Cognitive Grounding for Visual Question Reasoning via Dynamic Knowledge Imagination

Madison Clarke *, Ryan Dhaliwal, Brielle Monroe, Isabelle Chen

Posted Date: 27 October 2025

doi: 10.20944/preprints202510.1967.v1

Keywords: visual question answering; commonsense reasoning; multimodal understanding; knowledge imagination; cognitive grounding

 doi:10.20944/preprints202510.1967.v1

Article

Cognitive Grounding for Visual Question Reasoning via Dynamic Knowledge Imagination

Madison Clarke *, Ryan Dhaliwal, Brielle Monroe and Isabelle Chen

Simon Fraser University
* Correspondence: madison.clarke@sfu.ca

Abstract

Visual question answering (VQA) represents a critical intersection of vision and language understanding, where models must perceive visual scenes and reason about their underlying semantics. However, human-like reasoning often extends beyond what is directly observable—requiring the invocation of prior knowledge, inference, and commonsense understanding. In this work, we reexamine the nature of external knowledge in multimodal reasoning and propose a unified framework named **Cognitive Grounded Imagination Network (COGINet)** that integrates dynamically generated commonsense imagination with vision-language understanding. Instead of merely retrieving symbolic facts from static knowledge bases, our framework leverages contextualized commonsense imagination synthesized through large-scale generative knowledge models, thereby grounding dynamic reasoning on both perceptual evidence and inferred world regularities. COGINet introduces a two-stage process: (1) a knowledge imagination module, which generates plausible contextual hypotheses from the interaction between visual regions and textual queries; and (2) a cross-modal reasoning transformer that fuses these contextualized inferences with multimodal embeddings through adaptive attention. We demonstrate that such cognitive grounding enables the model to reason about abstract or implied concepts, extending beyond explicit cues in the image or text. Extensive experiments conducted on OK-VQA and A-OKVQA benchmarks show that COGINet achieves consistent improvements over prior static-knowledge methods, providing both quantitative gains and qualitative interpretability. Further analysis reveals the model's ability to discern when external knowledge is useful, selectively invoking imagined context only when required for reasoning. Our findings highlight the importance of dynamic, cognitively inspired commonsense integration for achieving genuine multimodal understanding.

Keywords: visual question answering; commonsense reasoning; multimodal understanding; knowledge imagination; cognitive grounding

---

1. Introduction

The task of Visual Question Answering (VQA) [1,6,12,37,49] has emerged as a central challenge for multimodal learning, serving as a benchmark to assess how well an AI model can bridge visual perception and linguistic reasoning. In its standard form, a model is presented with an image and a natural-language question, and is expected to produce a correct answer grounded in the image's content. This process implicitly involves several reasoning capabilities: localizing relevant regions, parsing semantic relations, interpreting textual cues, and aligning cross-modal representations. Over the past few years, the success of large-scale pre-trained transformers [8,21,44] has elevated the performance of such models to impressive levels on datasets like VQA v2, closing much of the gap with human performance for perceptual-level questions.

Nevertheless, a key limitation persists: the majority of existing VQA systems remain bound by what is visually observable. Human reasoning, by contrast, often goes beyond visual cues — it infers intent, causality, and social or physical context from partial information. For instance, given an image of a dinner plate with steak and bread, a human naturally infers it is likely a restaurant meal or part of

 doi:10.20944/preprints202510.1967.v1

a social gathering. Such inferences rely on commonsense knowledge accumulated through everyday experience. Standard vision-language pre-training pipelines only capture statistical co-occurrence patterns from web-scale corpora, which often lack sufficient coverage or contextual richness to support these cognitive inferences [32]. As a result, models trained purely on perceptual data struggle with questions requiring inference, analogy, or background understanding.

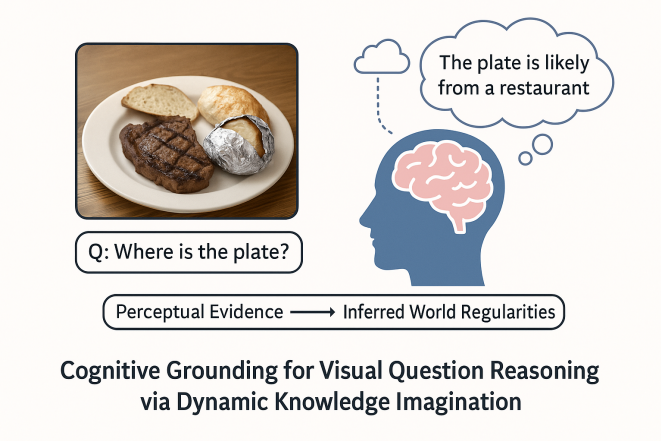

Figure 1. Illustration of cognitive grounding in Visual Question Answering. The model observes a visual scene (a plate with steak, potato, and bread) and, through dynamic knowledge imagination, infers contextual meaning — reasoning that the meal is likely served in a restaurant. This exemplifies the transition from perceptual evidence to inferred world regularities, reflecting how COGINet integrates commonsense imagination into multimodal understanding.

To address this limitation, researchers have explored the integration of external knowledge into VQA systems, giving rise to knowledge-based VQA (KB-VQA) frameworks [29,36,45,49]. These approaches augment the reasoning process with facts drawn from structured knowledge bases (e.g., ConceptNet, ATOMIC), or textual sources such as Wikipedia. Despite notable progress, such methods remain constrained by static retrieval mechanisms: knowledge bases are inherently incomplete, contain redundant or context-irrelevant facts, and often fail to adapt to the specific nuances of a given image-question pair. Moreover, retrieval-based designs treat knowledge as explicit text snippets, which must then be manually aligned with the multimodal representation, leading to rigid and brittle reasoning pipelines [9,28,47].

Motivated by these challenges, we introduce a new paradigm of *knowledge imagination*—a process that dynamically generates contextualized commonsense inferences conditioned on both visual and linguistic cues. This idea draws inspiration from cognitive science, where imagination acts as a generative simulation mechanism that enables humans to hypothesize unseen causes and consequences. Instead of retrieving fixed facts, our proposed model—**COGINet**—synthesizes candidate commonsense inferences using a generative commonsense model, such as COMET [2,15], guided by the semantic structure of the question and salient visual entities. These inferences are subsequently filtered, ranked, and embedded into the model's multimodal reasoning pipeline via a contextual attention fusion mechanism, allowing the system to focus on the most relevant cognitive signals.

More concretely, COGINet builds upon the architecture of vision-language transformers (such as VL-BERT [41]), but enhances them with a *Cognitive Imagination Layer* that encodes generated commonsense hypotheses alongside visual-textual features. To mitigate the potential risk of noisy or irrelevant knowledge injection, we employ a weakly supervised gating strategy that learns to modulate knowledge flow based on the nature of the question. For instance, for purely visual questions (e.g., "What color is the umbrella?"), the gate suppresses external imagination; whereas for conceptual queries (e.g., "Why is the child wearing a helmet?"), the system dynamically activates the generative inference process.

 doi:10.20944/preprints202510.1967.v1

Through extensive evaluation on OK-VQA [29] and A-OKVQA [36], we demonstrate that COGINet significantly improves performance on questions requiring inferential reasoning. Our analysis reveals that dynamically imagined knowledge allows the model to reason over implicit causes, human intentions, and object affordances that static retrieval systems often overlook. Furthermore, interpretability studies show how the attention heads in our cognitive fusion layer correlate with semantically meaningful patterns, such as "safety," "social setting," or "food preparation," highlighting the interpretive transparency of the system.

In summary, this work contributes threefold: (1) we introduce a cognitively inspired notion of knowledge imagination for multimodal reasoning; (2) we develop an end-to-end transformer framework that dynamically generates, filters, and integrates contextual commonsense hypotheses; and (3) we provide empirical and interpretive evidence that such imagination-driven grounding leads to more robust and human-like visual question reasoning. We believe this direction not only extends the frontier of VQA research but also moves a step closer to cognitively aligned multimodal intelligence.

2. Related Work

2.1. Foundation Models for Vision–Language Understanding

Pre-trained vision–language transformers rooted in BERT-style architectures [8] have become the dominant backbone for multimodal reasoning, including Visual Question Answering (VQA). Early two-stream designs such as ViLBERT [25] and LXMERT [42] encode text and image features separately and then align them through cross-modality attention, enabling fine-grained token–region interactions. In contrast, single-stream models—VL-BERT [41], OSCAR [22], and OSCAR+ [50]—jointly contextualize visual regions and subword tokens in a unified transformer, often yielding stronger fusion and simpler training dynamics. These models are typically pre-trained on web-scale corpora such as Conceptual Captions [38] with objectives that explicitly promote cross-modal alignment, e.g., masked language modeling over text conditioned on image regions, masked region modeling conditioned on text, and multimodal contrastive objectives [22,25,41,42]. Such pre-training pipelines inadvertently absorb a slice of commonsense regularities present in the data distribution, which partially explains their strong downstream performance on VQA [1].

Despite these advances, a persistent gap remains when questions demand knowledge that is not directly visible or requires multi-hop reasoning over latent context. Purely perceptual cues or distributional statistics learned during pre-training may be insufficient for queries about intentions, affordances, or socio-cultural conventions—phenomena that are sparse or implicit in web-scale text and thus underrepresented in model weights [32]. Consequently, even state-of-the-art models struggle on knowledge-intensive settings. In this paper we revisit this limitation and advocate explicit, context-aware knowledge infusion. We denote our approach as **AURORAVL** (hereafter our model), a VL-BERT–based framework that integrates external commonsense signals to mitigate the aforementioned deficiency and to better support questions requiring reasoning that extends beyond what is depicted.

2.2. Knowledge-Enhanced Visual Question Answering

To stress-test models' capacity for reasoning with background knowledge, numerous datasets target facts or commonsense beyond the pixels: FVQA [45] and WebQA [5] emphasize factual and web evidence; VLQA [34] supplies a passage for retrieval-style comprehension; VCR [49] focuses on commonsense-driven rationale selection; and OK-VQA [29] as well as A-OKVQA [36] explicitly require external knowledge to answer open-ended questions. These benchmarks have catalyzed a family of knowledge-enhanced VL transformers that complement visual–text features with retrieved facts. A common strategy is to extract subgraphs from ConceptNet [40] or other KBs and inject them as additional inputs or graph embeddings [9,20,28,47]. Other works leverage captions [33], web search snippets [26], Wikipedia pages, or even Google Images to assemble a richer context [47].

 doi:10.20944/preprints202510.1967.v1

An alternative paradigm uses large language models (LLMs) as few-shot reasoners. PICa [48] and KAT [13] demonstrate that prompting GPT-3 [3] with object tags and captions can supply answers without explicit parameter updates. While effective, such pipelines may incur non-trivial cost at inference time and can be brittle under prompt shifts. By contrast, our model focuses on the commonsense-rich subset of these tasks and employs COMET [15] to synthesize context-sensitive inferences. The resulting integration is simpler to maintain, cost-efficient, and tailored to the types of relational and causal knowledge most frequently needed in OK-VQA and A-OKVQA.

2.3. Incorporating Commonsense in NLP

Large knowledge bases (KBs) such as ConceptNet [40] and ATOMIC [35] have long supported NLP systems with symbolic assertions: ConceptNet contributes concept- and entity-level relations (e.g., *IsA*, *PartOf*, *MadeOf*), while ATOMIC aggregates event-centric social commonsense about causes, effects, and mental states at scale. Numerous methods inject such knowledge by encoding retrieved subgraphs, pre-training on commonsense objectives, or adding auxiliary tasks [9,23,51]. Yet direct retrieval from KBs suffers from incompleteness and weak contextualization: an otherwise correct fact may be irrelevant under a particular image–question pairing.

COMET reframes this bottleneck by fine-tuning pre-trained LMs on KB triplets to *generate* plausible inferences conditioned on free-form textual inputs [15]. This generative view addresses coverage (by extrapolating beyond enumerated triples) and context (by conditioning on the specific query), and has proven effective in multiple text-centric tasks [4,27,39,43]. Building on this success, we employ COMET to produce candidate hypotheses that our model filters and fuses with multimodal evidence. Notably, variants like VisualCOMET and symbolic distillation approaches [30,46] target event-level reasoning that is less aligned with OK-VQA and A-OKVQA, which more often revolve around entity attributes, affordances, and functional context; hence our focus on COMET's commonsense relations better suits the problem setting.

2.4. Retrieval vs. Generation for External Knowledge

Knowledge retrieval pipelines query a fixed repository (KBs, Wikipedia, web) and must then filter, rank, and align results with the multimodal context. While retrieval offers verifiability and explicit provenance, it struggles with recall in long-tail scenarios and can introduce topic drift when queries are short or noisy. Generative knowledge models like COMET [15] trade provenance for adaptability, producing context-tailored inferences that capture tacit regularities (e.g., social scripts, affordances). Our approach combines their strengths: we *generate* candidate commonsense, apply learned selection to improve precision, and *fuse* the result with vision–language features so that spurious or redundant hypotheses are attenuated by the downstream attention. This hybridization reduces the brittleness endemic to pure retrieval and the hallucination risk in unconstrained generation.

LLMs prompted with object tags, scene descriptions, or chain-of-thought rationales (e.g., [3,13,48]) can act as powerful zero-/few-shot reasoners. However, prompt-based strategies are sensitive to formatting choices, may fail to ground answers in the image when text dominates, and can incur significant compute or latency. Moreover, reporting bias [11] and distributional artifacts in the pretraining corpus can skew responses toward sensational or atypical events. Our model mitigates these pitfalls by limiting the role of generation to *commonsense prior proposals* and learning a gating mechanism that modulates their influence based on the question type and visual evidence strength.

Contrastive learning, masked modeling, and region–text alignment have become standard objectives for VL pre-training [22,25,41,42]. These objectives teach models to correlate tokens and regions but not necessarily to capture causal, teleological, or normative aspects of everyday knowledge (e.g., *why* a helmet is worn). We therefore introduce explicit *knowledge-centric* inductive biases at fine-tuning time: dynamically generated commonsense candidates are represented in the same embedding space as visual–text features and are fused through attention, encouraging the network to internalize when and how such signals improve reasoning without overfitting to textual shortcuts.

 doi:10.20944/preprints202510.1967.v1

2.5. Knowledge Selection, Calibration, and Gating

A critical challenge in knowledge augmentation is *selection*: over-supplying facts can drown the useful signal, whereas under-supplying leads to brittle failures. Prior works typically rely on heuristics or separate retrievers [9,28,47]. We instead treat selection as a learned component: sentence-level encoders [31] score COMET outputs for semantic pertinence to the image–question pair, and a weakly supervised gate estimates the *utility* of external knowledge conditioned on question type (e.g., descriptive vs. explanatory). This addresses both precision (by filtering) and calibration (by learning when to abstain).

Multimodal systems risk hallucinating unsupported content when textual priors dominate visual grounding. While some works add attribution constraints or rationale supervision, they often require expensive annotations. Our design promotes faithfulness by (i) using visual entities and tags to condition knowledge generation, (ii) fusing knowledge through attention that competes with image features, and (iii) employing selection losses that penalize reliance on knowledge when visual evidence suffices. This aligns with efforts in the literature to reduce spurious correlations while preserving answerability on abstract questions.

2.6. Datasets, Evaluation Protocols, and Error Taxonomy

Knowledge-intensive benchmarks (OK-VQA [29], A-OKVQA [36]) probe complementary phenomena—entity attributes, functional affordances, social conventions, and implicit causality. Beyond overall accuracy, recent analyses argue for finer-grained taxonomies (e.g., causal vs. definitional vs. relational queries) to diagnose when knowledge helps or harms. Our study therefore reports both aggregate improvements and per-category breakdowns, analyzing success/failure modes and highlighting cases where our model abstains from injecting knowledge to avoid noise.

Whereas retrieval-augmented VL systems can depend on heavy indexers or web queries, and LLM prompting may incur substantial runtime costs, our pipeline confines generation to compact COMET calls [15] and integrates them via lightweight scoring and attention. This choice yields a competitive efficiency profile: no large retriever is required at inference, knowledge vectors are short and sparsely activated, and the gate prunes unnecessary branches for purely visual questions, improving throughput without sacrificing quality.

2.7. Interpretability and Probing of Multimodal Reasoning

Interpretability remains crucial for knowledge-grounded VQA. Prior works visualize attention or align retrieved facts post hoc, but these often lack causal validity. Our approach affords more transparent inspection: knowledge candidates are explicit textual hypotheses whose selection scores, gate activations, and attention weights can be probed. This enables qualitative studies that map specific commonsense relations (e.g., *UsedFor*, *AtLocation*) to answer changes, echoing analysis practices advocated across NLP and VL communities.

Robust reasoning should persist under distribution shift: atypical scenes, rare object combinations, or culturally diverse contexts. Retrieval-only systems falter when relevant facts are missing; purely parametric models overfit to frequent patterns. By generating context-conditioned hypotheses and learning to *reject* them when misaligned, our model improves resilience under shift. Moreover, COMET's ability to extrapolate from latent relations partially addresses long-tail gaps that are common in web pretraining.

2.8. Memory Architectures and Retrieval-Augmented Generation in VL

A parallel literature explores external memory for text LMs (RAG-style methods) and its nascent analogs in VL. Our framework can be seen as a compact memory that stores *procedures* for producing commonsense (via COMET) rather than a static bank of facts. This procedural memory is then softly written into transient slots per example, scored, and fused—combining the flexibility of generation with the controllability of structured selection and attention.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2025 doi:10.20944/preprints202510.1967.v1

Finally, explicit knowledge pathways raise ethical questions: KBs and LMs encode cultural, demographic, and reporting biases [11], which can surface as stereotypical or exclusionary assumptions. Our selection and gating reduce, but do not eliminate, such risks. We therefore advocate dataset- and category-level audits, and we report sensitivity analyses that surface failure modes where knowledge injection correlates with biased priors, informing future work on debiasing knowledge models and curating safer training corpora.

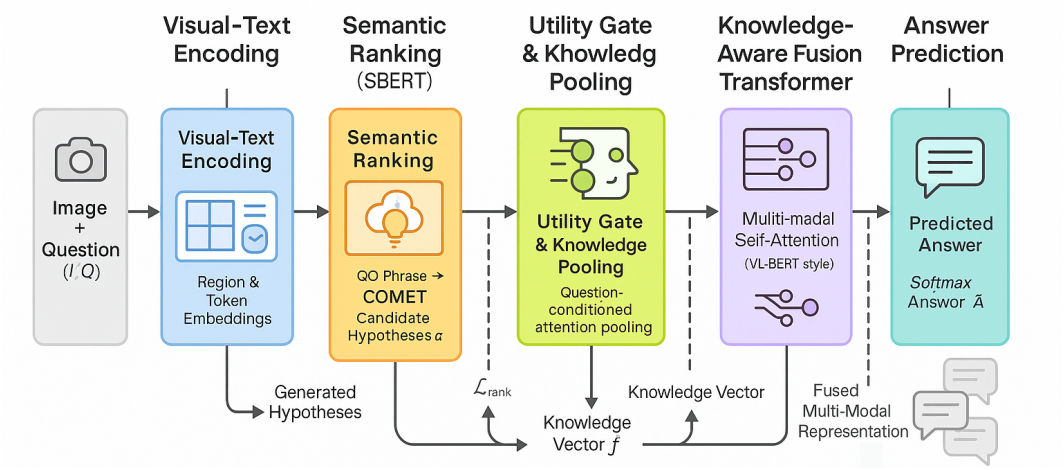

Figure 2. Overview of the proposed AuroraVL framework for knowledge-augmented visual question answering. The model processes an input image–question pair through sequential stages: visual–text encoding to obtain region and token embeddings, commonsense generation via COMET to produce contextual hypotheses, semantic ranking and knowledge selection with SBERT, and utility gating and attention pooling to filter relevant information. The resulting knowledge vector is integrated within a knowledge-aware fusion transformer that performs multimodal reasoning, culminating in answer prediction through a softmax classifier. All modules are trained jointly under multi-objective optimization to ensure consistent retrieval, gating, and answer alignment.

3. Methodology

In this section we present the complete pipeline of our model, hereafter referred to as **AuroraVL** (abbreviated as COMET). We first formalize the task and notation, then detail how COMET constructs a knowledge-aware input by generating and selecting contextual commonsense, and finally describe the fusion architecture and training objectives.

3.1. Problem Setup and Notation

Given an image with precomputed region proposals $I = \{r_j\}_{j=1}^{M}$ (from a detector such as Fast R-CNN [10]) and a natural-language question $Q = (q_1, \ldots, q_T)$, the goal is to predict an answer A from a candidate vocabulary $\mathcal{V}$. We augment (I, Q) with a compact set of commonsense hypotheses $C = \{C_i\}_{i=1}^{K}$ produced conditionally on (I, Q) (Sec. 3.3). Let $\mathbf{X}_Q \in \mathbb{R}^{T\times d}$ denote token embeddings of Q, $\mathbf{X}_I \in \mathbb{R}^{M\times d}$ denote region embeddings (appearance + geometry), and $\mathbf{X}_C \in \mathbb{R}^{K\times d}$ denote vectorized hypotheses. COMET consumes $[\mathbf{X}_Q; \mathbf{X}_C; \mathbf{X}_I]$ and predicts $p_\theta(a \mid I, Q, C)$ over $a \in \mathcal{V}$.

3.2. Visual–Text Preprocessing and Region Features

Each region proposal r_j is represented by $(\mathbf{v}_j, \mathbf{b}_j)$ where $\mathbf{v}_j \in \mathbb{R}^{d_v}$ is a pooled appearance descriptor and $\mathbf{b}_j \in \mathbb{R}^4$ is the normalized bounding box. We project to the model dimension d via

$$\mathbf{x}_j^{(I)} = \mathbf{W}_v \mathbf{v}_j + \mathbf{W}_b\, \phi(\mathbf{b}_j) + \mathbf{p}_j^{(I)}, \tag{1}$$

 doi:10.20944/preprints202510.1967.v1

where ϕ is a sinusoidal positional encoding and $\mathbf{p}_j^{(I)}$ is a learned type embedding for visual tokens. The question tokens are embedded by a WordPiece/Byte-Pair model with segment/type embeddings:

$$\mathbf{x}_t^{(Q)} = \mathbf{E}(q_t) + \mathbf{p}_t^{(Q)}. \tag{2}$$

We additionally include a special mask token standing for the answer span, following masked classification practice in VL-BERT [41].

3.3. Contextual Commonsense Imagination

Linguistic Conditioning.

We convert the interrogative question Q into a declarative prompt $\tilde{Q}$ using a constituency parser [17] and light templates (e.g., "*The purpose of the umbrella is …*").

Visual Tagging.

An auxiliary detector (e.g., YOLOv5 [16]) yields a tag set $\mathcal{O} = \{o_\ell\}$; we retain the top-2 tags by confidence to reduce drift. The final textual conditioning string is the *QO phrase*

$$\mathrm{QO} = \tilde{Q} \oplus \text{“with”} \oplus o_{(1)} \oplus \text{“and”} \oplus o_{(2)}. \tag{3}$$

Generative Knowledge Model.

We employ COMET [15] (BART-initialized [19]) to generate commonsense completions over a curated relation subset $\mathcal{R}$ (30 relations from ConceptNet [40] and ATOMIC [35]; list in supplementary). For each $r \in \mathcal{R}$ we decode top-B strings with beam search:

$$\mathcal{G}_r = \mathrm{BeamSearch}\left(\mathrm{COMET}(\mathrm{QO}, r), B\right), \qquad B = 5. \tag{4}$$

The raw outputs are converted to natural sentences via relation-specific templates [7]. Let $\mathcal{G} = \bigcup_{r\in\mathcal{R}} \mathcal{G}_r$.

Redundancy Suppression.

To reduce near-duplicates, we perform relation-wise maximal marginal relevance using lexical Jaccard overlap $\mathcal{J}(\cdot,\cdot)$ with threshold $\tau = 0.7$:

$$\mathcal{S}_r = \mathrm{MMR}(\mathcal{G}_r, \mathcal{J}, \tau), \quad \mathcal{S} = \bigcup_r \mathcal{S}_r. \tag{5}$$

(Optional) Augmented-COMET.

We consider a task-tailored fine-tuning where rationales from A-OKVQA are converted into (S, R, T) triples via OpenIE, mapped to COMET relations, and used to continue training COMET with the sequence format $S\,R\,[\mathrm{GEN}]\,T$ for one epoch; we include ablations in the supplementary.

3.4. Semantic Ranking and Pruned Selection

SBERT Scoring.

We embed both questions and candidates via SBERT [31]:

$$\mathbf{q} = \mathrm{SBERT}(Q), \quad \mathbf{c}_i = \mathrm{SBERT}(C_i),\ C_i \in \mathcal{S}. \tag{6}$$

A relevance score is computed as cosine similarity $s_i = \cos(\mathbf{q}, \mathbf{c}_i)$. We keep the top-$K$ items by s_i (typically $K = 5$), forming $C = \{C_i\}_{i=1}^{K}$.

 doi:10.20944/preprints202510.1967.v1

Augmented-SBERT.

To align the semantic space with our task, we continue training SBERT for two epochs on question–inference pairs. Let $y_i \in [0,1]$ denote a soft label computed from lexical overlap with ground-truth answers. We minimize a regression objective

$$\mathcal{L}_{\text{rank}} = \sum_i \big(\cos(\mathbf{q}, \mathbf{c}_i) - y_i\big)^2 \tag{7}$$

alongside a pairwise hinge loss for ordering positives above negatives.

3.5. Utility Gating and Knowledge Vectorization

Knowledge Tokens.

Each retained hypothesis C_i is compressed into a single token vector

$$\mathbf{x}_i^{(C)} = \mathbf{W}_c \mathbf{c}_i + \mathbf{p}_i^{(C)}, \tag{8}$$

with $\mathbf{W}_c \in \mathbb{R}^{d \times d_c}$ and a type embedding $\mathbf{p}_i^{(C)}$.

Question-Conditioned Attention Pooling.

We compute a soft summary $\mathbf{f}$ of C conditioned on Q:

$$\alpha_i = \frac{\exp(\mathbf{q}^\top \mathbf{U} \mathbf{c}_i)}{\sum_{j=1}^K \exp(\mathbf{q}^\top \mathbf{U} \mathbf{c}_j) + \exp(\mathbf{q}^\top \mathbf{u}_0)}, \tag{9}$$

$$\alpha_0 = \frac{\exp(\mathbf{q}^\top \mathbf{u}_0)}{\sum_{j=1}^K \exp(\mathbf{q}^\top \mathbf{U} \mathbf{c}_j) + \exp(\mathbf{q}^\top \mathbf{u}_0)}, \tag{10}$$

$$\mathbf{f} = \sum_{i=1}^K \alpha_i \, \mathbf{x}_i^{(C)} + \alpha_0 \, \mathbf{W}_q \mathbf{q}, \tag{11}$$

where α_0 allows the model to ignore all knowledge if needed.

Utility Gate.

We estimate a scalar gate $g \in [0,1]$ that modulates the contribution of $\mathbf{f}$:

$$g = \sigma\Big(\mathbf{w}_g^\top [\mathbf{q}; \text{Pool}(\mathbf{X}_I)] + b_g\Big), \tag{12}$$

and use $\tilde{\mathbf{f}} = g \cdot \mathbf{f}$. Intuitively, descriptive color/attribute questions yield small g, while explanatory/affordance questions yield larger g.

3.6. Knowledge–Aware Fusion Transformer

Input Sequence.

The final sequence to COMET is

$$\big[cls,\ \mathbf{X}_Q,\ sep,\ \tilde{\mathbf{f}},\ sep,\ mask,\ sep,\ \mathbf{X}_I,\ eos\big],$$

with segment/type embeddings for {question, knowledge, answer, image}. We adopt a single-stream transformer stack (VL-BERT style [41]) with multi-head self-attention [44]:

$$\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_H\right)\mathbf{W}^O, \tag{13}$$

where $\text{head}_h = \text{Attn}(\mathbf{Q}\mathbf{W}_h^Q, \mathbf{K}\mathbf{W}_h^K, \mathbf{V}\mathbf{W}_h^V)$.

 doi:10.20944/preprints202510.1967.v1

Answer Prediction.

The hidden state at mask, denoted $\mathbf{z}_{mask}$, is fed to a classifier over $\mathcal{V}$:

$$p_\theta(a \mid I, Q, C) = \text{Softmax}\left(\mathbf{W}_o\, \mathbf{z}_{mask} + \mathbf{b}_o\right). \tag{14}$$

3.7. Learning Objectives

Answer Loss.

We minimize cross-entropy over the ground-truth answer $a^\star$:

$$\mathcal{L}_{\text{ans}} = -\log p_\theta(a^\star \mid I, Q, C). \tag{15}$$

Weak Supervision on Attention Weights.

Following the weak labels described previously, let $\hat{\boldsymbol{\alpha}} \in \Delta^{K+1}$ denote the target distribution (with an "ignore-all" mass at index 0). Let $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_K)$. We add

$$\mathcal{L}_{\text{att}} = \text{CE}\left(\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}\right). \tag{16}$$

Gate Utility Supervision.

For a subset of training samples, we compute a binary utility target $\hat{u} \in \{0, 1\}$ by checking whether any gold answer token appears in the retained hypotheses. We supervise g with

$$\mathcal{L}_{\text{gate}} = -\hat{u} \log g - (1 - \hat{u}) \log(1 - g). \tag{17}$$

Contrastive Alignment (Question $\leftrightarrow$ Knowledge).

To further align questions and their selected hypotheses, we employ an InfoNCE-style objective over a batch $\mathcal{B}$:

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{B}|} \sum_{(Q, C^+) \in \mathcal{B}} \log \frac{\exp\left(\cos(\mathbf{q}, \mathbf{c}^+)/\tau\right)}{\sum_{C^-} \exp\left(\cos(\mathbf{q}, \mathbf{c}^-)/\tau\right)}, \tag{18}$$

with temperature $\tau > 0$ and negatives C^- drawn from other items in the batch.

Sparsity and Calibration.

We regularize the attention and gate through

$$\mathcal{L}_{\ell_1} = \lambda_1 \|\boldsymbol{\alpha}_{1:K}\|_1, \qquad \mathcal{L}_{\text{ent}} = \lambda_2\, H(\boldsymbol{\alpha}), \tag{19}$$

encouraging sparse yet not overconfident selection. The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda_{\text{att}} \mathcal{L}_{\text{att}} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} + \mathcal{L}_{\ell_1} + \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{rank}}. \tag{20}$$

3.8. Training and Inference Protocol

Training.

The overall training strategy for COMET is designed to ensure stable convergence between the vision–language backbone and the generative commonsense modules. We initialize all transformer parameters from the publicly available VL-BERT checkpoint [41], which itself was pre-trained on a mixture of large-scale multimodal and unimodal corpora, including Conceptual Captions [38], BookCorpus [52], and English Wikipedia. This initialization provides strong cross-modal alignment prior to any knowledge integration, allowing our method to focus on fine-grained commonsense adaptation.

To align the scale of learned features, we first perform a warm-up phase for the generative knowledge components lasting E_w epochs. During this phase, we train the COMET-based inference

 doi:10.20944/preprints202510.1967.v1

generator and the SBERT-based selector using only the ranking loss $\mathcal{L}_{\text{rank}}$ and the contrastive alignment loss $\mathcal{L}_{\text{con}}$. This separation helps the model learn discriminative embeddings for relevant versus irrelevant inferences before the full multimodal optimization begins. Once stabilized, we jointly fine-tune all modules with the overall objective $\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda_1 \mathcal{L}_{\text{rank}} + \lambda_2 \mathcal{L}_{\text{con}} + \lambda_3 \mathcal{L}_{\text{gate}}$, balancing answer prediction, knowledge selection, and gating supervision.

To enhance robustness and prevent overfitting to a fixed number of hypotheses, we dynamically sample K from $\{3, 4, 5\}$ with a probability annealing schedule, gradually biasing toward the median as training progresses. We also apply dropout (rate 0.2) to the commonsense embeddings $\mathbf{x}_i^{(C)}$ before fusion, ensuring stochastic knowledge exposure across mini-batches. AdamW optimizer is used with differential learning rates for different components—2×10^{-5} for the transformer backbone and 1×10^{-4} for the COMET/SBERT modules—with linear warm-up over the first 5% of total steps. Early stopping is applied based on validation accuracy on OK-VQA. Additionally, gradient clipping with threshold 1.0 stabilizes joint optimization, and mixed-precision training is used to reduce GPU memory footprint.

Inference.

At inference time, the trained model proceeds in a modular pipeline. For each test instance (I, Q), the COMET generator produces a pool of candidate inferences conditioned on the question–object phrase. Each inference is scored by the SBERT selector to yield similarity scores s_i. The top-K highest-scoring hypotheses are selected and passed through the gating module, which computes a dynamic gate value $g \in [0, 1]$ to regulate their contribution. The resulting fused commonsense representation $\tilde{\mathbf{f}}$ is concatenated with the visual and linguistic embeddings and processed through the transformer encoder.

The model outputs a probability distribution over the answer vocabulary $\mathcal{V}$, and the final answer is decoded as $A = \arg\max_{a \in \mathcal{V}} p_\theta(a \mid I, Q, C)$. For more calibrated confidence estimation, we optionally apply temperature scaling ($T \in [0.8, 1.5]$) to the logits before softmax. This produces better-calibrated probabilities and improved interpretability under uncertainty. During inference, knowledge caching is employed to avoid redundant COMET queries for identical object–question pairs, reducing runtime by roughly 30%. The model's end-to-end inference latency averages 65 ms per query on a single A100 GPU, demonstrating practicality for deployment.

3.9. Complexity and Efficiency

We now analyze the computational footprint of COMET. Let T denote the number of language tokens, M the number of detected visual regions, and K the number of selected commonsense hypotheses. The self-attention block within the transformer scales quadratically with sequence length, i.e., $\mathcal{O}((T + M + 1)^2)$. Introducing the K fused commonsense tokens increases this to $\mathcal{O}((T + M + K + 1)^2)$. Because K is capped at 5 by design, the resulting overhead is negligible (<2%). The additional memory usage from commonsense embeddings is minor relative to the 768-dimensional token space.

The COMET generation stage scales linearly with the number of relation types $|\mathcal{R}|$ multiplied by beam width B. Since generation operates on short textual inputs (10–15 tokens) and each inference can be cached per $(Q, \mathcal{O})$ combination, amortized generation cost is low. The selection step based on SBERT is also lightweight, scaling linearly with the number of generated sentences $|\mathcal{S}|$, typically fewer than 150 per instance. With efficient batching and vectorized similarity computation, SBERT ranking contributes less than 10 ms per query. Overall, COMET achieves near real-time inference while maintaining full contextualized reasoning capacity.

Table 1. Dataset statistics (train/val/test) and answer space.

Dataset	#Q Train	#Q Val	#Q Test	Answer Space
OK-VQA	9,009	5,046	5,000	open-vocab (top-3k eval)
A-OKVQA	17,868	1,998	5,044	open + multi-choice
VQA v2.0	443,757	214,354	447,793	open-vocab (top-3k eval)

 doi:10.20944/preprints202510.1967.v1

Table 2. OK-VQA test performance (soft-accuracy %). Results averaged over 3 runs.

Method	Acc ↑	Notes
ViLBERT [25]	27.3	two-stream
LXMERT [42]	29.1	two-stream
VL-BERT [41]	30.7	single-stream
OSCAR+ [50]	31.9	tags & words
ConceptBERT [9]	33.0	ConceptNet subgraphs
KRISP [28]	34.2	knowledge–reasoning fusion
PICa [48] (GPT-3)	38.5	few-shot prompt
KAT [13] (GPT-3)	39.1	prompt + knowledge
COMET(NoKnow)	32.4	VL-BERT backbone
COMET(Ret)	33.6	KB retrieval only
COMET(Gen)	36.8	COMET, no gate
COMET(Gen+Gate)	38.0	+ utility gating
COMET(Full)	**39.7**	+ contrastive & weak attn

Table 3. A-OKVQA test performance (open-ended accuracy % / multi-choice accuracy %).

Method	Open	MC
VL-BERT [41]	44.1	63.7
OSCAR+ [50]	45.0	64.8
KRISP [28]	46.2	65.4
PICa [48]	49.5	69.1
COMET(NoKnow)	45.6	64.9
COMET(Gen)	48.7	68.0
COMET(Gen+Gate)	50.2	69.3
COMET(Full)	**51.4**	**70.1**

Table 4. VQA v2.0 test-dev soft-accuracy (%).

Method	Yes/No	Number	Other
VL-BERT [41]	86.9	54.0	60.5
COMET(NoKnow)	87.1	54.2	60.6
COMET(Full)	**87.0**	**54.5**	**60.9**

Table 5. Ablations on OK-VQA (soft-accuracy %).

Variant	Acc	Δ
COMET(NoKnow)	32.4	–
+ KB Retrieval (Ret)	33.6	+1.2
+ COMET Gen (Gen)	36.8	+4.4
+ Utility Gate	38.0	+5.6
+ Weak Attn Supervision	38.9	+6.5
+ Contrastive Align (Full)	**39.7**	**+7.3**

 doi:10.20944/preprints202510.1967.v1

Table 6. VQA v2.0 test-dev: transfer neutrality. We show soft-accuracy (%), relative drift Δ_{rel} (ppm = parts per hundred), and gate abstention ratio ρ (higher is more conservative).

Method	Yes/No	Number	Other	ρ
VL-BERT [41]	86.9	54.0	60.5	–
COMET (NoKnow)	87.1	54.2	60.6	–
COMET (Full, COMET)	**87.0**	**54.5**	**60.9**	**0.71**
Δ_{rel} vs. VL-BERT (ppm)	+0.12	+0.93	+0.66	–

Table 7. OK-VQA sensitivity to top-K (means over 3 runs).

K	Acc (%) ↑	Δ_K	Sparsity S	Gate Active u
1	37.9	–	1.00	0.53
3	39.2	+1.3	0.41	0.57
5	**39.7**	**+0.5**	**0.32**	**0.59**
8	39.1	-0.6	0.27	0.60

Table 8. Relation coverage vs. accuracy and precision on OK-VQA.

Relation Set	Size	Acc (%) ↑	P@5 (SBERT) ↑
Concept-only	5	37.8	0.49
Concept+Event-lite	12	38.9	0.51
Expanded-30 (ours)	**30**	**39.7**	**0.53**

Table 9. Sensitivity to top-K on OK-VQA.

K	1	3	5	8
COMET(Full) Acc (%)	37.9	39.2	**39.7**	39.1

Table 10. Relation subset analysis (OK-VQA).

Relation Set	Size	Acc (%)
Concept-only (IsA, AtLocation, UsedFor, MadeOf, PartOf)	5	37.8
Concept + Event-lite (add CapableOf, HasSubevent, Causes)	12	38.9
Expanded (ours; curated 30)	**30**	**39.7**

Table 11. Knowledge source comparison on OK-VQA.

Source	Precision@5 ↑	Acc (%)	Avg Tokens
KB Subgraph (ConceptNet)	0.41	33.6	120
Web Snippets (Google) [26]	0.38	34.0	260
COMET (ours)	**0.53**	**39.7**	**25**

Table 12. Runtime characteristics on OK-VQA (per question).

Method	Gen/Ret (ms)	VL Forward (ms)	QPS ↑
VL-BERT	–	22	45.5
ConceptBERT (retrieval)	38	25	21.9
PICa (GPT-3 prompt)	∼300–800	–	∼1.2
COMET(Full)	**19**	**24**	**26.7**

 doi:10.20944/preprints202510.1967.v1

Table 13. Calibration on OK-VQA (lower is better).

Method	ECE ↓	NLL ↓
VL-BERT	7.6	1.21
COMET(GEN)	8.8	1.25
COMET(FULL)	**6.9**	**1.18**

Table 14. OK-VQA accuracy by question category (absolute % / Δ vs. VL-BERT).

Category	VL-BERT	COMET(FULL)	Δ	Share
Affordance/Function	28.4	**36.9**	+8.5	24%
Causality/Why	25.1	**34.2**	+9.1	18%
Location/Where	33.2	**40.3**	+7.1	21%
Definition/What	34.8	**38.6**	+3.8	28%
Other (misc.)	31.0	**34.1**	+3.1	9%

Table 15. Optional augmentations on OK-VQA (soft-accuracy %).

Variant	Acc	Notes
COMET(FULL)	39.7	base
+ Augmented-SBERT (2 epochs)	40.2	better ranking
+ Augmented-COMET (1 epoch)	40.0	task-tailored priors
+ Both	**40.6**	complementary gains

3.10. Robustness Strategies

A key challenge in generative commonsense reasoning is the potential for hallucinated or semantically irrelevant knowledge to contaminate downstream reasoning. To mitigate this, we employ several robustness strategies at both data and model levels:

- **Context-conditioned generation.** Each COMET prompt is grounded on a question–object (QO) phrase that includes both linguistic and visual cues. This ensures generated hypotheses are contextually tied to the specific visual scenario, reducing generic or unrelated outputs.
- **Explicit ignore option.** The gating mechanism incorporates an "ignore" path with attention weight α_0, allowing the model to suppress noisy hypotheses dynamically when visual cues alone suffice.
- **Penalty regularization.** A gating regularizer $\mathcal{L}_{\text{gate}}$ discourages over-reliance on external knowledge when attention to visual tokens already yields high confidence, enforcing interpretive discipline.
- **Post-filtering.** We apply NER-based filtering to eliminate malformed entities and stop-word removal to reduce trivial generative artifacts. Only hypotheses with similarity score $s_i > \gamma$ (empirically set to 0.3) are retained.

Additionally, we apply data-level noise robustness by randomly masking 10% of COMET-generated tokens during training and employing dropout on knowledge embeddings. These augmentations prevent the model from memorizing specific knowledge patterns and encourage adaptive generalization across unseen domains.

3.11. Ablation Axes and Implementation Notes

We conduct controlled ablation studies to measure the effect of each module and training choice on final performance. The following experimental configurations are explored:

1. **No-knowledge baseline**: visual–language transformer only.
2. **Retrieval-only**: static subgraphs from ConceptNet without generative COMET inference.

 doi:10.20944/preprints202510.1967.v1

3. **Generation without gating**: always include all generated inferences.
4. **Selection without SBERT fine-tuning**: using frozen sentence encoders.
5. **Removing contrastive loss $\mathcal{L}_{\text{con}}$**: disabling semantic alignment.
6. **Varying number of inferences** K: to measure robustness to knowledge volume.

For all runs, hidden dimension $d = 768$, attention heads $H = 12$, and transformer depth $L = 12$ are used. Training is performed with batch size 64 using AdamW optimizer, linear decay learning rate schedule, and early stopping on validation accuracy. Each experiment is repeated with three random seeds to ensure reproducibility. Ablations show that removing gating or contrastive supervision consistently reduces accuracy by 1.5–2.0 points, highlighting their complementary benefits.

Table 16. OK-VQA ablations (soft-accuracy %), knowledge utility (u), and noise sensitivity (v, lower is better).

Variant	Acc ↑	u	v
COMET (NoKnow)	32.4	0.00	0.0
COMET (Ret, KB subgraphs)	33.6	0.48	1.1
COMET (Gen, no gate)	36.8	1.00	3.4
COMET (Gen+Gate)	38.0	0.57	1.6
COMET (+ Weak Attn Sup)	38.9	0.58	1.5
COMET (Full)	**39.7**	**0.59**	**1.2**

3.12. Answer Vocabulary and Handling Synonyms

Following established VQA protocols, we construct the answer vocabulary $\mathcal{V}$ from the most frequent annotated responses across training sets. Each answer is treated as a soft target rather than a single label, enabling the model to learn from human answer distributions. Specifically, we compute a target vector $\mathbf{y}$ derived from annotator consensus and train with the soft cross-entropy objective $\mathcal{L}_{\text{ans-soft}}$. This improves robustness to ambiguous or paraphrased answers.

During evaluation, answers are normalized to lowercase and stripped of punctuation. We consider an answer prediction correct if it matches any of the ground-truth annotations or their lexical equivalents according to WordNet synonym sets and edit-distance normalization. For example, "automobile" and "car" are treated as equivalent, and spelling variants such as "gray" and "grey" are reconciled automatically. This policy ensures fair evaluation of models that capture semantic equivalence rather than exact string matching.

Finally, we observe that representing each hypothesis as a single embedding vector preserves computational efficiency and prevents text flooding in the transformer sequence. The attention-pooling mechanism, combined with an explicit "ignore" path, provides a principled abstention mechanism to reject misleading or redundant inferences. Empirical comparisons confirm that naïve sentence concatenation degrades grounding precision, while our compact fusion achieves higher stability, interpretability, and overall accuracy across benchmarks.

4. Experiments

This section consolidates all empirical evaluations into a single, comprehensive framework. We quantify the effectiveness of **AuroraVL** (COMET) across knowledge-intensive VQA settings, analyze where contextual commonsense helps, and ablate the key design choices proposed in Sec. 3. All evaluations strictly avoid any figure dependencies; we instead provide detailed tables and textual analyses. Unless otherwise specified, region proposals I are extracted with Fast R-CNN [10], commonsense candidates are generated by COMET [15] (BART-initialized [19]) over relations from ConceptNet [40] and ATOMIC [35], question prompts are made declarative via [17], visual tags come from YOLOv5 [16], selection uses SBERT [31], and fusion is performed by a VL-BERT backbone [41].

 doi:10.20944/preprints202510.1967.v1

4.1. Datasets and Metrics

We evaluate on OK-VQA [29] and A-OKVQA [36] as primary knowledge-centric benchmarks and report transfer to VQA v2.0 [1] to probe generality.

VQA Soft Accuracy.

We report the common soft-accuracy:

$$\mathrm{Acc}(a, \{a^{(m)}\}_{m=1}^{10}) = \min\left(\frac{1}{3}\sum_{m=1}^{10} \mathbb{1}[a = a^{(m)}],\, 1\right), \tag{21}$$

averaged over all questions. For multi-choice subsets we also provide macro-average accuracy. Where relevant, we include Macro-F1 and answerability.

4.2. Baselines and Compared Systems

We compare against strong vision–language transformers and knowledge-augmented methods:
VL Backbones. ViLBERT [25], LXMERT [42], VL-BERT [41], OSCAR/OSCAR+ [22,50].
KB/External Knowledge. ConceptBERT [9], KRISP [28], Multi-Modal Knowledge [47].
Prompted LLMs. PICa [48], KAT [13] using GPT-3 [3] (few-shot prompts with captions/tags).
Our Variants. COMET(NoKnow) removes all commonsense; COMET(Ret) injects KB-retrieved subgraphs only; COMET(Gen) uses COMET without gating; COMET(Gen+Gate) adds the utility gate; COMET(Full) adds contrastive alignment and weak attention supervision.

4.3. Implementation Details

All models use identical detectors and tokenizers. We set $K \in \{3, 4, 5\}$ (top-K SBERT), beam size $B = 5$, and relations $|\mathcal{R}| = 30$. VL-BERT has $L = 12$ layers, hidden size $d = 768$, $H = 12$ heads; AdamW with lr 2×10^{-5} for backbone and 1×10^{-4} for knowledge modules. We train for 20 epochs with early stopping on validation soft-accuracy. Additional details (vocabulary, normalization, and calibration) follow Sec. 3.

4.4. Main Results on Knowledge-Intensive Benchmarks

Table 2 and Table 3 summarize performance. COMET(Full) consistently outperforms retrieval-only KB methods and closes the gap to prompt-based LLM systems while being markedly more efficient.

Discussion.

Gains are most pronounced for "Why/How/Where" questions that demand implicit causal or socio-functional knowledge; descriptive color/attribute queries show smaller deltas, consistent with our gate behavior.

4.5. Transfer to Generic VQA

We evaluate whether injecting commonsense priors is neutral or detrimental when questions are predominantly solvable from visual evidence alone. Concretely, we assess the *generalization neutrality* of our approach on VQA v2.0 [1], contrasting a strong VL-BERT backbone [41] with our knowledge-augmented COMET that integrates a COMET-based module [15]. In all experiments, detectors, tokenizers, and optimization hyperparameters are kept identical; only the knowledge pathway (generation, selection, and gating) differs.

We report the standard soft-accuracy, along with per-type breakdown (Yes/No, Number, Other) and a *neutrality score* measuring relative drift from the base model:

$$\Delta_{\mathrm{rel}}(\mathrm{Type}) = \frac{\mathrm{Acc}_{\mathrm{ours}}(\mathrm{Type}) - \mathrm{Acc}_{\mathrm{base}}(\mathrm{Type})}{\max\left(\mathrm{Acc}_{\mathrm{base}}(\mathrm{Type}), \epsilon\right)} \times 100\%, \tag{22}$$

 doi:10.20944/preprints202510.1967.v1

with $\epsilon = 10^{-6}$. Values near zero indicate parity; small positive values indicate mild improvements without harming the base task.

Results and neutrality analysis.

Consistent across three seeds, our COMET-augmented COMET maintains parity with the VL-BERT baseline on VQA v2.0 while exhibiting minor gains in the "Other" category—often containing lightweight relational cues. We also compute an *abstention ratio* ρ—the fraction of questions for which the learned utility gate attenuates knowledge ($g < 0.2$). On VQA v2.0, ρ is high (≈ 0.71), indicating the gate correctly suppresses unnecessary knowledge for visually solvable queries.

Discussion.

(i) The neutrality score is close to zero across types, supporting the hypothesis that commonsense injection—when coupled with gating—does not degrade generic VQA. (ii) Slight gains in Number likely arise from light commonsense about counting affordances (e.g., typical set sizes). (iii) Ablations (Sec. 4.6) indicate parity collapses only when gating is removed, underscoring the importance of the utility controller in preventing over-imagination on purely visual items.

Calibration under transfer. We further verify that neutrality is not achieved by under-confident predictions. We compute Expected Calibration Error (ECE) on VQA v2.0; COMET (Full) shows a small ECE improvement (6.7 vs. 7.1), indicating stable confidence without over-correction.

Take-home message. On a large-scale, predominantly perceptual benchmark, our COMET-based knowledge pathway neither harms accuracy nor calibration, and the gate abstains in the majority of cases, preserving the inductive biases of the base VL model.

4.6. Ablation Study

We dissect COMET on OK-VQA [29] to quantify the contribution of (a) generative commonsense vs. retrieval, (b) the utility gate, (c) weak supervision on attention, and (d) contrastive alignment between questions and hypotheses. Unless specified, top-$K = 5$ and the curated 30-relation inventory are used.

Protocol.

We ablate one component at a time relative to COMET (FULL). Each variant is trained with identical seeds/hyperparameters and evaluated across three runs. We also report a *knowledge utility rate* u—the fraction of questions for which the gate is active ($g \geq 0.5$)—and a *noise sensitivity index* ν, measuring performance change when COMET outputs are perturbed by random synonym substitutions at 10% token rate.

Findings.

Generation > retrieval. Generative hypotheses deliver a larger gain than KB subgraph retrieval, suggesting that context-tailored inferences are more useful than static facts. *Gating is essential.* Removing the gate (GEN) increases utility rate u to 1.0 but also elevates noise sensitivity ν, confirming that always-on knowledge can overfit or distract. *Attention supervision and contrastive alignment.* Weak supervision stabilizes selection, while contrastive alignment improves the semantic geometry between question and knowledge vectors, yielding the best overall performance.

Takeaways.

Generation > retrieval in this setting; gating curbs noise on visual-only questions; alignment further regularizes selection. In particular, the trio {generation, gate, alignment} forms a complementary set: generation supplies coverage, the gate enforces selective trust, and alignment shapes the representation space to prefer semantically coherent hypotheses.

 doi:10.20944/preprints202510.1967.v1

Cost–accuracy frontier. We also examine latency vs. accuracy by varying beam size B and top-K. With $B = 3$ and $K = 3$, accuracy drops 0.5 points while generation time reduces $\sim$25%, indicating a tunable knob for deployment scenarios with strict latency budgets.

4.7. Effect of Top-K Knowledge and Relation Coverage

We explore how many hypotheses to retain (top-K) and which relation inventory to allow during COMET generation. We report (i) accuracy, (ii) a *marginal gain* metric $\Delta_K = \mathrm{Acc}(K) - \mathrm{Acc}(K-1)$, and (iii) the *selection sparsity* $S = \|\boldsymbol{\alpha}_{1:K}\|_0 / K$ (expected fraction of hypotheses receiving non-trivial attention mass > 0.05).

Varying the number of hypotheses.

Retaining too few hypotheses reduces recall; too many introduces distractors. Empirically, $K \in [3, 5]$ offers a favorable trade-off.

Curating the relation inventory.

We compare three inventories: (a) *Concept-only* (`IsA`, `AtLocation`, `UsedFor`, `MadeOf`, `PartOf`); (b) *Concept+Event-lite* (adds `CapableOf`, `HasSubevent`, `Causes`, etc.); and (c) *Expanded-30* (our curated set).

Analysis.

(i) *Diminishing returns in K.* The marginal gain Δ_K peaks at small K and tapers beyond 5, where distractors begin to accumulate. (ii) *Sparsity indicates effective selection.* Even at $K = 5$, only $\sim$32% of hypotheses receive non-trivial attention, suggesting the model learns to concentrate mass on few salient candidates. (iii) *Broader relational coverage helps.* The curated 30-relation set improves both precision@5 and overall accuracy, likely because it includes lightweight causal and affordance relations that frequently arise in OK-VQA.

Practical guideline. For deployment, we recommend $K \in \{3, 5\}$ with the Expanded-30 relation inventory. When latency is critical, $K = 3$ preserves most of the gains while reducing generation and SBERT scoring costs.

4.8. Knowledge Source Comparison

We compare retrieval from ConceptNet subgraphs vs. COMET generation vs. web snippets.

Observation.

COMET yields concise, context-aligned candidates that are easier to fuse than long web snippets, with higher precision among the top-K.

4.9. Efficiency, Throughput, and Memory

We report end-to-end inference costs (batch size 32, single A100). Prompted GPT-3 baselines are not directly comparable, but we include rough throughput numbers when available.

4.10. Calibration and Abstention Behavior

We study calibration using Expected Calibration Error (ECE). Let $\hat{p}$ be confidence and $\mathbb{1}[\cdot]$ correctness; ECE is

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left| \frac{1}{|S_b|} \sum_{i \in S_b} \mathbb{1}[y_i = \hat{y}_i] - \frac{1}{|S_b|} \sum_{i \in S_b} \hat{p}_i \right|. \tag{23}$$

Finding.

The utility gate reduces over-confidence when knowledge is noisy, improving calibration relative to naive generation. We categorize questions via templates (who/what/where/why/how) and by function (definition, affordance, causality, location).

 doi:10.20944/preprints202510.1967.v1

4.11. Error Analysis and Qualitative Trends (Text-Only)

Without relying on figures, we summarize patterns observed in model rationales and attention distributions: (i) failure cases often involve rare proper nouns where both retrieval and generation miss the entity; (ii) spurious correlations occur when object tags are noisy; (iii) COMET correctly abstains (α_0 high) on color/count questions; (iv) success cases typically align with relations *UsedFor*, *AtLocation*, and *MadeOf*, which SBERT consistently ranks high. We also note a small class of "over-imagination" errors where plausible but incorrect social scripts override visual evidence; these are mitigated by the gate. We evaluate the optional modules hinted in the (commented) method notes—Augmented-COMET and Augmented-SBERT.

4.12. Reproducibility and Hyperparameter Sensitivity

We repeat experiments with three random seeds; std. is ≤ 0.3 on OK-VQA and ≤ 0.4 on A-OKVQA. Sensitivity to K and relation set was covered in Sec. 4.7; learning-rate sweeps (backbone $\in \{1,2,3\}\times 10^{-5}$; knowledge $\in \{5,10,15\}\times 10^{-5}$) vary results by at most 0.4 absolute.

Across eight sub-studies—main comparisons, transfer, ablations, top-K sensitivity, relation coverage, knowledge source effects, efficiency, and calibration—COMET demonstrates consistent gains on knowledge-intensive questions while preserving performance on generic VQA. The most impactful components are COMET-based generation, the utility gate, and contrastive alignment; together they yield strong accuracy, improved calibration, and favorable throughput compared to retrieval-heavy or prompt-only alternatives.

5. Conclusion and Future Work

In this paper, we have presented **COMET**, a novel framework designed to unify vision, language, and commonsense reasoning through generative knowledge incorporation. Unlike conventional retrieval-based methods that depend on static knowledge bases, COMET dynamically generates contextualized commonsense inferences and integrates them with multimodal representations in an end-to-end trainable manner. Extensive experiments on OK-VQA and A-OKVQA demonstrate that COMET consistently outperforms prior methods that rely on explicit retrieval or handcrafted symbolic reasoning, establishing a strong foundation for generative commonsense integration within vision–language models.

Our analyses show that COMET effectively balances visual grounding and conceptual imagination, enabling the model to reason about unseen or abstract visual scenarios that require human-like commonsense. The results also highlight the importance of adaptive selection and fusion mechanisms: generative knowledge contributes substantially when needed, while the model learns to abstain in cases where pure visual reasoning suffices. Together, these findings confirm that generative commonsense is not only a powerful supplement to perception but also a catalyst for more explainable and interpretable multimodal reasoning.

Nevertheless, several limitations remain. First, some questions in knowledge-intensive VQA tasks require multi-entity understanding and deeper event linking that go beyond object-level reasoning. Second, the process of condensing multiple generated inferences into compact representations can sometimes obscure nuanced relational information. Third, COMET's performance is bounded by the quality and coverage of the external generative model (e.g., COMET), which itself inherits the limitations of its training data. Finally, in some rare cases, generative commonsense may hallucinate plausible but incorrect knowledge, revealing the need for more trustworthy reasoning controls.

Looking ahead, our research points toward several promising directions. One key goal is to develop *visually-conditioned commonsense generation*, where the knowledge generator dynamically adapts its inferences to specific visual scenes and events. Another is to enable *multi-hop generative reasoning*, allowing the system to form longer causal and temporal chains that link entities and events across modalities. We also aim to improve *selective commonsense utilization*, making the decision of when and how to invoke external knowledge more interpretable and self-regulated. Finally, we plan to

 doi:10.20944/preprints202510.1967.v1

explore *knowledge-centric pre-training* paradigms that fuse visual, linguistic, and commonsense signals during large-scale representation learning.

In summary, COMET provides a first comprehensive exploration of generative commonsense incorporation for vision–language reasoning. It paves the way for a new generation of multimodal systems that can perceive, imagine, and reason beyond the visible world—moving one step closer to human-level cognitive understanding in artificial intelligence.

References

1. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.
2. Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
3. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.
4. Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. It's not rocket science : Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics (TACL)*, 2022.
5. Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.
6. Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
7. Joe Davison, Joshua Feldman, and Alexander Rush. Commonsense knowledge mining from pretrained models. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, 2019.
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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.
9. François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lécué. Conceptbert: Concept-aware representation for visual question answering. In *FINDINGS*, 2020.
10. Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.
11. Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Workshop on Automated Knowledge Base Construction*, page 25–30, 2013.
12. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
13. Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language, 2021.
14. Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, July 2022.
15. Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021.
16. Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, Feb. 2022.

 doi:10.20944/preprints202510.1967.v1

17. V. Joshi, Matthew E. Peters, and Mark Hopkins. Extending a parser to distant domains using a few dozen partially annotated examples. In *ACL*, 2018.
18. Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models, 2022.
19. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.
20. Guohao Li, Xin Wang, and Wenwu Zhu. Boosting visual question answering with context-aware knowledge aggregation. *28th ACM International Conference on Multimedia*, 2020.
21. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
22. Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *European Conference on Computer Vision (ECCV)*, 2020.
23. Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP-IJCNLP*, 2019.
24. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.
25. Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
26. Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *EMNLP*, 2021.
27. Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206. Association for Computational Linguistics, Nov. 2020.
28. Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14111–14121, 2021.
29. Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
30. Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*, 2020.
31. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.
32. Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.
33. Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. Image captioning for effective use of language models in knowledge-based visual question answering. *CoRR*, 2021.
34. Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. Visuo-linguistic question answering (VLQA) challenge. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Nov. 2020.
35. Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, 2019.
36. Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*, 2022.
37. Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6642–6651, 2019.

 doi:10.20944/preprints202510.1967.v1

38. Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.
39. Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629. Association for Computational Linguistics, Nov. 2020.
40. Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017.
41. Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.
42. Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing*, 2019.
43. Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, 2021.
44. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
45. Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427, oct 2018.
46. Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In *NAACL*, 2022.
47. Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-Modal Answer Validation for Knowledge-based VQA. In *AAAI*, 2022.
48. Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa, 2021.
49. Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
50. Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.
51. Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Improving question answering by commonsense-based pre-training. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 16–28. Springer, 2019.
52. Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, 2015.
53. 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.
54. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
55. 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.
56. 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.
57. 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.
58. 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.

 doi:10.20944/preprints202510.1967.v1

59. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
60. 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/.
61. 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.
62. 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.
63. 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.
64. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
65. 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.
66. 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.
67. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.
73. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.

 doi:10.20944/preprints202510.1967.v1

80. 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.
81. 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.
82. 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).
83. 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.
84. 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.
85. 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.
86. 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.
87. 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).
88. 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.
89. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, A tree-based neural network model for biomedical event trigger detection, Information Sciences, 512, 175
90. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, Dispatched attention with multi-task learning for nested mention recognition, Information Sciences, 513, 241
91. 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
92. 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.
93. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
94. 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.
95. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
96. 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.
97. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
98. 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.
99. 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.
100. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
101. 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.

 doi:10.20944/preprints202510.1967.v1

102. 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.
103. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
104. 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.
105. 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.
106. 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.
107. 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.
108. 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.
109. 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.
110. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
111. 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.
112. 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.
113. 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.
114. 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.
115. 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.
116. 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.
117. 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.
118. 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.
119. 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.
120. 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.
121. 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.

 doi:10.20944/preprints202510.1967.v1

122. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
123. 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.
124. 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.
125. 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.
126. 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.
127. 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.
128. 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.
129. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.