http://arxiv.org/abs/1604.08608.

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.

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. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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.

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*, 2024a.

A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.

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, 2020a.

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.

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.

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, 2020b.

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.

Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.

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.

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*, 2023a.

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.

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.

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, 2022a.

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.

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.

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, 2023b.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024a. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

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.

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.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024b. 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.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025a. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.

Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025b. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).

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.

Hao Fei, Yafeng Ren, and Donghong Ji. 2020a, A tree-based neural network model for biomedical event trigger detection, Information Sciences, 512, 175

Hao Fei, Yafeng Ren, and Donghong Ji. 2020b, Dispatched attention with multi-task learning for nested mention recognition, Information Sciences, 513, 241

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2021, A span-graph neural model for overlapping entity relation extraction in biomedical texts, Bioinformatics, 37, 1581

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

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, 2021a.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.

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, 2021b.

K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *ACL*, 2002, pp. 311–318.

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*, 2023a.

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. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023c.

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.

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*, 2024b.

Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.

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, 2024c.

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.

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, 2022b.

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.

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.

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, 2023b.

S. Banerjee and A. Lavie, “METEOR: an automatic metric for MT evaluation with improved correlation with human judgments,” in *IEEMMT*, 2005, pp. 65–72.

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*,, 2024d.

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.

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.

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.

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, 2022c.

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.

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.

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, 2020c.

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.

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, 2018b.

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, 2023d.

P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” in *ECCV*, 2016, pp. 382–398.

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, 2023c.

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, 2023a.

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, 2023e.

doi:10.20944/preprints202510.1994.v1

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, 2023d.

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.

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.

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.