Augmented Vision-Language Models: A Systematic Review

Anthony C Davis *tony.davis@jhuapl.edu*
Johns Hopkins University

Burhan Sadiq *bsadiq1@jhu.edu*
Johns Hopkins University

Tianmin Shu *tianmin.shu@jhu.edu*
Johns Hopkins University

Chien-Ming Huang *chienming.huang@jhu.edu*
Johns Hopkins University

Reviewed on OpenReview: *https://openreview.net/forum?id=DFnPi77v6J*

Abstract

Recent advances in visual-language machine learning models have demonstrated exceptional ability to use natural language and understand visual scenes by training on large, unstructured datasets. However, this training paradigm cannot produce interpretable explanations for its outputs, requires retraining to integrate new information, is highly resource-intensive, and struggles with certain forms of logical reasoning. One promising solution involves integrating neural networks with external symbolic information systems, forming neural symbolic systems that can enhance reasoning and memory abilities. These neural symbolic systems provide more interpretable explanations to their outputs and the capacity to assimilate new information without extensive retraining. Utilizing powerful pre-trained Vision-Language Models (VLMs) as the core neural component, augmented by external systems, offers a pragmatic approach to realizing the benefits of neural-symbolic integration. This systematic literature review aims to categorize techniques through which visual-language understanding can be improved by interacting with external symbolic information systems.

1 Introduction

1.1 Motivation

Vision-Language Models (VLMs) represent a significant leap forward in artificial intelligence (AI), showing remarkable abilities to interpret complex visual scenes and generate coherent natural language descriptions, powering advancements in tasks such as visual question answering (VQA) (Alayrac et al., 2022) and image/video captioning (Radford et al., 2021). Trained on vast web-scale datasets, these models excel at mapping between visual inputs and textual concepts. However, this end-to-end training paradigm inherently limits their capabilities in several critical ways. VLMs produce outputs without clear justifications, making them difficult to trust or debug without specialized tools (Rudin et al., 2021; Stan et al., 2024). Integrating new factual knowledge or correcting errors typically requires resource-intensive retraining. Furthermore, despite their semantic understanding, VLMs often struggle with tasks that require precise logical deduction, mathematical calculation (for example, accurate object counting), verifiable factual recall of entities within an image, and complex spatial reasoning (Khajuria et al., 2024; Zhang et al., 2025b). These limitations hinder their deployment in high-stakes applications that require precision, reliability, and adaptability. The concept of augmenting VLMs with external information systems has evolved into several distinct paradigms that offer practical solutions to VLM limitations. These augmentation approaches can be broadly categorized by the type of external resource utilized and how it interfaces with the VLM.

Retrieval-based augmentation has emerged as one of the most widespread approaches, with Retrieval Augmented Generation (RAG) (Lewis et al., 2021) becoming a common paradigm in both research and commercial applications. These methods retrieve relevant information from external sources to provide context for the VLM's processing. The retrieval mechanisms vary considerably: dense vector-based retrieval uses learned embeddings to find semantically similar content, often employing pretrained encoders like CLIP or custom embedding models; traditional term-based methods like BM25 provide lexical matching capabilities; and structured retrieval from knowledge graphs enables access to explicitly encoded relationships and facts (See Figure 1 for an example of knowledge graph retrieval). While some approaches fine-tune the VLM jointly with the retriever to improve retrieval relevance and integration (Chen et al., 2022b; Rao et al., 2023; Yuan et al., 2023b), many implementations use frozen VLMs with off-the-shelf retrieval systems, demonstrating the flexibility of this augmentation strategy. The retrieved information can be integrated at different stages: as additional input context (prompt augmentation), during the model's reasoning process, or to validate and refine outputs.

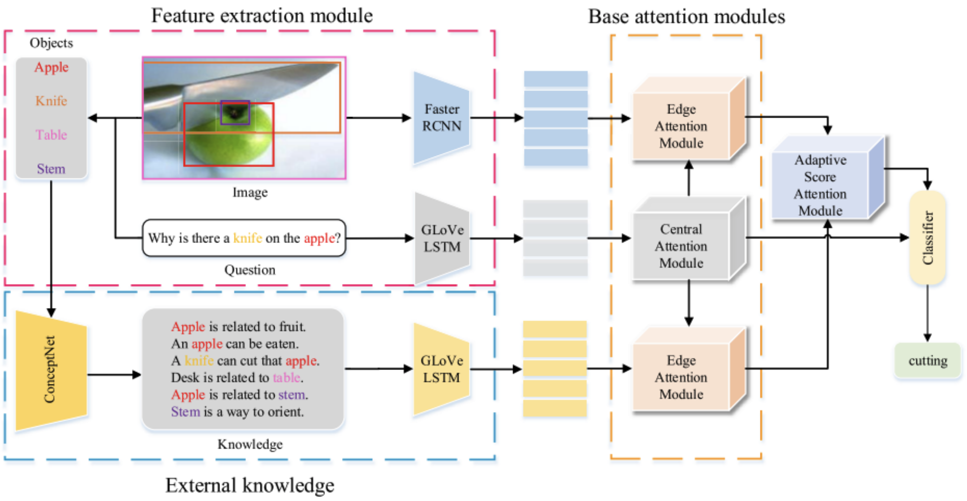
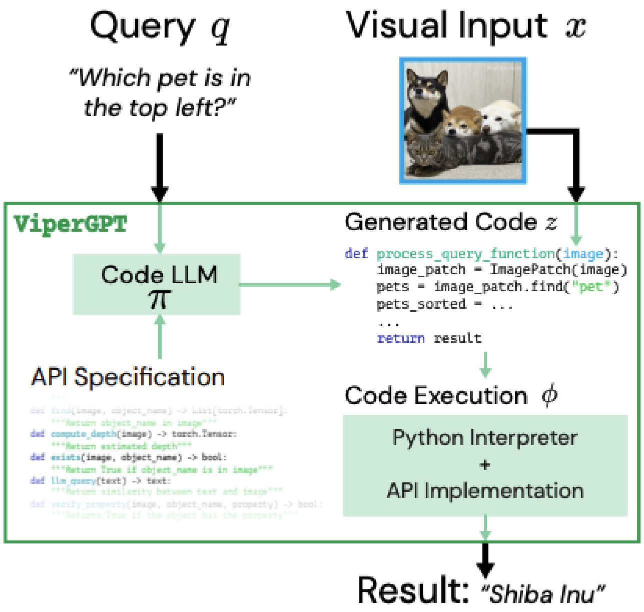


Figure 1: Architecture of the Knowledge-based Augmentation Network (KAN) by Zhang et al. (2020). The system extracts visual features via an object detector module and retrieves external knowledge from ConceptNet with labeled relationships and reliability scores.

Symbolic computation augmentation represents another major category where VLMs interface with external computational tools and reasoning engines. Program synthesis approaches enable VLMs to generate executable code (e.g., Python scripts, SQL queries, or domain-specific languages. See figure 2 for an example) that operates on structured representations or queries external systems, with the execution results informing the VLM's outputs. Symbolic reasoning engines such as logic solvers, planning systems, and specialized reasoning frameworks can be invoked to perform precise logical operations that complement the VLM's pattern recognition capabilities. The rapidly evolving paradigm of tool use (Schick et al., 2023; Qin et al., 2023) treats diverse external capabilities (calculators, APIs, specialized vision modules, web browsers) as tools that the VLM can dynamically invoke based on task requirements. Additionally, symbolic graph operations allow VLMs to manipulate structured representations like scene graphs or knowledge graphs through operations such as graph traversal, node matching, or relational reasoning, bridging perceptual understanding with structured symbolic manipulation.

These augmentation strategies offer a pragmatic path forward by building upon the sophisticated visual and language understanding already present in state-of-the-art VLMs, rather than replacing them. Through augmentation, a single VLM can be adapted to diverse tasks without needing to master every capability internally. Importantly, VLMs can be trained when and how to invoke these external resources, discovering effective strategies for combining their internal representations with external capabilities. This approach enables targeted mitigation of specific weaknesses, such as mathematical computation through calculators, factual accuracy through knowledge bases, and complex spatial reasoning through specialized geometric



```
def process_query_function(image):
    image_patch = ImagePatch(image)
    pets = image_patch.find("pet")
    pets_sorted = ...
    ...
    return result
```

```
    """
def find(image, object_name) -> List[torch.Tensor]:
    """Return object_name in image"""
def compute_depth(image) -> torch.Tensor:
    """Return estimated depth"""
def exists(image, object_name) -> bool:
    """Return True if object_name is in image"""
def llm_query(text) -> text:
    """Return similarity between text and image"""
def verify_property(image, object_name, property) -> bool:
    """Returns True if the object has the property"""
```

Figure 2: The ViperGPT framework by Surís et al. (2023). Given a visual query and an image/video input, ViperGPT uses a code-generation model (GPT-3 Codex) to generate Python code that composes various vision modules through an API. The generated program makes explicit function calls to vision capabilities (e.g., `find`, `compute_depth`, `count`) and uses Python's built-in logical and mathematical operators to reason about the results.

modules (Surís et al., 2023; Marino et al., 2020; Yi et al., 2018). This systematic review examines how these augmentation techniques are realized in practice, analyzing their implementation strategies, capabilities, and the specific VLM limitations they address.

1.2 Augmented Vision-Language Models: Definition and Scope

We define an Augmented Model as a system where external information or computational processes are actively integrated with a neural model's inference operations (before, during, or after its forward pass) to enhance its capabilities. This survey focuses specifically on inference-time augmentation, where external resources are accessed during model execution, not training-time data augmentation. However, the AVLMs we survey may still undergo finetuning to improve their ability to utilize these external resources effectively. This inference augmentation is distinct from prompting techniques like chain-of-thought or in-context examples (Wei et al., 2023; Zhao et al., 2023a), which elicit latent reasoning capabilities without accessing external data or tools.

A Vision-Language Model, for the purpose of this review, is a machine learning model that jointly processes and understands visual and textual modalities, either through generation tasks (e.g., VQA, captioning), alignment tasks (e.g., image-text retrieval, zero-shot classification), or both. This encompasses models that output natural language text as well as those that learn joint representations of vision and language.

Therefore, an AVLM is an VLM integrated with external symbolic information systems, APIs, databases, or other computational tools. An AVLM may involve modifications in VLM neural architecture, or it may involve pre- or post-processing of inputs or outputs of the VLM. Regardless, these integrations aims to overcome the inherent limitations of standalone VLMs and represent a particularly compelling implementation of the augmented neural system concept. A glossary of key terms in this survey is included in Appendix C.

1.3 Related Work and Knowledge Gap

The quest to enhance neural models, particularly in the vision-language domain, by incorporating external knowledge or symbolic reasoning has spurred significant research, reflected in several existing surveys. Reviews on knowledge-enhanced multimodal learning (Lymperaiou & Stamou, 2022; Zhao et al., 2023b; Wajid et al., 2023) investigate integrating factual knowledge, often via knowledge graphs or retrieval augmentation, to improve tasks like captioning and VQA. Concurrently, surveys exploring neuro-symbolic approaches (Aditya et al., 2019; Senior et al., 2023; Hitzler et al., 2022; Khan et al., 2024) examine the broader challenge of combining neural perception with symbolic reasoning, often focusing on graph neural networks, spatio-temporal logic, or commonsense knowledge integration for better scene understanding and reasoning. Specific areas like VQA have also been surveyed (Jamshed & Fraz, 2021; Mostafa et al., 2020), tracing the evolution towards models capable of more complex reasoning, sometimes touching upon the need for external knowledge or structured representations.

While these surveys provide valuable context by covering knowledge integration, neuro-symbolic methods, and advances in VQA reasoning, they do not specifically offer a systematic review focused on the augmentation of VLMs through interaction with diverse external symbolic systems and tools. Existing reviews often focus on specific knowledge types (e.g., knowledge graphs) or broader neuro-symbolic theory. There is a knowledge gap in understanding the landscape of techniques specifically designed to connect modern VLMs with external symbolic resources in a flexible, often learned manner (i.e., tool use). Particularly, there is a lack of systematic analysis regarding how these augmentation techniques address core VLM challenges, such as their noted difficulties with precise spatial reasoning (Wang et al., 2024c; Zhang et al., 2025b). Augmentation via external tools or information sources presents a potential pathway to compensate for such weaknesses by providing structured spatial information or enabling interactions with geometric reasoners, at least until VLM architectures intrinsically improve in these areas.

This systematic literature review aims to fill this gap by specifically categorizing and analyzing techniques where VLMs interact with external symbolic information systems or tools to enhance their vision-language understanding capabilities. We seek to provide a structured overview of how these augmentations are implemented, what types of external systems are used, and how they address the limitations of standard VLMs, with a particular interest in emerging tool-use paradigms and their application to challenging visual reasoning tasks.

2 Overview: Three Stages of Vision-Language Fusion

The papers surveyed demonstrate a variety of techniques for augmenting vision-language models with external symbolic information systems. The selection of these studies is the result of a systematic literature search conducted according to the PRISMA guidelines (Page et al., 2021), which involved querying academic databases with specific keywords and applying rigorous inclusion/exclusion criteria to identify relevant publications (see Appendix A). This process ensures that the surveyed works specifically target inference-time augmentation and filter out approaches like pure prompting or training-time knowledge integration. To structure this diverse landscape, we categorize the surveyed approaches based on three key characteristics:

- *When* the external interaction occurs relative to the VLM's processing pipeline. We distinguish between Early Fusion (integrating external data at the input stage, influencing initial representations), Middle Fusion (interfacing with external systems during the VLM's internal reasoning or generation steps), and Late Fusion (using the VLM's initial output to trigger external processing, validation, or refinement).

- *What* type of external information or computation is leveraged. This includes Retrieval (accessing pre-existing facts or knowledge from sources like knowledge graphs or text corpora) and Symbolic Computation (generating new information through logical deduction, program execution, or specialized computational tools), or a combination of both.

- *How* the fusion is specifically implemented, detailing the particular mechanisms used in each approach.

This review primarily organizes findings according to the temporal fusion stage (When), as this significantly impacts how external information influences the VLM. Within each temporal category (Early, Middle, Late), we further analyze the type of external interaction (What) and discuss notable implementation details (How). While some sophisticated methods may blend characteristics, this framework provides a structured lens for comparing the underlying principles, capabilities, and trade-offs of different augmentation techniques. The following sections elaborate on the findings for each category, referencing the detailed categorizations presented in the Appendix tables (Tables 2 through 5).

3 Early Fusion Methods

Early fusion methods augment the VLM by incorporating external information directly at the input stage, before the core VLM begins its internal processing. This is often the conceptually simplest approach, treating external information as additional context and potentially requiring no VLM architecture changes. Its main advantage is implementation simplicity, offering a direct way to provide context. However, it faces challenges related to the relevance and noise of retrieved information. For example, some implementations use generated image captions as retrieved context which may introduce information loss. The choice between simple prompt augmentation and more structured retrieval encoding depends on the desired level of integration and complexity tolerance. These methods primarily fall into retrieval-based or, less commonly, symbolic computation-based categories, as detailed in Appendix Table 2.

3.1 Retrieval-Based Early Fusion

The most common early fusion strategy involves retrieving relevant information from external sources and providing it alongside the primary visual and textual inputs. A primary technique is **Prompt Augmentation**, where retrieved textual context is directly appended to the input prompt, exemplified by Retrieval Augmented Generation (RAG) (Lewis et al., 2021). This retrieved text can originate from various sources. Text/Fact Retrieval draws information from text corpora or knowledge graphs (KGs), using approaches ranging from pre-trained encoders like CLIP without further training to fine-tuning the retriever, possibly jointly with the VLM, for better relevance (see Table 2, column "Retrieval FT"). See Figure 3 for an example of text retrieval using a pretrained vision-language encoder. Reranking retrieved results is often employed to enhance quality (Qu et al., 2024; Liu et al., 2024; Wen et al., 2024). Retrieved KG triplets can also be formatted as text for the prompt (see Table 2, column "Prompt Augmentation"). An alternative form of prompt augmentation uses Image Caption Augmentation, where textual descriptions (captions, labels, Optical Character Recognition (OCR)) are first generated from the visual input, and this text is then used for retrieval or directly added to the prompt, with some methods jointly training the caption generator and retriever (see Table 2, column "Image Caption"). While simplifying the problem to text-based retrieval, this approach risks information loss during captioning.

Instead of appending raw text, another approach uses **Retrieval Encoders** to encode the retrieved information (e.g., KG subgraphs, text passages) into separate embedding vectors. These embeddings then condition the VLM, often through attention mechanisms (Yuan et al., 2023b; Weng et al., 2024; Chen et al., 2022a; Salemi et al., 2023a), Long Short Term Memory models (LSTMs) (Wu et al., 2016), or memory modules (Hu et al., 2022). This allows for a more structured integration of knowledge. Specifically, KG subgraphs can be encoded using Graph Neural Networks (GNNs) (see Table 2, column "Subgraph Enc") or fused with scene graphs (see Table 2, column "KG Conv"). Multimodal KGs can also provide richer representations (Jiang & Meng, 2023).

3.2 Symbolic Computation Early Fusion

Integrating the results of symbolic computations at the input stage is rare in the surveyed literature. The primary example identified (Potapov et al., 2019) involves transforming the visual input into a symbolic scene graph. This structured representation, potentially processed by an external symbolic reasoning engine like OpenCog, serves as input or conditioning for the VLM. This approach explicitly introduces symbolic structure early on but depends heavily on robust perception-to-symbol conversion modules.

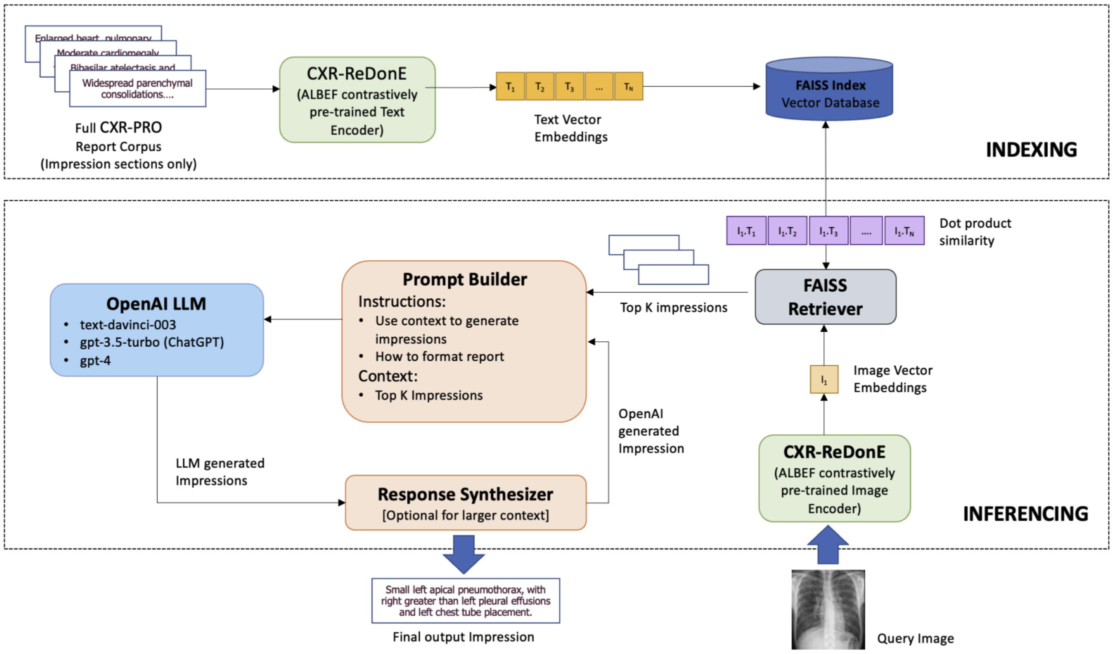

Enlarged heart, pulmonary
Moderate cardiomegaly
Bibasilar atelectasis and
Widespread parenchymal consolidations....
Full CXR-PRO Report Corpus (Impression sections only)
CXR-ReDonE (ALBEF contrastively pre-trained Text Encoder)
T_1 T_2 T_3 ... T_N
Text Vector Embeddings
FAISS Index Vector Database
INDEXING
$I_1.T_1$ $I_1.T_2$ $I_1.T_3$ .... $I_1.T_N$
Dot product similarity
FAISS Retriever
Top K impressions
Prompt Builder
Instructions:
• Use context to generate impressions
• How to format report
Context:
• Top K Impressions
OpenAI LLM
• text-davinci-003
• gpt-3.5-turbo (ChatGPT)
• gpt-4
LLM generated Impressions
Response Synthesizer [Optional for larger context]
OpenAI generated Impression
I_1
Image Vector Embeddings
CXR-ReDonE (ALBEF contrastively pre-trained Image Encoder)
INFERENCING
Small left apical pneumothorax, with right greater than left pleural effusions and left chest tube placement.
Final output Impression
Query Image

Figure 3: Architecture for Retrieval Augmented Chest X-Ray Report Generation by Ranjit et al. (2023). Text embeddings from radiology impressions are indexed in a vector database. For an input X-ray image, its embedding, generated by a contrastively pretrained vision-language encoder (CXR-ReDonE), is used to retrieve the most similar text (impressions or sentences) from the database. This retrieved text then forms the context for a prompt, along with specific instructions, which is fed to an LLM (e.g., OpenAI GPT models) to generate the final radiology report impression. This process is illustrated for both indexing and inferencing stages.

4 Middle Fusion Methods

Middle fusion techniques integrate external information or symbolic computation *during* the VLM's inference stage, allowing interaction with the model's intermediate representations before the final response is generated. Intermediate representations can mean either token-based such as in autoregressive generation, or embedding-based as in dense retrieval techniques. This enables more dynamic and potentially iterative integration compared to early fusion, where external data influences internal processing, reasoning steps, or feature refinement. By allowing external information and symbolic processes to interact with the VLM's internal state, these methods enable context-aware reasoning and iterative refinement. This often involves more complex architectures and training but holds promise for leveraging both neural pattern recognition and symbolic manipulation more effectively. The rise of tool use and agent-based frameworks within this category points towards VLMs acting less as monolithic predictors and more as components in larger reasoning systems, echoing paradigms like Kahneman's System 1 (neural intuition) and System 2 (deliberate symbolic reasoning) (Kahneman, 2011; Booch et al., 2020). These methods, categorized in Appendix Table 3, often involve feedback loops or specialized modules operating alongside main VLM components.

4.1 Retrieval-Based Middle Fusion

These methods retrieve external information based on intermediate VLM states and fuse it back into the ongoing computation. One approach is **Dense Retrieval**, which uses dense vector similarity between intermediate VLM representations and a knowledge corpus to find relevant information (often images or text) that is then fused back into the model's layers, typically via attention (Wang et al., 2022b; Lin et al., 2023b; Jia et al., 2023). Another major approach leverages **Graph-Based Retrieval**, primarily using KGs. This includes methods where intermediate visual or textual features trigger **KG Querying**; the retrieved subgraphs or facts are processed (often with GNNs) and fused with VLM representations, sometimes after extracting visual subgraphs first (see Table 3, column "KG Prompt Augmentation"). Figure 4 illustrates a middle fusion approach that constructs coupled scene and concept graphs, using shared entities as mediums for cross-modal knowledge exchange. Other graph-based methods use **Similarity Measures** between internal VLM representations and KG elements to guide reasoning or weighting, rather than directly injecting KG structure (Wu et al., 2024a; Chae & Kim, 2022; Li et al., 2019; ming Xian et al., 2023; Marino et al., 2020). A significant group focuses on **Concept/Scene Graph Fusion**, explicitly combining internally generated scene graphs with external concept graphs (e.g., from ConceptNet (Speer et al., 2018)), often using GNNs on the combined graphs (see Table 3, column "Concept/Scene Fusion"). More complex structures like **Multimodal KGs** (MMKGs) (Xi et al., 2024; Shi et al., 2022; Santiesteban et al., 2024; Ouyang et al., 2024; Liu et al., 2021) or **Hypergraphs** (Heo et al., 2022; Wang et al., 2024b) are also integrated using specialized graph networks. Finally, **Reinforcement Learning** (RL) can be used to learn policies for querying or integrating external knowledge based on the current state (Bougie et al., 2018).

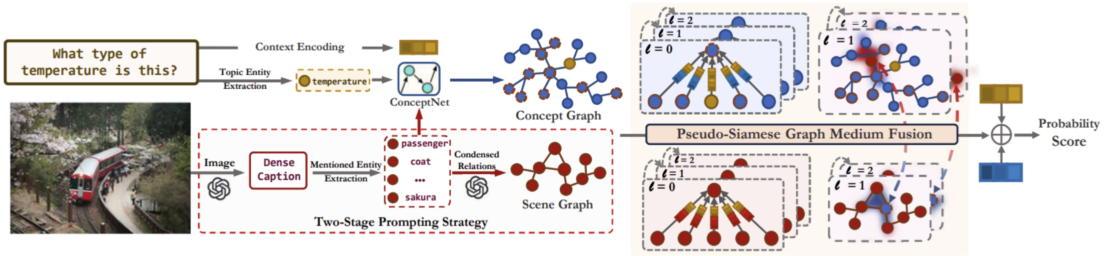


Figure 4: The MAIL (Modality-Aware Integration with LLMs) framework by Dong et al. (2024). The system employs a two-stage prompting strategy: first generating a dense caption through a visual LLM, then constructing a scene graph by extracting spatial and object features as triples (e.g., *(person, wearing, coat)*). These scene graph entities are linked with external knowledge from ConceptNet to form a coupled concept graph containing real-world facts (e.g., *(coat, used_for, warn)*). A pseudo-siamese graph medium fusion module processes both graphs through parallel graph attention networks with different weights, using shared mentioned entities as mediums to enable cross-modal exchange while preserving intra-modal information.

4.2 Symbolic Computation Middle Fusion

These methods incorporate symbolic reasoning, calculations, or tool use within the VLM's processing pipeline. One key technique is **Program Synthesis**, where the VLM generates intermediate programs (e.g., functional programs, Python code) operating on symbolic input representations or querying external tools; the execution result influences subsequent VLM processing (Zhang et al., 2022c; 2023e; Hu et al., 2023b), (Shirai et al., 2023, see Figure 5), (Zhang et al., 2023b; Li et al., 2021; Mishra et al., 2024; Xue et al., 2024). Another approach involves integrating **Symbolic Logic Engines**, translating intermediate VLM representations into facts or queries processed by engines like differentiable first-order logic (Zhang et al., 2025a), Answer Set Programming (ASP) (Riley & Sridharan, 2019; Mitchener et al., 2021), Description Logic (Tsatsou et al., 2021), planning domain definition languages (PDDL) (Zhang et al., 2022b; 2023d), temporal logic (Choi et al., 2024), specialized neurosymbolic languages like Scallop (Li et al., 2023d; Huang et al., 2021), or embedding propositional logic operations (Li et al., 2023c). **Vector Symbolic Architectures (VSAs)** represent symbols and perform operations using high-dimensional vectors within the neural architecture (Montone et al., 2017; Kovalev et al., 2021). Some methods perform **Symbolic Graph Operations** directly on graph representations (scene graphs, KGs) during processing, like guided walks or routing (Li et al., 2022c; Liang et al., 2020; Wu et al., 2023; Zhao, 2015; Yang et al., 2020; Zhang et al., 2023f; Hudson & Manning, 2019; Cao et al., 2021). Increasingly popular is **Tool Use**, where the VLM dynamically calls external tools (calculators, APIs, vision algorithms, drawing tools) based on its intermediate state, integrating the tool's output (see Table 3, column "Tool Use"). Lastly, **Self Play** involves using the VLM within a simulated environment where it interacts, uses tools (potentially itself), and learns from feedback (Misiunas et al., 2024).

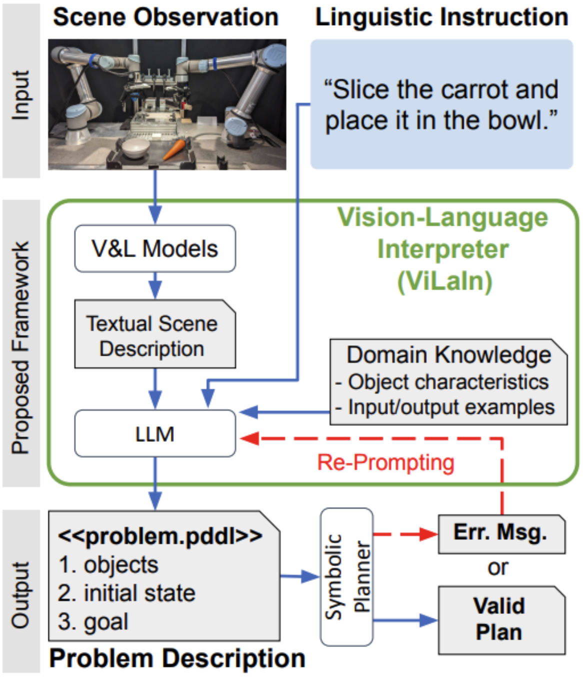


Figure 5: Overview of the ViLaIn approach for VLM planning of robotic actions Shirai et al. (2023). The vision-language interpreter (ViLaIn) generates a problem description from a linguistic instruction and scene observation. The symbolic planner finds an optimal plan from the generated problem description.

4.3 Combined Retrieval and Symbolic Computation Middle Fusion

These advanced methods integrate both retrieval and symbolic computation during the forward pass. Many employ **Agent** architectures where the VLM acts as a controller, deciding when to retrieve information and when to use symbolic tools (including sub-agents or code execution) to achieve a goal (see Table 3, column "Agents"). **Other Approaches** combine retrieval (e.g., from ontologies, KGs) with symbolic reasoning (e.g., probabilistic logic, program synthesis, graph walks, concept binding) in bespoke ways for specific tasks like embodied QA, riddle solving, or rumor detection (Besbes et al., 2015; Aditya et al., 2016; Aditya, 2017; Aditya & Baral, 2016; Tan et al., 2021; Liu et al., 2023a; Stammer et al., 2024; Vatashsky & Ullman, 2018; Gao et al., 2023c; 2024).

5 Late Fusion Methods

Late fusion methods apply external information retrieval or symbolic computation *after* the VLM has generated an initial output. This external step typically serves to validate, refine, explain, or augment the VLM's output using structured knowledge or precise tools. Late fusion provides a powerful mechanism for verification, refinement, and explanation by applying structured knowledge or precise computations to the VLM's generated output. It leverages the VLM's ability to produce a plausible initial response, which then guides a more targeted external process. This approach is particularly well-suited for enhancing reliability and interpretability, as symbolic steps can act as explicit checks or provide traceable reasoning paths. The main dependency is the quality of the initial VLM output; if it is too vague or incorrect, the subsequent external process may be misguided. These techniques are cataloged in Appendix Table 4.

5.1 Retrieval-Based Late Fusion

Here, the VLM's output triggers a targeted retrieval query. In **Dense Retrieval**, the initial VLM output (e.g., answer, rationale) queries a dense retrieval system. The retrieved information (text, facts) is then used to refine the output or provide supporting evidence (Song et al., 2022a;b; Shi et al., 2024). Alternatively, using **Knowledge Graph Retrieval**, the VLM's output (e.g., generated caption, predicted relationships) queries a KG. Retrieved facts or subgraphs refine the output, for instance, by adjusting probabilities or improving relationship predictions (Gao et al., 2022b; Huang et al., 2020; Xiao & Fu, 2022).

5.2 Symbolic Computation Late Fusion

This involves applying symbolic tools or logic engines to the VLM's output. **Program Synthesis** generates programs based on the VLM's output for analysis, validation, or transformation. Examples include generating Python code to verify VQA answers via vision APIs, treating symbolic programs as latent variables, or generating Structured Query Language (SQL) queries from the output (see Table 4, column "Program Synth"). The influential Neural-Symbolic VQA (NS-VQA) approach (Yi et al., 2018), executing programs on scene representations post-prediction, is often adapted. **Symbolic Engines** feed the VLM's output (or derived symbolic representations) into formal logic engines (e.g., Prolog, ASP, Probabilistic Soft Logic) for consistency checking, inference, or validation (Sethuraman et al., 2021; Aditya et al., 2018; Eiter et al., 2022; 2021; Cunnington et al., 2024), or use PDDL for planning (Xu et al., 2022). **Tool Use** involves calling external tools or APIs based on the VLM's output for specialized functions, verification, or generating structured data (Yuan et al., 2023a; Cesista et al., 2024; Cesista, 2024; Zhang, 2023). **Symbolic Graph Operations** perform manipulations on graph representations derived from the VLM's output, such as reasoning over action chains or graph traversals (Li et al., 2023a; Zhan et al., 2021; Saqur & Narasimhan, 2020; Johnston et al., 2023). **Other Approaches** include applying symbolic solvers to latent representations (Singh, 2018), using VLM output confidence to trigger human interaction or further symbolic checks (Bao et al., 2023), or updating conversational memory based on the response (Verheyen et al., 2023).

5.3 Combined Retrieval and Symbolic Computation Late Fusion

These methods combine both retrieval and symbolic computation after the initial VLM output. Typically, the VLM output is parsed into a logical form, relevant domain knowledge (facts or programs) is retrieved, and a symbolic reasoner (e.g., probabilistic logic, ASP) derives the final answer (Sachan, 2020; Basu et al., 2020). The AQuA framework (Basu et al., 2020) is depicted in Figure 6.

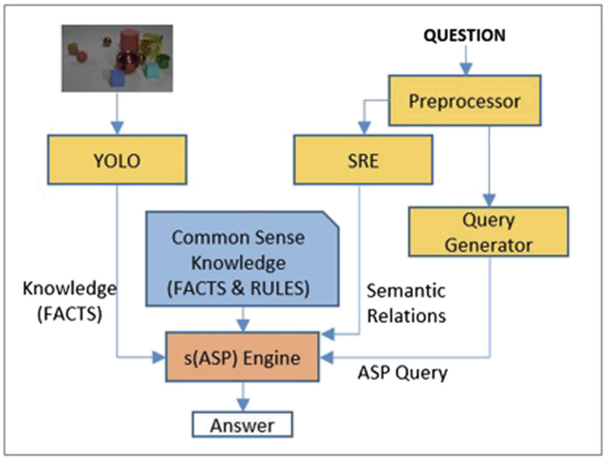


Figure 6: The architecture of the AQuA framework Basu et al. (2020). It consists of five main modules: (i) YOLO for object detection and feature extraction, (ii) a Preprocessor for the natural language question, (iii) a Semantic Relation Extractor (SRE), (iv) a Query Generator based on semantic analysis, and (v) a retrieval based Commonsense Knowledge module leveraging. The system utilizes an ASP engine for symbolic reasoning.

6 Discussion

The studies reviewed in this paper underscore the growing success of incorporating external symbolic information into vision-language models across different fusion paradigms. Here we discuss key observations, challenges, and potential directions for future research stemming from these findings.

6.1 Domain-Specific Limitations and Augmentation Solutions

The integration of external symbolic systems in vision-language models addresses fundamental perceptual and reasoning limitations that distinguish visual understanding from purely textual tasks. We organize domain-centric limitations of VLMs along five phenomena that recur across the surveyed studies: spatial, temporal, knowledge grounding, physical commonsense, and action/embodiment (STKPA). Each category of limitation has driven the development of specific augmentation patterns that leverage the complementary strengths of neural perception and symbolic computation. Orthogonal to this axis is how external structure is injected, via retrieval, program/tool execution, or logic/constraint checking, and when it is injected (early, middle, or late fusion; see Sections 3– 5). The matrix in Table 1 summarizes typical augmentation patterns observed for each phenomenon. Across the 264 papers in our corpus, each work is aligned with at least one STKPA category, as shown in Tables 2 to 4. We will discuss each domain and highlight some example augmentation solutions and their associated datasets.

6.1.1 Spatial numeracy and geometry

VLMs struggle with tasks requiring exact measurements or counts within visual scenes, with error rates exceeding 30% on simple counting tasks (Wang et al., 2024c). This limitation stems from the continuous nature of visual features in neural representations conflicting with discrete spatial reasoning requirements.

Effective augmentation systems externalize discrete structure through two primary families of approaches. First, program synthesis methods compile questions to programs that operate on structured scene representations. Neural-Symbolic VQA (Yi et al., 2018) achieved 99.8% accuracy on CLEVR through explicit program execution over scene graph representations, effectively saturating the benchmark with a 15+ percentage point improvement over end-to-end neural baselines. Second, visual API orchestration methods like ViperGPT (Surís et al., 2023) generate Python code that composes heterogeneous computer vision modules such as detectors, object/semantic segmenters, or depth estimators, each providing complementary symbolic information about the visual scene. This compositional approach allows VLMs to leverage specialized perception modules' strengths while maintaining flexibility through learned orchestration policies.

Spatial reasoning benchmarks have been fundamental in evaluating AVLMs' ability to understand geometric relationships and counting (Table 5). The CLEVR family of datasets (Johnson et al., 2016) provides synthetic but precisely controlled evaluation of spatial reasoning, with extensions like Super-CLEVR (Li et al., 2022e) testing domain robustness and CLEVR-POC (Abraham et al., 2024) introducing partial observability challenges. Scene graph datasets including Visual Genome (Krishna et al., 2016) and VCD (Shen et al., 2024) evaluate models on real-world spatial relationships between objects. More recent benchmarks like SOK-Bench (Wang et al., 2024a) combine spatial and knowledge requirements, while specialized datasets test whether VLMs can reason about times and locations (Zhang et al., 2023a) or answer complex visual information-seeking questions (Chen et al., 2023).

6.1.2 Temporal and causal ordering

In video understanding, VLMs frequently misinterpret temporal sequences and causal relationships, with accuracy drops of 20-40% on tasks requiring temporal ordering compared to static image reasoning (Yi et al., 2019). The absence of explicit temporal reasoning mechanisms in standard architectures necessitates augmentation with structured temporal representations.

Middle-fusion pipelines address this by tracking entities and actions through specialized tools (detection, tracking, pose estimation, optical flow) to construct event sequences that programs or planners can reason over. Logic-based approaches apply temporal operators and constraint checks to predicted timelines, with

	Retrieval	**Symbolic Computation**
Spatial	KG fact retrieval to bias relations/attributes (Marino et al., 2020; Li et al., 2020); dense region/sentence retrieval for spatial cues (Chen et al., 2022b; Lin et al., 2022); graph-based retrieval of related objects/relations (Wang et al., 2022c; Zhu et al., 2020b)	Executable visual programs over scene graphs (NS-VQA) (Yi et al., 2018); Python tool orchestration for counting/geometry (detector/segmenter/depth) (Surís et al., 2023); visual program distillation (Hu et al., 2023b); ASP/logic verification of spatial predicates (Basu et al., 2020; Eiter et al., 2022; Sethuraman et al., 2021); probabilistic neural–symbolic constraints (Vedantam et al., 2019)
Temporal	Knowledge-guided caption/entity retrieval for event context (Xi et al., 2024; Hou et al., 2020); ontology lookup of task/event relations (Jiang et al., 2020)	LLM-orchestrated video tools (tracking, OCR/ASR, shot segmentation) (Fan et al., 2024); program traces over events (Zhang et al., 2023e); symbolic activity reasoning with tool-augmented execution (Wu et al., 2023); temporal logic/order constraints and ASP checks (Choi et al., 2024; Eiter et al., 2022)
Knowledge grounding	Outside-knowledge retrieval (text/KG, reranking) for OK-VQA (Marino et al., 2019; Chen et al., 2022b; Wen et al., 2024); multi-source multimodal retrieval (entities, captions, KG) (Zhu et al., 2020b; Salemi et al., 2023a; Gui et al., 2021)	Code/program generation to query KGs/APIs or compose operators (Subramanian et al., 2023; Zhang, 2023); agentic tool-use for information seeking (Hu et al., 2023c); PSL/formal-logic consistency and entailment (Aditya et al., 2016; Sethuraman et al., 2021; Eiter et al., 2022)
Physical commonsense	Affordance/functional KB retrieval (e.g., ConceptNet) (Speer et al., 2018; Marino et al., 2020); knowledge-aided captioning (Huang et al., 2020)	Differentiable/explicit physics for feasibility (Huang et al., 2023); vision-tool composition for metric/-geometry checks (Surís et al., 2023); 3D symbolic grounding of objects/relations (Hsu et al., 2023); rule/constraint checking for stability, containment, and causal function (Aditya et al., 2018; Vedantam et al., 2019; Zhao, 2015)
Action / embodiment	External maps/knowledge for navigation and planning (Li et al., 2022d; Ni et al., 2023); VL-action models transferring web knowledge for control (Brohan et al., 2023); robotic manipulation knowledge bases (Gao et al., 2023b)	Planner interfaces (PDDL/ASP) invoked from VLM outputs (Shirai et al., 2023; Zhang et al., 2022b); plan–execute–observe loops grounded by VLMs (Zhang et al., 2023b; Gao et al., 2023b); tool documentation for zero-shot tool use (Hsieh et al., 2023); planner validity/safety constraints and neuro-symbolic agent frameworks (Xu et al., 2022; Cunnington et al., 2024)

Table 1: Reasoning domains and associated augmentation mechanisms with example implementations

temporal logic systems (Choi et al., 2024) showing 15-25% improvements on temporal ordering tasks. Retrieval of script knowledge or prototypical event progressions biases hypotheses toward plausible sequences, while late-fusion verifiers test temporal feasibility against these priors. These augmented approaches separate perceptual feature extraction from logical temporal inference, allowing each component to operate in its optimal representational space.

Temporal reasoning evaluation is primarily anchored by CLEVRER (Yi et al., 2019) (Table 5), which extends spatial reasoning into the temporal domain through collision events and causal chains in synthetic videos. The Compositional 4D dataset (Wang et al., 2024d) further challenges models with four-dimensional scene understanding requiring both temporal and physical reasoning. While fewer dedicated temporal datasets exist compared to other domains, temporal reasoning is often implicitly tested in embodied agent benchmarks and video understanding tasks that require tracking state changes over time.

6.1.3 Knowledge-intensive grounding

The challenge of linking language to specific visual entities requiring external knowledge (identifying landmarks, recognizing famous individuals, understanding fine-grained categories) represents a fundamental limitation where learned pattern matching proves insufficient due to sparse training data for highly specific instances (Kalai et al., 2025).

Multimodal knowledge graph navigation addresses this through explicit retrieval and graph-based reasoning. VQA-GNN (Wang et al., 2022c) demonstrated 4.6% improvement on GQA by coupling internally generated scene graphs with external concept graphs through graph neural networks that enable message passing between visual features and symbolic knowledge. KAT (Gui et al., 2021) achieved 53.1% on OK-VQA, a 5.1% improvement over PICa which relied solely on GPT-3's parametric knowledge (Yang et al., 2022). The integration of structured knowledge graphs like ConceptNet enables more accurate entity recognition through explicit retrieval of facts about identity, function, and long-tail categories, rather than relying solely on learned pattern matching.

Knowledge-based VQA represents the largest category of benchmarks for AVLMs (Table 5). The progression from general VQA (Agrawal et al., 2015) to fact-based FVQA (Wang et al., 2016; Lin et al., 2023c) and knowledge-aware KVQA (Shah et al., 2019) reflects increasing demands for external knowledge. OK-VQA (Marino et al., 2019) and its successor (Reichman et al., 2023) have become standard benchmarks requiring knowledge beyond visual content. Specialized variants target encyclopedic knowledge (Mensink et al., 2023), cultural domains (Agarwal et al., 2024), named entities (Lerner et al., 2022; Qiu et al., 2024), and synthetic knowledge generation (Su et al., 2024). Advanced reasoning benchmarks like A-OKVQA (Schwenk et al., 2022) and Visual Riddles (Bitton-Guetta et al., 2024) combine knowledge requirements with complex compositional reasoning, while CRIC (Gao et al., 2019) tests compositional reasoning on vision and commonsense.

6.1.4 Physical commonsense and affordances

VLMs often lack robust understanding of physical properties and object affordances visible in scenes, commonly misunderstanding material properties, physical constraints, or predicted behavior under manipulation (Yi et al., 2019; Wang et al., 2024d). This gap necessitates specialized bridging mechanisms that translate between continuous visual features and discrete symbolic representations.

Two augmentation routes have proven effective. First, symbolic vector or graph representations bridge continuous perception and discrete physical predicates. Vector Symbolic Architectures (Montone et al., 2017; Kovalev et al., 2021) represent spatial relationships as high-dimensional vectors supporting symbolic operations like binding and unbinding within the neural architecture itself. Graph neural networks operating on coupled scene-concept graphs (Dong et al., 2024; Song et al., 2023) enable cross-modal knowledge exchange, with ConceptNet integration showing 8-12% improvements on tasks requiring physical commonsense reasoning. Second, tool-enabled pipelines call physics or geometry modules (depth estimation, meshing, simple simulators) to test feasibility or measure quantities before answering. The structured nature of external knowledge allows models to access explicit relationships like "glass IsA fragile" or "liquid HasProperty flows_downward" that may not be reliably encoded in learned parameters.

Physical reasoning evaluation is less explicitly represented in current benchmarks but appears implicitly across multiple datasets (Table 5). The Compositional 4D dataset (Wang et al., 2024d) specifically targets understanding of physical dynamics and material properties across time. Visual Commonsense Reasoning (VCR) (Zellers et al., 2018) requires understanding of physical plausibility and social dynamics, while explainable reasoning benchmarks (Cao et al., 2019) often involve physical world knowledge. The relative scarcity of dedicated physical commonsense benchmarks represents a gap in current evaluation frameworks, despite this being a critical limitation of VLMs.

6.1.5 Action and embodiment

In robotics and embodied AI applications, the challenge of translating visual scenes and natural language instructions into executable action sequences requires precision and verifiability that standalone VLMs cannot provide. Augmentation patterns treat VLMs as interpreters that generate symbolic action specifications for downstream execution.

Methods like ViLaIn (Shirai et al., 2023) generate PDDL specifications that symbolic planners execute with guaranteed optimality properties. Systems employing Answer Set Programming (Zhang et al., 2022b) or other formal planning languages leverage logical consistency checking. Many approaches implement plan–execute–observe–reflect loops that re-plan on mismatch, with retrieval supplying action/operator libraries and environment maps. Tool interfaces expose planners and low-level skills through well-defined APIs, while safety and validity are enforced through logic layers checking preconditions, effects, and constraints. This vision-to-symbol translation separates flexible perception from rigorous planning, enabling reliable task execution in physical environments.

Embodied AI benchmarks evaluate AVLMs' ability to translate understanding into executable actions (Table 5). Web-based environments like WebArena (Zhou et al., 2023) and ScreenAgent (Niu et al., 2024) test agents on realistic computer control tasks, while Spider2-V (Cao et al., 2024) focuses on data science workflows. Robotic manipulation benchmarks (Gao et al., 2023b) evaluate physical grounding and tool use in real-world settings. These agent-focused evaluations represent a shift from passive question-answering to active, goal-directed interaction with environments, requiring integration of perception, reasoning, and action execution.

6.2 Common Architectural Patterns in AVLMs

While the surveyed papers employ diverse implementation strategies, our analysis reveals three fundamental architectural patterns that have emerged as dominant paradigms for augmenting vision-language models. These patterns represent crystallized design solutions that address specific computational challenges in vision-language understanding. These patterns distribute differently across fusion stages, with retrieval-based approaches primarily operating at the input stage, while symbolic computation patterns span both middle and late fusion paradigms.

6.2.1 The Retrieval-Integration Pipeline Pattern

The retrieval-reasoning pipeline has evolved from simple keyword matching to sophisticated neural retrieval systems. The evolution of this pattern over the past decade (2016-2024) reflects increasing sophistication in how external knowledge is indexed, accessed and integrated in the following ways:

- **Early Simple Retrieval (2016-2018)**: Direct keyword matching or TF-IDF based retrieval from knowledge bases, with retrieved facts concatenated to prompts (Wang et al., 2015; Narasimhan & Schwing, 2018)
- **Learned Dense Retrieval (2019-2020)**: Introduction of learned embeddings using pretrained encoders like CLIP, enabling semantic similarity search (Li et al., 2020; Zhang et al., 2020)
- **Joint VLM-Retriever Training (2021-2022)**: End-to-end training of retriever and VLM components, optimizing retrieval for downstream task performance (Chen et al., 2022b; Gui et al., 2021)
- **Tool-Augmented Retrieval (2023-2024)**: Retrieval as one tool among many, with VLMs learning when and what to retrieve (Yan & Xie, 2024; Hao et al., 2024b)

Critical design decisions in this pattern include the choice of embedding space, where CLIP-based retrieval demonstrates strong performance on vision-language tasks (Gui et al., 2021), retrieval granularity with sentence-level retrieval showing advantages for factoid questions (Chen et al., 2022b), and integration mechanism where attention-based fusion consistently outperforms simple concatenation (Yuan et al., 2023b; Weng et al., 2024).

6.2.2 The Intermediary Program Pattern

This pattern treats program synthesis as a bridge between neural perception and symbolic computation, with the VLM generating executable code that operates on structured representations or invokes external tools. The generated programs serve as interpretable reasoning chains that can be verified, debugged, and modified. Importantly, we distinguish between domain-specific program synthesis and general-purpose code generation, as they represent different points on the expressiveness-tractability spectrum.

This pattern's evolution has followed a similar progression to the retrieval-reasoning pipeline, moving from static patterns to dynamic, learned behaviors:

- **Static Logic on VLM Outputs (2018-2019)**: Early approaches applied predefined symbolic logic engines (ASP, Prolog) to VLM-extracted scene graphs, requiring manual rule specification (Aditya et al., 2018; Riley & Sridharan, 2019)

- **Domain-Specific Program Synthesis (2020-2022)**: VLMs learned to generate programs in constrained domain-specific languages (DSLs) with guaranteed executability. NS-VQA (Yi et al., 2018) synthesizes functional programs over a fixed set of visual primitives, while Zhang et al. (2025a) generates first-order logic expressions

- **General-Purpose Code Generation (2022-2023)**: Shift to generating Python or SQL code with broader expressiveness but without execution guarantees. ViperGPT (Surís et al., 2023) generates unrestricted Python code composing vision APIs, while Gupta & Kembhavi (2022) produces Python programs for visual reasoning

- **Dynamic Tool Orchestration (2023-2024)**: VLMs as orchestrators selecting and composing heterogeneous tools including APIs, specialized models, and code execution environments (Hu et al., 2023c; Wu et al., 2024b; Liu et al., 2023b)

The key architectural components include: (1) Program specification language - DSLs offer tractability with limited expressiveness (NS-VQA's 20 primitives achieve 99.8% on CLEVR (Yi et al., 2018)), while general-purpose languages enable broader capabilities but require error handling; (2) Execution environment - ranging from symbolic executors for DSLs to sandboxed Python interpreters with vision library access; (3) Error handling mechanisms - recent approaches like Mishra et al. (2024) incorporate execution feedback for iterative refinement. ViperGPT demonstrates the general-purpose approach's flexibility, achieving strong performance across diverse visual reasoning tasks through unrestricted Python generation (Surís et al., 2023).

6.2.3 The Graph Fusion Pattern

Graph-mediated fusion explicitly models relationships between visual elements and external knowledge through graph structures, enabling structured reasoning over combined perceptual and symbolic information. This pattern typically involves three stages: graph construction (from visual input and/or external knowledge), graph alignment (connecting visual and symbolic graphs), and graph neural network processing for joint reasoning.

Key Design Variations:

- **Scene-Concept Graph Coupling**: Methods like VQA-GNN (Wang et al., 2022c) and MAIL (Dong et al., 2024) construct parallel scene graphs (from images) and concept graphs (from knowledge bases), using shared entities as bridges. Graph attention networks enable cross-modal message passing while preserving intra-modal structure, with VQA-GNN showing 4.6% improvement on GQA through this approach (Wang et al., 2022c)

- **Multimodal Knowledge Graph Integration**: Approaches incorporating MMKGs (Xi et al., 2024; Liu et al., 2021) where nodes contain both visual exemplars and textual descriptions, enabling richer cross-modal grounding through joint embedding spaces

- **Dynamic Graph Construction**: Methods that construct query-specific subgraphs rather than using fixed graph structures, with Li et al. (2022b) demonstrating improved efficiency through adaptive graph pruning.

Critical implementation choices include graph representation (heterogeneous vs. homogeneous nodes), alignment mechanisms (entity matching vs. learned attention), and message passing strategies (synchronous vs. asynchronous updates). The graph-mediated pattern provides a principled way to preserve structural information while enabling neural reasoning, making it particularly effective for tasks requiring explicit relational understanding between visual elements and conceptual knowledge.

6.2.4 Cross-Pattern Observations

Several key insights emerge from analyzing these patterns:

1. **Temporal Distribution**: Retrieval-based approaches concentrate in early fusion (where they modify inputs), while program and graph-based patterns distribute across middle and late fusion stages (where they can interact with intermediate representations or refine outputs).

2. **Complementarity**: The most successful recent systems combine patterns - for example, using retrieval to gather relevant facts, then applying program synthesis for precise computation over retrieved information (Castrejon et al., 2024; Lu et al., 2023).

3. **Interpretability-Simplicity Trade-off**: Program-based approaches offer highest interpretability through executable traces but require more specialized, task-specific training. Graph-based methods provide moderate interpretability through explicit relational structure. Pure retrieval offers limited interpretability but is simplest to implement.

4. **Scalability Characteristics**: Retrieval scales well with knowledge base size using approximate nearest neighbor search (Chen et al., 2022b), program synthesis complexity grows with the size of the DSL or API set (Yi et al., 2018), while graph methods face quadratic complexity in number of nodes, requiring approximation techniques for large graphs (Wang et al., 2022c).

These architectural patterns provide a technical foundation for designing augmented vision-language systems, with the choice of pattern depending on task requirements for accuracy, interpretability, and computational efficiency.

6.3 Future Directions for Vision-Centric Augmentation Research

Cross-Modal Knowledge Integration Through Bidirectional Grounding: Rather than treating retrieved information as passive context, future research should explore augmentation techniques that leverage cross-modal correspondences more deeply. For example, retrieved knowledge about typical spatial layouts (e.g., "monitors are usually on desks") could constrain visual parsing, while visual evidence could trigger targeted knowledge retrieval. Multimodal knowledge graphs where nodes represent visual concepts with both image exemplars and textual descriptions (Jiang & Meng, 2023; Xi et al., 2024) could enable richer cross-modal reasoning through joint embedding spaces. This bidirectional interaction between symbolic knowledge and visual processing could improve both grounding accuracy and reasoning efficiency by ensuring consistency between what the model sees and what external knowledge suggests should be present.

Embodied Vision-Language Systems with Reinforcement Learning for Tool Use: The integration of VLMs with physical embodiment and interactive environments presents unique augmentation opportunities beyond one-shot prediction. Embodied systems can iteratively refine understanding through interaction: moving cameras to new viewpoints, manipulating objects to reveal hidden properties, or executing actions to test hypotheses about the physical world (Gao et al., 2023b). Recent advances in reinforcement learning for visual tool use, exemplified by approaches like VTool-R1 (Wu et al., 2025), demonstrate how RL can train VLMs to dynamically select and compose vision tools based on environmental feedback and task requirements. Future research should develop reward functions that balance exploration (trying new tool combinations) with exploitation (using known effective strategies), while incorporating human preferences to ensure safe and interpretable tool usage patterns. This includes learning when to request human intervention for ambiguous visual scenes or safety-critical decisions in autonomous systems.

Interpretable Visual Program Synthesis for Safety-Critical Applications: While current visual programming approaches demonstrate success on controlled benchmarks (Surís et al., 2023; Gupta & Kembhavi, 2022), deploying these systems in safety-critical domains requires advances that combine program synthesis with interpretability mechanisms. Future systems must handle partial observability through techniques like uncertainty-aware program generation that explicitly represents confidence in different execution paths. (Bao et al., 2023; Vedantam et al., 2019; Chae & Kim, 2022) The challenge of providing interpretability becomes paramount in high-stakes applications. Systems should produce visual reasoning chains that show which image regions were examined, what external tools were invoked, and how intermediate results combined to reach conclusions. This requires developing visualization techniques that render program execution traces overlaid on images with attention heatmaps, tool call annotations, and confidence scores. For medical imaging, autonomous driving, or industrial inspection, these interpretable execution traces serve dual purposes, enabling domain experts to verify correctness and providing auditable records for regulatory compliance. The key is balancing completeness (showing all reasoning steps for full transparency) with comprehensibility (avoiding overwhelming users with excessive detail through hierarchical or interactive visualization approaches).

Scalable Visual Program Libraries and Transfer Learning: Current visual programming approaches often require task-specific program templates or limited APIs. Future research should develop methods for building compositional program libraries that grow through experience, enabling VLMs to synthesize increasingly complex vision pipelines by combining learned subroutines. This includes meta-learning approaches that discover reusable visual reasoning patterns across tasks and transfer learning techniques that adapt programs from source domains (where supervision is abundant) to target domains (where it is scarce). The vision-language community would benefit from standardized APIs and benchmark tasks for visual program synthesis, analogous to how HuggingFace standardized model interfaces for NLP.

Adaptive Augmentation Based on Task Uncertainty: Rather than applying fixed augmentation strategies, future AVLMs should dynamically determine when and how to invoke external resources based on uncertainty estimation. For instance, a model confident in its counting ability for sparse scenes might bypass external tools, while requesting symbolic computation for cluttered environments. This adaptive approach requires developing calibrated uncertainty measures for different aspects of visual reasoning (spatial relationships, object recognition, attribute prediction) and learning policies that optimize the trade-off between accuracy gains and computational costs of augmentation.

7 Conclusion

Vision-Language Models have revolutionized AI's ability to connect vision and language, yet standalone models struggle with factual accuracy, complex reasoning, adaptability, and interpretability. This systematic review charted the landscape of Augmented Vision-Language Models (AVLMs), which overcome these limitations by integrating VLMs with external symbolic information systems and computational tools. We surveyed a diverse range of techniques, categorizing them by fusion timing (early, middle, late) and the nature of augmentation (retrieval, symbolic computation, combined), revealing a clear consensus: augmenting VLMs significantly boosts performance and interpretability on knowledge-intensive and reasoning-heavy tasks by synergizing neural pattern recognition with symbolic precision (Marino et al., 2020; Vedantam et al., 2019; Bitton-Guetta et al., 2024; Yan & Xie, 2024; Hu et al., 2023c). A particularly powerful paradigm emerging from this landscape is tool use, which offers a flexible and unifying abstraction for AVLM design. This approach frames the VLM as an intelligent orchestrator trained to select and utilize external capabilities (such as knowledge bases, calculators, code execution, specialized algorithms, formal reasoners) encapsulated as "tools," enabling modularity and scalability. Significant challenges remain in managing interaction complexity, ensuring scalability and efficiency, guaranteeing robustness against unreliable external inputs, developing comprehensive evaluation methods, and refining the tool integration mechanisms themselves. Nevertheless, the advancement of AVLMs, particularly through the lens of tool use, represents a crucial progression towards more capable, reliable, and trustworthy AI systems that effectively blend neural perception with symbolic reasoning, allowing them to not only see and describe the world but also reason about it with greater depth, accuracy, and transparency.

A Methodology

This section describes the process of gathering relevant articles for this survey, following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines. The goal of this approach is to avoid bias when selecting what papers to review, focusing on the merits of the paper and the relevancy to the topic of AVLMs.

A.1 Search Strategy

A.1.1 Databases and Search Queries

We utilized two primary databases for our literature search:

- **Google Scholar**: Known for its extensive coverage of scholarly publications across disciplines.
- **Semantic Scholar**: Provides advanced search capabilities and citation analysis, facilitating the identification of semantically relevant works.

A.1.2 Search Terms

We formulated specific search queries to capture studies related to augmented vision-language models interacting with symbolic systems during inference. The search strategy used the strengths of both databases by employing an iterative process of testing and refining the search query until the resulting set of papers was adequately relevant. Google Scholar is more sensitive to the inclusion of keywords, and so we used a combination of Boolean operators to refine the results effectively.

The search query used in Google Scholar was:

```
"("augmented" OR "knowledge" OR "knowledge graphs" OR
"knowledge augmentation" OR "commonsense knowledge" OR
"commonsense reasoning" OR "tool use" OR
"retrieval augmented" OR "retrieval-augmented" OR
"external knowledge" OR "neural symbolic" OR
"neural-symbolic" OR "symbolic")
AND
("vision-language" OR "vision language" OR
"visual question answering" OR "image question answering" OR
"video question answering" OR "image caption" OR
"video caption" OR "image text" OR "spatial reasoning" OR
"visual reasoning")
AND
("neural network" OR "machine learning" OR
"artificial intelligence" OR "deep learning")
-"virtual reality" -"augmented reality"
```

Semantic Scholar is less sensitive to keywords and more of a semantic search, so for this database, we employed a set of targeted queries to capture key aspects of our research focus:

- "Commonsense reasoning in visual question answering"
- "Knowledge graphs for image or video captioning"
- "External knowledge in visual reasoning"
- "Neural-symbolic vision-language models"
- "Tool use in vision-language tasks"

- "Retrieval-augmented image question answering"
- "Symbolic reasoning in AI for vision"
- "Commonsense in image-text models"
- "Neural-symbolic visual question answering"
- "Multimodal knowledge graph LLM"

A.2 Inclusion and Exclusion Criteria

To ensure the relevance and quality of the studies included in this review, we established clear inclusion and exclusion criteria.

A.2.1 Inclusion Criteria

- **Relevance**: Studies that describe machine learning models integrating external symbolic information systems during inference.
- **Language**: Publications written in English.
- **Implementation Focus**: Papers providing detailed descriptions of implementation methods rather than purely conceptual or theoretical discussions.
- **Vision-Language Tasks**: Research focusing on tasks such as visual question answering, image captioning, and video captioning where the input is imagery and/or text and the output is natural language text.

A.2.2 Exclusion Criteria

We excluded studies that did not align with the focus of this review, such as:

- **Prompting Techniques**: Research solely on prompt engineering or techniques that rely on internal reasoning patterns without external data augmentation (e.g., chain-of-thought prompting).
- **Self-Prompting/Recursive Prompting**: Methods that involve iterative querying without integration of external symbolic information systems.
- **Synthetic Data Generation**: Studies focusing on generating synthetic data to improve model performance without external symbolic system interaction.
- **Architectural Modifications Without External Integration**: Papers discussing model architectures like vision encoder adapters for large language models that do not involve external symbolic systems during inference.
- **Training with Structured Knowledge**: Research that involves training models with external knowledge but does not allow for the external knowledge to be modified or read during inference (e.g., methods where external knowledge is embedded in model parameters).

A.3 Selection Process

The selection process involved several iterative steps to refine and identify the most relevant studies.

A.3.1 Initial Search Results

- **Google Scholar**: The search yielded **980 papers** after filtering by category and removing irrelevant results based on titles and abstracts.
- **Semantic Scholar**: The targeted queries returned **1,332 papers**.

A.3.2 Total Papers Collected

In total, **2,312 papers** were collected from both databases.

A.3.3 Relevance Scoring

In alignment with the theme of augmented models, we utilized the **GPT-4o OpenAI model (gpt-4o-2024-08-06)** to assist in the relevance assessment:

- **Automated Categorization**: GPT-4o was prompted to categorize each paper and assign a relevance score ranging from 1 to 10 based on the alignment with the review topic.
- **Threshold for Inclusion**: Papers scoring less than **8 out of 10** were excluded from further consideration.
- **Iteration and Validation**: The relevance scoring process was iterated, and we ensured that all highly relevant papers were retained, even if they narrowly missed the initial threshold.

A.3.4 Manual Screening

- **Total Papers After Automated Filtering**: **616 papers** remained after applying the relevance threshold.
- **Full-Text Assessment**: We conducted a thorough manual review of the full text of these papers.
- **Final Selection**: After removing duplicates and papers not meeting the inclusion criteria, **264 papers** were selected for detailed analysis. See Figure 7.

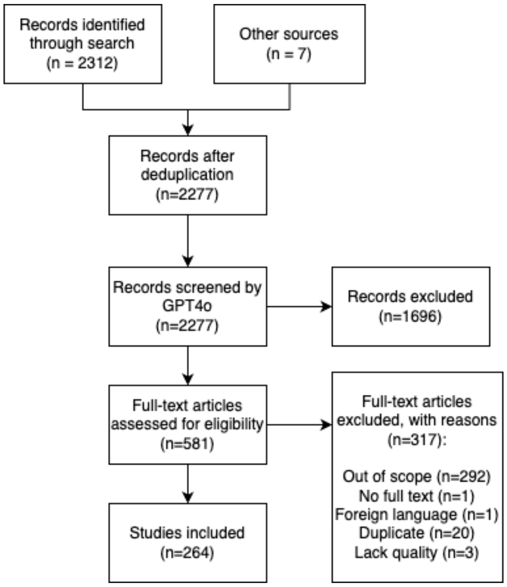

Records identified through search (n = 2312)
Other sources (n = 7)
Records after deduplication (n=2277)
Records screened by GPT4o (n=2277)
Records excluded (n=1696)
Full-text articles assessed for eligibility (n=581)
Full-text articles excluded, with reasons (n=317):
Out of scope (n=292)
No full text (n=1)
Foreign language (n=1)
Duplicate (n=20)
Lack quality (n=3)
Studies included (n=264)

Figure 7: PRISMA Flowchart

A.4 Data Extraction and Synthesis

From the selected studies, we extracted pertinent information to facilitate a comprehensive understanding of the methods:

- **Integration Techniques**: Description of how external systems were integrated with vision-language models, classified into early fusion, middle fusion, and late fusion methods.

- **Types of External Symbolic Systems**: Categorization of external symbolic systems used, such as knowledge graphs, symbolic logic engines, and program synthesis tools.
- **Tasks Addressed**: Identification of the specific vision-language tasks tackled by each study, including visual question answering, image captioning, and others.
- **Implementation Details**: Detailed examination of the models' architectures, including the interaction mechanisms with external symbolic systems during inference.

A.5 Quality Assessment

We assessed the quality of the included studies based on:

- **Clarity of Methodology**: Transparency and reproducibility of the methods described.
- **Experimental Rigour**: Adequacy of experimental design, including dataset usage, evaluation protocols, and statistical significance of results.
- **Contribution to the Field**: The extent to which the study advanced understanding or provided innovative solutions in augmented vision-language models.

A.6 Limitations

While we aimed for a comprehensive review, certain limitations exist:

- **Publication Bias**: Unpublished works or those not indexed in the selected databases may have been missed.
- **Language Restriction**: Non-English publications were excluded, which may omit relevant research conducted in other languages.
- **Dynamic Field**: Given the rapidly evolving nature of machine learning research, new studies may have emerged after the completion of our search.
- **AI Bias**: The use of GPT4o in filtering of papers could potentially remove relevant search results.

By following this systematic approach, we ensured a thorough and unbiased selection of relevant literature, providing a solid foundation for the subsequent analysis and discussion in this review.

B Categorization Tables

This section contains the tables categorizing the surveyed papers based on the fusion method (Early, Middle, Late) and the type of augmentation (Retrieval, Symbolic Computation, Combined). It also includes a table summarizing relevant datasets. These tables correspond to the synthesis presented in the main Results section.

Table 2: Early Fusion Methods in Vision-Language Model Augmentation. Bracketed tags denote predominant reasoning domain categories: **S**patial, **T**emporal, **K**nowledge grounding, **P**hysical commonsense, **A**ction/Embodiment

Retrieval			
Prompt Augmentation		Querying KG	**Retrieval Encoders**
Image Caption	Retrieval FT	Prompt Augmentation	Subgraph Enc
(Gao et al., 2022a) [**K**] (An et al., 2024) [**K**] (Li et al., 2018) [**S,K**] (Fabian et al., 2023) [**K**] (Sharifymoghaddam et al., 2024) [**K**] (Ghosal et al., 2023) [**K**] (Khademi et al., 2023) [**K**] (Fu et al., 2023) [**K**] (Dey et al., 2021) [**K**] (Lin et al., 2022) [**K**] (Yu et al., 2019) [**S,K**] (an Liu et al., 2024) [**K**] (Mogadala, 2019) [**K**] (Lin & Byrne, 2022) [**K**] (Garcia-Olano et al., 2021) [**K**] (Salemi et al., 2023b) [**K**] (Luo et al., 2021) [**K**] (Vo et al., 2022) [**K**] (Hao et al., 2024b) [**K**] (Gui et al., 2021) [**K**] (Chen et al., 2021a) [**K**] (Liang et al., 2021) [**K**] (Lerner et al., 2023) [**K**]	(Kan et al., 2023) [**K**] (Ranjit et al., 2023) [**K**] (Qu et al., 2024) [**K**] (Liu et al., 2024) [**K**] (Yan & Xie, 2024) [**K**] (Xu et al., 2024a) [**K**] (Khaliq et al., 2024) [**K**] (Xuan et al., 2024) [**K**] (Wen et al., 2024) [**K**] (Iscen et al., 2023) [**K**] (Joshi et al., 2024) [**K**] (Chen et al., 2022b) [**K**] (Gur et al., 2021) [**K**] (Cui et al., 2024) [**K**] (Zhu et al., 2023) [**K**] (Hao et al., 2024a) [**K**]	(Ravi et al., 2022) [**K**] (Narasimhan & Schwing, 2018) [**K**] (Vickers et al., 2021) [**K**] (Guo et al., 2022) [**K**] (Wang et al., 2015) [**S,K**] (Natu et al., 2023) [**K**] (Jhalani et al., 2024) [**K**] (Barezi & Kordjamshidi, 2024) [**K**] (Zhang et al., 2024) [**K**] (Zhang et al., 2023g) [**K**] (Wang et al., 2023) [**K**] (Ogawa et al., 2024) [**T**] (Chen et al., 2022c) [**K**] (Kan et al., 2021) [**S,K**] (Gan et al., 2023) [**K**] (Yang et al., 2019) [**T,K**]	(Li et al., 2020) [**K**] (Rao et al., 2023) [**K**] (Zhang et al., 2020) [**K**] (Lee & Kim, 2021) [**K**] (Li et al., 2022a) [**K**] (Lin et al., 2023a) [**K**] (Torino et al., 2020) [**S,K**] (Wang et al., 2022a) [**K**] (Qu et al., 2020) [**T,K**] (Padhi et al., 2024) [**K**] (Shevchenko et al., 2021) [**K**] (Gardères et al., 2020) [**K**] (Jing et al., 2023) [**K**] (Mondal et al., 2024) [**K**] (Lee et al., 2024) [**K**]

Retrieval			
Retrieval Encoders (Continued)			
KG Encoding		Encoder Architectures	
KG Conv	MMKG Attn	Attention	LSTM
(Chen et al., 2021b) [**K**] (Ziaeefard & Lécué, 2020) [**K**] (Yu et al., 2020) [**K**] (Zhu et al., 2020b) [**K**] (Hussain et al., 2022) [**K**] (Li & Moens, 2022) [**K**] (Ye et al., 2021) [**S**]	(Jiang & Meng, 2023) [**K**]	(Yuan et al., 2023b) [**K**] (Weng et al., 2024) [**K**] (Chen et al., 2022a) [**K**] (Salemi et al., 2023a) [**K**]	(Wu et al., 2016) [**S,K**]

Retrieval	**Symbolic**
Retrieval Encoders (cont.)	
Memory	Symbolic
(Hu et al., 2022) [**K**]	(Potapov et al., 2019) [**S**]

Table 3: Middle Fusion Methods in Vision-Language Model Augmentation. Bracketed tags denote predominant reasoning domain categories: **S**patial, **T**emporal, **K**nowledge grounding, **P**hysical commonsense, **A**ction/Embodiment

Retrieval			
Dense Retrieval	Graph		
	KG Prompt Augmentation	KG/NN Similarity	Concept/Scene Fusion
(Wang et al., 2022b) [**K**] (Lin et al., 2023b) [**K**] (Jia et al., 2023) [**K**]	(Li et al., 2017) [**K**] (Li et al., 2023b) [**K**] (Zheng et al., 2021) [**K**] (Su et al., 2018) [**K**] (Narasimhan et al., 2018) [**K**] (Zhang et al., 2023c) [**K**] (Singh et al., 2019) [**K**] (Jiang et al., 2020) [**T,A,K**] (Yu et al., 2023) [**K**] (Du et al., 2022) [**K**] (Li et al., 2024a) [**K**] (Yin et al., 2023) [**K**] (Ma et al., 2022) [**K**] (Zhu et al., 2020a) [**S,K**] (Cao et al., 2019) [**S,K**] (Li et al., 2022b) [**K**] (Zheng et al.) [**A,K**] (Wei et al., 2022) [**K**] (Narayanan et al., 2021) [**K**]	(Wu et al., 2024a) [**K**] (Chae & Kim, 2022) [**K**] (Li et al., 2019) [**K**] (ming Xian et al., 2023) [**K**] (Marino et al., 2020) [**K**]	(Yang et al., 2023) [**S,K**] (Wang et al., 2022c) [**S,K**] (Khan et al., 2022b) [**S,K**] (Khan et al., 2022a) [**S,K**] (Zhu, 2022) [**S,K**] (Wen & Peng, 2021) [**S,K**] (Song et al., 2023) [**K,P**] (Li et al., 2022d) [**A,K**] (Zhang et al., 2021) [**S,K**] (Dong et al., 2024) [**S,K**] (Gao et al., 2023a) [**S,A**] (Zhang et al., 2022a) [**K**] (Xu et al., 2021) [**T,K**] (Li et al., 2024b) [**K**] (Hou et al., 2020) [**T,K**] (Gu et al., 2019) [**S,K**]

Retrieval			**Symbolic Computation**
Graph		RL	Program Synthesis
MMKGs	Hypergraphs		
(Xi et al., 2024) [**T,K**] (Shi et al., 2022) [**S,K**] (Santiesteban et al., 2024) [**K**] (Ouyang et al., 2024) [**K**] (Liu et al., 2021) [**K**]	(Heo et al., 2022) [**K**] (Wang et al., 2024b) [**K**]	(Bougie et al., 2018) [**A,K**]	(Zhang et al., 2022c) [**K**] (Zhang et al., 2023e) [**S,K**] (Hu et al., 2023b) [**S,K**] (Shirai et al., 2023) [**A,S**] (Zhang et al., 2023b) [**A**] (Li et al., 2021) [**S,K**] (Mishra et al., 2024) [**K**] (Xue et al., 2024) [**S,K**]

Symbolic Computation			
Logic Engines	VSA	Symbolic Graph Ops	Tool Use
(Zhang et al., 2025a) [**K**] (Riley & Sridharan, 2019) [**K**] (Mitchener et al., 2021) [**A**] (Tsatsou et al., 2021) [**K**] (Zhang et al., 2022b) [**A**] (Choi et al., 2024) [**T,K**] (Li et al., 2023c) [**S**] (Li et al., 2023d) [**K**] (Huang et al., 2021) [**K**] (Zhang et al., 2023d) [**A**]	(Montone et al., 2017) [**S**] (Kovalev et al., 2021) [**S**]	(Li et al., 2022c) [**K**] (Liang et al., 2020) [**S,K**] (Wu et al., 2023) [**T,K**] (Zhao, 2015) [**P,S,K**] (Yang et al., 2020) [**S**] (Zhang et al., 2023f) [**K**] (Hudson & Manning, 2019) [**S**] (Cao et al., 2021) [**S,K**]	(Hu et al., 2024) [**S**] (Fan et al., 2024) [**T**] (Liu et al., 2023b) [**A**] (Hu et al., 2023c) [**K**] (Wu et al., 2024b) [**K**]

Symbolic Computation		**Combined Retr & Symb**
Self Play	Agents	Other
(Misiunas et al., 2024) [**K**]	(Niu et al., 2024) [**A**] (Castrejon et al., 2024) [**K**] (Lu et al., 2023) [**S,K**] (Hsieh et al., 2023) [**K**] (Xu et al., 2024b) [**A,K**] (Yang et al., 2024) [**T**]	(Besbes et al., 2015) [**K**] (Aditya et al., 2016) [**K,P**] (Aditya, 2017) [**S,K**] (Aditya & Baral, 2016) [**K**] (Tan et al., 2021) [**A,K**] (Liu et al., 2023a) [**K**] (Stammer et al., 2024) [**K**] (Vatashsky & Ullman, 2018) [**S,K**] (Gao et al., 2023c) [**S,K**] (Gao et al., 2024) [**S,K**]

Table 4: Late Fusion Methods in Vision-Language Model Augmentation. Bracketed tags denote predominant reasoning domain categories: **S**patial, **T**emporal, **K**nowledge grounding, **P**hysical commonsense, **A**ction/Embodiment

Retrieval		**Symbolic Computation**	
Dense	Knowledge Graph	Program Synth	Symbolic Engines
(Song et al., 2022a) [**K**] (Song et al., 2022b) [**K**] (Shi et al., 2024) [**S,T,K**]	(Gao et al., 2022b) [**S,K**] (Huang et al., 2020) [**K**] (Xiao & Fu, 2022) [**S,K**]	(Vedantam et al., 2019) [**S**] (Yi et al., 2018) [**S**] (Surís et al., 2023) [**S,P**] (Khandelwal et al., 2023) [**S**] (Subramanian et al., 2023) [**S,K**] (Gupta & Kembhavi, 2022) [**S**] (Bhaisaheb et al., 2023) [**S,K**]	(Sethuraman et al., 2021) [**S,K**] (Aditya et al., 2018) [**S**] (Eiter et al., 2022) [**S,K**] (Eiter et al., 2021) [**K**] (Cunnington et al., 2024) [**K**]

Symbolic Computation			**Combined**
Symbolic Graph Ops	Tool Use	Other	Combined
(Li et al., 2023a) [**A,T**] (Zhan et al., 2021) [**S**] (Saqur & Narasimhan, 2020) [**S,K**] (Johnston et al., 2023) [**S,K**]	(Yuan et al., 2023a) [**K**] (Cesista et al., 2024) [**K**] (Cesista, 2024) [**K**] (Zhang, 2023) [**K**]	(Xu et al., 2022) [**A**] (Singh, 2018) [**K**] (Bao et al., 2023) [**K**] (Verheyen et al., 2023) [**T,K**]	(Sachan, 2020) [**K**] (Basu et al., 2020) [**S,K**]

Table 5: Datasets Relevant to Augmented Vision-Language Models. Bracketed tags denote predominant reasoning domain categories: **S**patial, **T**emporal, **K**nowledge grounding, **P**hysical commonsense, **A**ction/Embodiment

Spatial Reasoning [S]		**Knowledge Based VQA** [K]	**Reasoning VQA** [S,K]
CLEVER	Scene Graph	KBVQA	Reasoning VQA
(Johnson et al., 2016) (Yi et al., 2019) (Li et al., 2022e) (Abraham et al., 2024) (Wang et al., 2024d)	(Krishna et al., 2016) (Shen et al., 2024)	(Agrawal et al., 2015) (Wang et al., 2016) (Lin et al., 2023c) (Shah et al., 2019) (Marino et al., 2019) (Reichman et al., 2023) (Su et al., 2024) (Mensink et al., 2023) (Jain et al., 2021) (Cao et al., 2020) (Sung et al., 2022) (Agarwal et al., 2024) (Qiu et al., 2024) (Lerner et al., 2022)	(Schwenk et al., 2022) (Gao et al., 2019) (Zellers et al., 2018) (Cao et al., 2019) (Bitton-Guetta et al., 2024)

Knowledge and Spatial [K,S]	**Agents** [A]	**Task Specific**	
Knowledge and Spatial	Agents	Robotics [A]	Other (Task) [K]
(Chen et al., 2023) (Wang et al., 2024a) (Zhang et al., 2023a)	(Niu et al., 2024) (Zhou et al., 2023) (Cao et al., 2024)	(Gao et al., 2023b)	(Hayashi et al., 2024) (Jin et al., 2024) (Hu et al., 2023a)

C Glossary of Key Terms

Table 6: Key Terms and Definitions

Term	Definition
Vision-Language Model (VLM)	Model jointly processing visual and textual modalities through tasks like VQA, captioning, or image-text retrieval.
Augmented VLM (AVLM)	VLM integrated with external symbolic systems, APIs, or tools during inference to overcome standalone limitations.
Neural-Symbolic System	Hybrid architecture combining neural pattern recognition with symbolic logical reasoning and knowledge representation.
Early Fusion	Integration of external information at input stage, before VLM internal processing begins.
Middle Fusion	Integration during VLM's inference, interacting with intermediate representations before final output.
Late Fusion	Integration after VLM generates initial output, typically for validation, refinement, or explanation.
Retrieval Augmented Generation (RAG)	Retrieving relevant external information to provide as context for model generation.
Knowledge Graph (KG)	Structured knowledge as directed graph with entity nodes and relationship edges, encoded as triplets.
Scene Graph	Structured visual scene representation with object nodes and spatial relationship edges.
Program Synthesis	Automatic generation of executable code (Python, SQL) by models for reasoning operations.
Tool Use	VLMs dynamically invoking external tools (calculators, APIs, vision modules) based on task needs.
Graph Neural Network (GNN)	Neural architecture for graph data, enabling message passing across nodes and edges.
Dense Retrieval	A retrieval method using learned dense vector embeddings to find semantically similar content through vector similarity metrics rather than keyword matching.
Visual Question Answering (VQA)	Task requiring natural language answers to questions about visual content.
ConceptNet	Multilingual common-sense knowledge graph with semantic concept networks.
Answer Set Programming (ASP)	Declarative programming paradigm for knowledge representation and logical constraint solving.
PDDL	Planning Domain Definition Language - a standardized language for expressing planning problems and domains in automated planning systems.

References

Savitha Sam Abraham, Marjan Alirezaie, and L. D. Raedt. Clevr-poc: Reasoning-intensive visual question answering in partially observable environments. pp. 3297–3313, 2024.

Somak Aditya. Explainable image understanding using vision and reasoning. pp. 5028–5029, 2017.

Somak Aditya and Chitta Baral. Deepiu : An architecture for image understanding. 2016.

Somak Aditya, Yezhou Yang, Chitta Baral, and Yiannis Aloimonos. Answering image riddles using vision and reasoning through probabilistic soft logic. *arXiv preprint*, arXiv:1611.05896v1, 2016.

Somak Aditya, Rudra Saha, Yezhou Yang, and Chitta Baral. Spatial knowledge distillation to aid visual reasoning. *arXiv preprint*, arXiv:1812.03631v2, 2018.

Somak Aditya, Yezhou Yang, and Chitta Baral. Integrating knowledge and reasoning in image understanding. *IJCAI 2019*, 2019.

Pulkit Agarwal, S. Sravanthi, and Pushpak Bhattacharyya. Indifoodvqa: Advancing visual question answering and reasoning with a knowledge-infused synthetic data generation pipeline. pp. 1158–1176, 2024.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. *arXiv preprint*, arXiv:1505.00468v7, 2015.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL `https://arxiv.org/abs/2204.14198`.

Wenbin An, Feng Tian, Jiahao Nie, Wenkai Shi, Haonan Lin, Yan Chen, Qianying Wang, Y. Wu, Guang Dai, and Ping Chen. Knowledge acquisition disentanglement for knowledge-based visual question answering with large language models. *ArXiv*, abs/2407.15346, 2024.

An an Liu, Chenxi Huang, Ning Xu, Hongshuo Tian, J. Liu, and Yongdong Zhang. Counterfactual visual dialog: Robust commonsense knowledge learning from unbiased training. *IEEE Transactions on Multimedia*, 26:1639–1651, 2024.

Yajie Bao, Tianwei Xing, and Xun Chen. Confidence-based interactable neural-symbolic visual question answering. *Neurocomputing*, 564:126991, 2023.

Elham J. Barezi and Parisa Kordjamshidi. Find the gap: Knowledge base reasoning for visual question answering. *arXiv preprint*, arXiv:2404.10226v1, 2024.

Kuntal Basu, Farhad Shakerin, and Gopal Gupta. AQuA: ASP-Based Visual Question Answering. In Ekaterina Komendantskaya and Yanhong A. Liu (eds.), *Practical Aspects of Declarative Languages (PADL 2020)*, volume 12007 of *Lecture Notes in Computer Science*. Springer, Cham, 2020. doi: 10.1007/978-3-030-39197-3_4. URL `https://doi.org/10.1007/978-3-030-39197-3_4`.

Ghada Besbes, H. B. Zghal, and H. Ghézala. An ontology-driven visual question-answering framework. *2015 19th International Conference on Information Visualisation*, pp. 127–132, 2015.

Shabbirhussain Bhaisaheb, Shubham Paliwal, Rajaswa Patil, Manasi S. Patwardhan, L. Vig, and Gautam M. Shroff. Program synthesis for complex qa on charts via probabilistic grammar based filtered iterative back-translation. pp. 2456–2470, 2023.

Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. *arXiv preprint*, arXiv:2407.19474v2, 2024.

Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jon Lenchner, Nick Linck, Andrea Loreggia, Keerthiram Murugesan, Nicholas Mattei, Francesca Rossi, and Biplav Srivastava. Thinking fast and slow in ai. *Proceedings of the AAAI Conference on Artificial Intelligence 2021, 35(17), 15042-15046*, 2020.

Nicolas Bougie, Limei Cheng, and R. Ichise. Combining deep reinforcement learning with prior knowledge and reasoning. *ACM SIGAPP Applied Computing Review*, 2018.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, K. Choromanski, Tianli Ding, Danny Driess, Kumar Avinava Dubey, Chelsea Finn, Peter R. Florence, Chuyuan Fu, Montse Gonzalez Arenas, K. Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, A. Irpan, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, S. Levine, H. Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, M. Ryoo, Grecia Salazar, Pannag R. Sanketi, P. Sermanet, Jaspiar Singh, Anika Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Q. Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Ted Xiao, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *ArXiv*, abs/2307.15818, 2023.

Qingxing Cao, Bailin Li, Xiaodan Liang, and Liang Lin. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network. *ArXiv*, abs/1909.10128, 2019.

Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 33:2758–2767, 2020.

Qingxing Cao, Wentao Wan, Keze Wang, Xiaodan Liang, and Liang Lin. Linguistically routing capsule network for out-of-distribution visual question answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1594–1603, 2021.

Ruisheng Cao, Fangyu Lei, Haoyuan Wu, Jixuan Chen, Yeqiao Fu, Hongcheng Gao, Xinzhuang Xiong, Hanchong Zhang, Yuchen Mao, Wenjing Hu, Tianbao Xie, Hongshen Xu, Danyang Zhang, Sida Wang, Ruoxi Sun, Pengcheng Yin, Caiming Xiong, Ansong Ni, Qian Liu, Victor Zhong, Lu Chen, Kai Yu, and Tao Yu. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? *arXiv preprint*, arXiv:2407.10956v1, 2024.

Lluis Castrejon, Thomas Mensink, Howard Zhou, Vittorio Ferrari, Andre Araujo, and Jasper Uijlings. Hammr: Hierarchical multimodal react agents for generic vqa. *arXiv preprint*, arXiv:2404.05465v2, 2024.

Franz Louis Cesista. Multimodal structured generation: Cvpr's 2nd mmfm challenge technical report. *arXiv preprint*, arXiv:2406.11403v2, 2024.

Franz Louis Cesista, Rui Aguiar, Jason Kim, and Paolo Acilo. Retrieval augmented structured generation: Business document information extraction as tool use. *arXiv preprint*, arXiv:2405.20245v1, 2024.

Jinyeong Chae and Jihie Kim. Uncertainty-based visual question answering: Estimating semantic inconsistency between image and knowledge base. *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2022.

Kezhen Chen, Qiuyuan Huang, Yonatan Bisk, Daniel J. McDuff, and Jianfeng Gao. Kb-vlp: Knowledge based vision and language pretraining. 2021a.

Kezhen Chen, Qiuyuan Huang, Daniel J. McDuff, Yonatan Bisk, and Jianfeng Gao. Krit: Knowledge-reasoning intelligence in vision-language transformer. 2022a.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. pp. 5558–5570, 2022b.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *ArXiv*, abs/2302.11713, 2023.

Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen. Zero-shot visual question answering using knowledge graph. *ArXiv*, abs/2107.05348, 2021b.

Zhuo Chen, Yufen Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z. Pan, Ningyu Zhang, and Wen Zhang. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, 2022c.

Minkyu Choi, Harsh Goel, Mohammad Omama, Yunhao Yang, Sahil Shah, and Sandeep Chinchali. Towards neuro-symbolic video understanding. 2024.

Wanqing Cui, Keping Bi, J. Guo, and Xueqi Cheng. More: Multi-modal retrieval augmented generative commonsense reasoning. *ArXiv*, abs/2402.13625, 2024.

Daniel Cunnington, Mark Law, Jorge Lobo, and Alessandra Russo. The role of foundation models in neuro-symbolic learning and reasoning. *ArXiv*, abs/2402.01889, 2024.

Arka Ujjal Dey, Ernest Valveny, and Gaurav Harit. External knowledge augmented text visual question answering. *ArXiv*, abs/2108.09717, 2021.

Junnan Dong, Qinggang Zhang, Huachi Zhou, D. Zha, Pai Zheng, and Xiao Huang. Modality-aware integration with large language models for knowledge-based visual question answering. pp. 2417–2429, 2024.

Qinyi Du, Qingqing Wang, Keqian Li, Jidong Tian, Liqiang Xiao, and Yaohui Jin. Calm: Commen-sense knowledge augmentation for document image understanding. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Thomas Eiter, N. Higuera, J. Oetsch, and Michael Pritz. A confidence-based interface for neuro-symbolic visual question answering. 2021.

Thomas Eiter, N. Higuera, J. Oetsch, and Michael Pritz. A neuro-symbolic asp pipeline for visual question answering. *Theory and Practice of Logic Programming*, 22:739 – 754, 2022.

Zalan Fabian, Zhongqi Miao, Chunyuan Li, Yuanhan Zhang, Ziwei Liu, A. Hern'andez, Andrés Montes-Rojas, Rafael S. Escucha, Laura Siabatto, Andr'es Link, Pablo Arbel'aez, R. Dodhia, and J. Ferres. Multimodal foundation models for zero-shot animal species recognition in camera trap images. *ArXiv*, abs/2311.01064, 2023.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *ArXiv*, abs/2403.11481, 2024.

Xingyu Fu, Shenmin Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, A. Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, D. Roth, and Bing Xiang. Generate then select: Open-ended visual question answering guided by world knowledge. *ArXiv*, abs/2305.18842, 2023.

Jingru Gan, Xinzhe Han, Shuhui Wang, and Qingming Huang. Open-set knowledge-based visual question answering with inference paths. *ArXiv*, abs/2310.08148, 2023.

Chen Gao, Si Liu, Jinyu Chen, Luting Wang, Qi Wu, Bo Li, and Qi Tian. Room-object entity prompting and reasoning for embodied referring expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:994–1010, 2023a.

Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *arXiv preprint*, arXiv:1908.02962v3, 2019.

Feng Gao, Q. Ping, G. Thattai, Aishwarya N. Reganti, Yingting Wu, and Premkumar Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5057–5067, 2022a.

Jensen Gao, Bidipta Sarkar, F. Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12462–12469, 2023b.

Jingying Gao, A. Blair, and M. Pagnucco. A symbolic-neural reasoning model for visual question answering. *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2023c.

Jingying Gao, Alan Blair, and Maurice Pagnucco. Explainable visual question answering via hybrid neural-logical reasoning. *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2024.

Yueqing Gao, Huachun Zhou, Lulu Chen, Yuting Shen, Ce Guo, and Xinyu Zhang. Cross-modal object detection based on a knowledge update. *Sensors (Basel, Switzerland)*, 22, 2022b.

Diego Garcia-Olano, Yasumasa Onoe, and J. Ghosh. Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection. *Companion Proceedings of the Web Conference 2022*, 2021.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lécué. Conceptbert: Concept-aware representation for visual question answering. In *Findings*, 2020. URL `https://api.semanticscholar.org/CorpusID:226284018`.

Deepanway Ghosal, Navonil Majumder, Roy Ka-Wei Lee, Rada Mihalcea, and Soujanya Poria. Language guided visual question answering: Elevate your multimodal language model using knowledge-enriched prompts. pp. 12096–12102, 2023.

Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. *arXiv preprint*, arXiv:1904.00560v1, 2019.

Liangke Gui, Borui Wang, Qiuyuan Huang, A. Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *ArXiv*, abs/2112.08614, 2021.

Yangyang Guo, Liqiang Nie, Yongkang Wong, Y. Liu, Zhiyong Cheng, and Mohan S. Kankanhalli. *A Unified End-to-End Retriever-Reader Framework for Knowledge-based VQA*. 2022.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *arXiv preprint*, arXiv:2211.11559v1, 2022.

Shir Gur, N. Neverova, C. Stauffer, S. Lim, Douwe Kiela, and A. Reiter. Cross-modal retrieval augmentation for multi-modal classification. *ArXiv*, abs/2104.08108, 2021.

Dongze Hao, Jian Jia, Longteng Guo, Qunbo Wang, Te Yang, Yan Li, Yanhua Cheng, Bo Wang, Quan Chen, Han Li, and Jing Liu. Knowledge condensation and reasoning for knowledge-based vqa. *ArXiv*, abs/2403.10037, 2024a.

Dongze Hao, Qunbo Wang, Longteng Guo, Jie Jiang, and Jing Liu. Self-bootstrapped visual-language model for knowledge selection and question answering. 2024b.

Kazuki Hayashi, †. YusukeSakai, Hidetaka Kamigaito, ‡. KatsuhikoHayashi, †. TaroWatanabe, Jean-Baptiste, Xincan Feng, Katsuhiko Hayashi, Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Noa García, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, Teruko Mitamura, A. dataset, D. Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, S. Hambardzumyan, Abhina Tuli, Levon Ghukasyan, Fariz Rahman, Hrant Topchyan, David Isayan, M. Harutyunyan, Tatevik Hakobyan, Albert Q. Jiang, Alexandre Sablayrolles, Arthur Men-sch, Chris Bamford, Devendra Singh, Diego Chaplot, Florian de las Casas, Gianna Bressand, Guil laume Lengyel, Lucile Lample, Lélio Renard Saulnier, Lavaud Marie-Anne, Pierre Lachaux, Teven Stock, Le Scao Thibaut, Thomas Lavril, Timothée Wang, Lacroix William, El Sayed, Mistral, Scott Kushal Kafle, Cohen, Shyamal Anadkat, Red Avila, Igor Babuschkin, S. Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Ir wan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, M. Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Su-Yuan Chen, Ruby Chen, Jason Chen, Mark Chen, B. Chess, Chester Cho, Hyung Casey Chu, Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Simón Niko Felix, Posada Fishman, Juston Forte, Is abella

Fulford, Leo Gao, Elie Georges, C. Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross Shixiang, Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, B. Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, I. Kanitscheider, Nitish Shirish, Tabarak Keskar, Logan Khan, J. Kilpatrick, Wook, Christina Kim, Yongjik Kim, Hendrik Kim, Jamie Kirch-ner, Matt Kiros, Daniel Knight, Kokotajlo Łukasz, A. Kondraciuk, Aris Kondrich, Kyle Kon-stantinidis, Gretchen Kosic, Vishal Krueger, Michael Kuo, Ikai Lampe, Teddy Lan, Jan Lee, Jade Leike, Daniel Leung, Chak Ming Levy, Li Rachel, Molly Lim, Stephanie Lin, Mateusz Lin, Theresa Litwin, Ryan Lopez, Patricia Lowe, Lue Anna, Kim Makanju, S. Malfacini, Todor Manning, Yaniv Markov, Bianca Markovski, Katie Martin, Andrew Mayer, Bob Mayne, Scott Mayer McGrew, Christine McKinney, Paul McLeavey, McMillan Jake, David McNeil, Aalok Medina, Jacob Mehta, Luke Menick, Andrey Metz, Pamela Mishchenko, Vinnie Mishkin, Evan Monaco, Daniel Morikawa, Tong Mossing, Mira Mu, Oleg Murati, David Murk, Ashvin Mély, Reiichiro Nair, Rajeev Nakano, Nayak Arvind, Richard Neelakantan, Hyeonwoo Ngo, Noh Long, Cullen Ouyang, Jakub O'Keefe, Alex Pachocki, J. Paino, Ashley Palermo, Giambat tista Pantuliano, Joel Parascandolo, Emy Parish, Alex Parparita, Mikhail Passos, Andrew Pavlov, Adam Peng, Filipe Perel-man, de Avila Belbute, Michael Peres, Petrov Henrique, Pondé, Michael Oliveira Pinto, Michelle Pokrass, Vitchyr H. Pong, Tolly Pow-ell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-der, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Tianhao Shengjia Zhao, Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, Roozbeh Mottaghi. 2022, Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-bert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, Punit Artem Korenev, Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-tinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-stein, Rashi Rungta, Kalyan Saladi, Liang Wang, Wei Zhao, Zhuoyu Wei, Jingming Liu, SimKGC, Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtow-icz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, J. Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, Tamara Berg, Modeling, Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, Wenhu Chen, Mmmu, Susan Zhang, Stephen Roller, Mikel Artetxe, Shuohui Chen, Christopher De-wan, Mona Diab, Xi Xian Li, Todor Victoria Lin, Myle Ott, Kurt Shuster, Punit Daniel Simig, Anjali Sridhar, Tianlu Wang, Luke Zettlemoyer. 2022a, Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. 2020, Bertscore, Yao Zhang, Haokun Chen, Ahmed Frikha, Yezi Yang, Denis Krompass, and Jindong Gu. Towards artwork explanation in large-scale vision language models. pp. 705–729, 2024.

Y. Heo, Eun-Sol Kim, Woo Suk Choi, and Byoung-Tak Zhang. Hypergraph transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. *ArXiv*, abs/2204.10448, 2022.

P. Hitzler, Md Kamruzzaman Sarker, and Aaron Eberhart. Neuro-symbolic spatio-temporal reasoning. 2022.

Jingyi Hou, Xinxiao Wu, Xiaoxun Zhang, Yayun Qi, Yunde Jia, and Jiebo Luo. Joint commonsense and relation reasoning for image and video captioning. pp. 10973–10980, 2020.

Cheng-Yu Hsieh, Sibei Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander J. Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Tool documentation enables zero-shot tool-usage with large language models. *ArXiv*, abs/2308.00675, 2023.

Joy Hsu, Jiayuan Mao, and Jiajun Wu. Ns3d: Neuro-symbolic grounding of 3d objects and relations. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2614–2623, 2023.

Xinyue Hu, Lin Gu, Qi A. An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, R. M. Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023a.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, K. Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9590–9601, 2023b.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint*, arXiv:2406.09403v3, 2024.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, C. Schmid, David A. Ross, and A. Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23369–23379, 2022.

Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David A Ross, Cordelia Schmid, and Alireza Fathi. Avis: Autonomous visual information seeking with large language model agent. *arXiv preprint*, arXiv:2306.08129v3, 2023c.

Feicheng Huang, Zhixin Li, Haiyang Wei, Canlong Zhang, and Huifang Ma. Boost image captioning with knowledge reasoning. *arXiv preprint*, arXiv:2011.00927v1, 2020.

Jiani Huang, Ziyang Li, Binghong Chen, Karan Samel, M. Naik, Le Song, and X. Si. Scallop: From probabilistic deductive databases to scalable differentiable reasoning. pp. 25134–25145, 2021.

Zhiao Huang, Feng Chen, Yewen Pu, Chun-Tse Lin, Hao Su, and Chuang Gan. Diffvl: Scaling up soft body manipulation using vision-language driven differentiable physics. *ArXiv*, abs/2312.06408, 2023.

Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine. *arXiv preprint*, arXiv:1907.03950v4, 2019.

Afzaal Hussain, Ifrah Maqsood, M. Shahzad, and M. Fraz. Multimodal knowledge reasoning for enhanced visual question answering. *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 224–230, 2022.

Ahmet Iscen, Mathilde Caron, A. Fathi, and C. Schmid. Retrieval-enhanced contrastive vision-text models. *ArXiv*, abs/2306.07196, 2023.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, P. Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

A. Jamshed and M. Fraz. Nlp meets vision for visual interpretation - a retrospective insight and future directions. In *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pp. 1–8, 2021.

Manas Jhalani, Annervaz K M, and Pushpak Bhattacharyya. Precision empowers, excess distracts: Visual question answering with dynamically infused knowledge in language models. *arXiv preprint*, arXiv:2406.09994v1, 2024.

Zhiwei Jia, P. Narayana, Arjun Reddy Akula, G. Pruthi, Haoran Su, Sugato Basu, and Varun Jampani. Kafa: Rethinking image ad understanding with knowledge-augmented feature adaptation of vision-language models. *ArXiv*, abs/2305.18373, 2023.

Chen Jiang, Masood Dehghan, and Martin Jägersand. Understanding contexts inside robot and human manipulation tasks through vision-language model and ontology system in video streams. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8366–8372, 2020.

Lei Jiang and Zuqiang Meng. Knowledge-based visual question answering using multi-modal semantic graph. *Electronics*, 2023.

Ruihan Jin, Ruibo Fu, Zhengqi Wen, Shuai Zhang, Yukun Liu, and Jianhua Tao. Fake news detection and manipulation reasoning via large vision-language models. *ArXiv*, abs/2407.02042, 2024.

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. Multi-source semantic graph-based multimodal sarcasm explanation generation. pp. 11349–11361, 2023.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint*, arXiv:1612.06890v1, 2016.

Penny Johnston, Keiller Nogueira, and Kevin Swingler. Ns-il: Neuro-symbolic visual question answering using incrementally learnt, independent probabilistic models for small sample sizes. *IEEE Access*, 11: 141406–141420, 2023.

Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia. Robust multi model rag pipeline for documents containing text, table & images. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 993–999, 2024.

Daniel Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, 2011.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL `https://arxiv.org/abs/2509.04664`.

Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15624–15634, 2023.

Xuan Kan, Hejie Cui, and Carl Yang. Zero-shot scene graph relation prediction through commonsense knowledge integration. pp. 466–482, 2021.

Mahmoud Khademi, Ziyi Yang, F. Frujeri, and Chenguang Zhu. Mm-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering. pp. 6571–6581, 2023.

Tarun Khajuria, Braian Olmiro Dias, and Jaan Aru. How structured are the representations in transformer-based vision encoders? an analysis of multi-object representations in vision-language models. *ArXiv*, abs/2406.09067, 2024.

M. A. Khaliq, P. Chang, M. Ma, B. Pflugfelder, and F. Mileti'c. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. *ArXiv*, abs/2404.12065, 2024.

M. J. Khan, Filip Ilievski, John G. Breslin, and Edward Curry. A survey of neurosymbolic visual reasoning with scene graphs and common sense knowledge. *Neurosymbolic Artificial Intelligence*, 2024.

Muhammad Jaleed Khan, J. Breslin, and E. Curry. Expressive scene graph generation using commonsense knowledge infusion for visual understanding and reasoning. pp. 93–112, 2022a.

Muhammad Jaleed Khan, J. Breslin, and E. Curry. Neusire: Neural-symbolic image representation and enrichment for visual understanding and reasoning. 2022b.

Apoorv Khandelwal, Ellie Pavlick, and Chen Sun. Analyzing modular approaches for visual question decomposition. *arXiv preprint*, arXiv:2311.06411v1, 2023.

A. Kovalev, M. Shaban, Evgeny Osipov, and A. Panov. Vector semiotic model for visual question answering. *Cognitive Systems Research*, 71:52–63, 2021.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint*, arXiv:1602.07332v1, 2016.

Jaeyun Lee and Incheol Kim. Vision–language–knowledge co-embedding for visual commonsense reasoning. *Sensors (Basel, Switzerland)*, 21, 2021.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint*, arXiv:2406.02030v2, 2024.

Paul Lerner, Olivier Ferret, C. Guinaudeau, H. Borgne, Romaric Besançon, José G. Moreno, and Jesús Lovón-Melgarejo. *ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities.* 2022.

Paul Lerner, O. Ferret, and C. Guinaudeau. Multimodal inverse cloze task for knowledge-based visual question answering. pp. 569–587, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL `https://arxiv.org/abs/2005.11401`.

Bojin Li, Yan Sun, Xue Chen, and Xiangfeng Luo. Hkfnet: Fine-grained external knowledge fusion for fact-based visual question answering. *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024a.

Guohao Li, Hang Su, and Wenwu Zhu. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. *ArXiv*, abs/1712.00733, 2017.

Guohao Li, Xin Wang, and Wenwu Zhu. Boosting visual question answering with context-aware knowledge aggregation. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

Hui Li, Peng Wang, Chunhua Shen, and Anton van den Hengel. Visual question answering as reading comprehension. *arXiv preprint*, arXiv:1811.11903v1, 2018.

Jiangmeng Li, Wenyi Mo, Wenwen Qiang, Bing Su, and Changwen Zheng. Supporting vision-language model inference with causality-pruning knowledge prompt. *ArXiv*, abs/2205.11100, 2022a.

Meng Li, Tianbao Wang, Jiahe Xu, Kairong Han, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Shiliang Pu, and Fei Wu. Multi-modal action chain abductive reasoning. pp. 4617–4628, 2023a.

Mingxiao Li and Marie-Francine Moens. Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering. pp. 10983–10992, 2022.

Qifeng Li, Xinyi Tang, and Yi Jian. Learning to reason on tree structures for knowledge-based visual question answering. *Sensors (Basel, Switzerland)*, 22, 2022b.

Qun Li, Fu Xiao, Le An, Xianzhong Long, and Xiaochuan Sun. Semantic concept network and deep walk-based visual question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15:1 – 19, 2019.

Qun Li, Fu Xiao, B. Bhanu, Biyun Sheng, and Richang Hong. Inner knowledge-based img2doc scheme for visual question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18:1 – 21, 2022c.

Xin Li, Yu Zhang, Weilin Yuan, and Junren Luo. Incorporating external knowledge reasoning for vision-and-language navigation with assistant's help. *Applied Sciences*, 2022d.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *ArXiv*, abs/2309.13625, 2023b.

Yili Li, Jing Yu, Keke Gai, and Gang Xiong. Iiu: Independent inference units for knowledge-based visual question answering. pp. 109–120, 2024b.

Yunxin Li, Baotian Hu, Yunxin Ding, Lin Ma, and M. Zhang. A neural divide-and-conquer reasoning framework for image retrieval from linguistically complex text. *ArXiv*, abs/2305.02265, 2023c.

Zhuowan Li, Elias Stengel-Eskin, Yixiao Zhang, Cihang Xie, Quan Tran, Benjamin Van Durme, and Alan Yuille. Calibrating concepts and operations: Towards symbolic reasoning on real images. *arXiv preprint*, arXiv:2110.00519v1, 2021.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. *arXiv preprint*, arXiv:2212.00259v2, 2022e.

Ziyang Li, Jiani Huang, and Mayur Naik. Scallop: A language for neurosymbolic programming. *arXiv preprint*, arXiv:2304.04812v1, 2023d.

Weixin Liang, Fei Niu, Aishwarya N. Reganti, G. Thattai, and Gökhan Tür. Lrta: A transparent neural-symbolic reasoning framework with modular supervision for visual question answering. *ArXiv*, abs/2011.10731, 2020.

Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. Maria: A visual experience powered conversational agent. *arXiv preprint*, arXiv:2105.13073v2, 2021.

Bingqian Lin, Zicong Chen, Mingjie Li, Haokun Lin, Hang Xu, Yi Zhu, Jian zhuo Liu, Wenjia Cai, Lei Yang, Shen Zhao, Chenfei Wu, Ling Chen, Xiaojun Chang, Yi Yang, L. Xing, and Xiaodan Liang. Towards medical artificial general intelligence via knowledge-enhanced multimodal pretraining. *ArXiv*, abs/2304.14204, 2023a.

Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint*, arXiv:2210.03809v2, 2022.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *ArXiv*, abs/2309.17133, 2023b.

Weizhe Lin, Zhilin Wang, and B. Byrne. Fvqa 2.0: Introducing adversarial samples into fact-based visual question answering. *ArXiv*, abs/2303.10699, 2023c.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. *ArXiv*, abs/2206.01201, 2022.

Jiawei Liu, Jingyi Xie, Fanrui Zhang, Qiang Zhang, and Zhengjun Zha. Knowledge-enhanced hierarchical information correlation learning for multi-modal rumor detection. *ArXiv*, abs/2306.15946, 2023a.

Luping Liu, Meiling Wang, Xiaohai He, L. Qing, and Honggang Chen. Fact-based visual question answering via dual-process system. *Knowl. Based Syst.*, 237:107650, 2021.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint*, arXiv:2311.05437v1, 2023b.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Rar: Retrieving and ranking augmented mllms for visual recognition. *arXiv preprint*, arXiv:2403.13805v1, 2024.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Y. Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *ArXiv*, abs/2304.09842, 2023.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. pp. 6417–6431, 2021.

Maria Lymperaiou and G. Stamou. A survey on knowledge-enhanced multimodal learning. *Artif. Intell. Rev.*, 57:284, 2022.

Xuan Ma, Xiaoshan Yang, and Changsheng Xu. Multi-source knowledge reasoning graph network for multimodal commonsense inference. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19:1 – 17, 2022.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *arXiv preprint*, arXiv:1906.00067v2, 2019.

Kenneth Marino, Xinlei Chen, Devi Parikh, A. Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14106–14116, 2020.

Thomas Mensink, J. Uijlings, Lluís Castrejón, A. Goel, Felipe Cadar, Howard Zhou, Fei Sha, A. Araújo, and V. Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3090–3101, 2023.

Guang ming Xian, Wencong Zhang, Fucai Lan, Yifan Lin, and Yanhang Lin. Multimodal knowledge triple extraction based on representation learning. In *2023 5th International Conference on Electronic Engineering and Informatics (EEI)*, pp. 684–689, 2023.

Aakansha Mishra, S. S. Miriyala, and V. N. Rajendiran. Learning representations from explainable and connectionist approaches for visual question answering. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6420–6424, 2024.

Tautvydas Misiunas, Hassan Mansoor, Jasper Uijlings, Oriana Riva, and Victor Carbune. Vqa training sets are self-play environments for generating few-shot pools. *ArXiv*, abs/2405.19773, 2024.

Ludovico Mitchener, David Tuckey, Matthew Crosby, and A. Russo. Detect, understand, act a neuro-symbolic hierarchical reinforcement learning framework. 2021.

Aditya Mogadala. Multi-view representation learning for unifying languages, knowledge and vision. 2019.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. pp. 18798–18806, 2024.

Guglielmo Montone, J. O'Regan, and A. Terekhov. Hyper-dimensional computing for a visual question-answering system that is trainable end-to-end. *ArXiv*, abs/1711.10185, 2017.

A. Mostafa, Hazem M. Abbas, and M. Khalil. Comparative study of visual question answering algorithms. In *International Conference on Communication and Electronics Systems*, pp. 1–6, 2020.

Medhini Narasimhan and Alexander G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. *arXiv preprint*, arXiv:1809.01124v1, 2018.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *arXiv preprint*, arXiv:1811.00538v1, 2018.

Abhishek Narayanan, Abijna Rao, Abhishek Prasad, and N. S. Vqa as a factoid question answering problem: A novel approach for knowledge-aware and explainable visual question answering. *Image Vis. Comput.*, 116:104328, 2021.

Sanika Natu, Shounak Sural, and Sulagna Sarkar. External commonsense knowledge as a modality for social intelligence question-answering. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3036–3042, 2023.

Zhe Ni, Xiao-Xin Deng, Cong Tai, Xin-Yue Zhu, Xiang Wu, Y. Liu, and Long Zeng. Grid: Scene-graph-based instruction-driven robotic task planning. *ArXiv*, abs/2309.07726, 2023.

Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. Screenagent: A vision language model-driven computer control agent. *ArXiv*, abs/2402.07945, 2024.

Tomohiro Ogawa, Kango Yoshioka, Ken Fukuda, and Takeshi Morita. Prediction of actions and places by the time series recognition from images with multimodal llm. *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pp. 294–300, 2024.

Kun Ouyang, Yi Liu, Shicheng Li, Ruihan Bao, Keiko Harimoto, and Xu Sun. Modal-adaptive knowledge-enhanced graph-based financial prediction from monetary policy conference calls with llm. *ArXiv*, abs/2403.16055, 2024.

Trilok Padhi, Ugur Kursuncu, Yaman Kumar, V. Shalin, and Lane Peterson Fronczek. Improving contextual congruence across modalities for effective multimodal marketing using knowledge-infused learning. *ArXiv*, abs/2402.03607, 2024.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, March 2021. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.n71. URL `https://doi.org/10.1136/bmj.n71`. Epub 2021-03-29.

A. Potapov, A. Belikov, V. Bogdanov, and Alexander Scherbatiy. Cognitive module networks for grounded reasoning. pp. 148–158, 2019.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023. URL `https://arxiv.org/abs/2307.16789`.

Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A. Crook, Y. Xu, Xin Luna Dong, Christos Faloutsos, Lei Li, Babak Damavandi, and Seungwhan Moon. Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. *ArXiv*, abs/2403.04735, 2024.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jiashuo Sun, and Jianfeng Dong. Alleviating hallucination in large vision-language models with active retrieval augmentation. *ArXiv*, abs/2408.00555, 2024.

Zhaowei Qu, Luhan Zhang, Xiaoru Wang, Bingyu Cao, Yueli Li, and Fu Li. Ksf-st: Video captioning based on key semantic frames extraction and spatio-temporal attention mechanism. *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 1388–1393, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. Retrieval augmented chest x-ray report generation using openai gpt models. *arXiv preprint*, arXiv:2305.03660v1, 2023.

Jiahua Rao, Zifei Shan, Long Liu, Yao Zhou, and Yuedong Yang. Retrieval-based knowledge augmented vision language pre-training. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. Vlc-bert: Visual question answering with contextualized commonsense knowledge. *arXiv preprint*, arXiv:2210.13626v1, 2022.

Benjamin Z. Reichman, Anirudh S. Sundar, Christopher Richardson, Tamara Zubatiy, Prithwijit Chowdhury, Aaryan Shah, Jack Truxal, Micah Grimes, Dristi Shah, Woo Ju Chee, Saif Punjwani, Atishay Jain, and Larry Heck. Outside knowledge visual question answering version 2.0. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Heather Riley and M. Sridharan. Integrating non-monotonic logical reasoning and inductive learning with deep learning for explainable visual question answering. *Frontiers in Robotics and AI*, 6, 2019.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges, 2021. URL `https://arxiv.org/abs/2103.11251`.

Mrinmaya Sachan. *Towards Literate Artificial Intelligence*. Phd thesis, Carnegie Mellon University, 2020. URL `https://doi.org/10.1184/R1/11898378.v1`.

Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023a.

Alireza Salemi, Mahta Rafiee, and Hamed Zamani. *Pre-Training Multi-Modal Dense Retrievers for Outside-Knowledge Visual Question Answering*. 2023b.

Sergio Sánchez Santiesteban, Sara Atito, Muhammad Awais, Yi-Zhe Song, and Josef Kittler. Improved image captioning via knowledge graph-augmented models. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4290–4294, 2024.

Raeid Saqur and Karthik Narasimhan. Multimodal graph networks for compositional generalization in visual question answering. 33, 2020.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023. URL `https://arxiv.org/abs/2302.04761`.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv preprint*, arXiv:2206.01718v1, 2022.

Henry Senior, G. Slabaugh, Shanxin Yuan, and L. Rossi. Graph neural networks in vision-language image understanding: A survey. *ArXiv*, abs/2303.03761, 2023.

Muralikrishnna G. Sethuraman, Ali Payani, Faramarz Fekri, and J. Clayton Kerce. Visual question answering based on formal logic. *arXiv preprint*, arXiv:2111.04785v1, 2021.

Sanket Shah, Anand Mishra, N. Yadati, and P. Talukdar. Kvqa: Knowledge-aware visual question answering. pp. 8876–8884, 2019.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. Unirag: Universal retrieval augmentation for multi-modal large language models. *ArXiv*, abs/2405.10311, 2024.

Xiangqing Shen, Yurun Song, Siwei Wu, and Rui Xia. Vcd: Knowledge base guided visual commonsense discovery in images. *arXiv preprint*, arXiv:2402.17213v1, 2024.

Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *arXiv preprint*, arXiv:2101.06013v1, 2021.

Weimin Shi, Denghong Gao, Yuan Xiong, and Zhong Zhou. Qr-clip: Introducing explicit knowledge for location and time reasoning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2024.

Zhan Shi, Yilin Shen, Hongxia Jin, and Xiao-Dan Zhu. Improving zero-shot phrase grounding via reasoning on external knowledge and spatial relations. pp. 2253–2261, 2022.

Keisuke Shirai, C. C. Beltran-Hernandez, Masashi Hamaya, Atsushi Hashimoto, Shohei Tanaka, Kento Kawaharazuka, Kazutoshi Tanaka, Yoshitaka Ushiku, and Shinsuke Mori. Vision-language interpreter for robot task planning. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2051–2058, 2023.

A. Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4601–4611, 2019.

Gursimran Singh. A bayesian approach to visual question answering. 2018.

Lingyun Song, Jianao Li, J. Liu, Yang Yang, Xuequn Shang, and Mingxuan Sun. Answering knowledge-based visual questions via the exploration of question purpose. *Pattern Recognit.*, 133:109015, 2022a.

Zijie Song, Zhenzhen Hu, and Richang Hong. Efficient and self-adaptive rationale knowledge base for visual commonsense reasoning. *Multimedia Systems*, 29:3017–3026, 2022b.

Zijie Song, Wenbo Hu, Hao Ye, and Richang Hong. How to use language expert to assist inference for visual commonsense reasoning. *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 521–527, 2023.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2018. URL `https://arxiv.org/abs/1612.03975`.

Wolfgang Stammer, Antonia Wüst, David Steinmann, and Kristian Kersting. Neural concept binder. *arXiv preprint*, arXiv:2406.09949v2, 2024.

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. Lvlm-interpret: An interpretability tool for large vision-language models, 2024. URL `https://arxiv.org/abs/2404.03118`.

Xin Su, Man Luo, Kris W Pan, Tien Pei Chou, Vasudev Lal, and Phillip Howard. Sk-vqa: Synthetic knowledge generation at scale for training context-augmented multimodal llms. *ArXiv*, abs/2406.19593, 2024.

Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. Learning visual knowledge memory networks for visual question answering. *arXiv preprint*, arXiv:1806.04860v1, 2018.

Sanjay Subramanian, Medhini Narasimhan, Kushal Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell, and Dan Klein. Modular visual question answering via code generation. *arXiv preprint*, arXiv:2306.05392v1, 2023.

J. Sung, Qiuyuan Huang, Yonatan Bisk, Subhojit Som, Ali Farhadi, Yejin Choi, and Jianfeng Gao. Ink : Intensive-neural-knowledge aligned image text retrieval. 2022.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint*, arXiv:2303.08128v1, 2023.

Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and F. Sun. Knowledge-based embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:11948–11960, 2021.

P. Torino, Elena Baralis, and Dott. Andrea Pasini. Semantics-aware vqa a scene-graph-based approach to enable commonsense reasoning. 2020.

D. Tsatsou, Konstantinos Karageorgos, A. Dimou, J. Rubiera, J. M. López, and P. Daras. Towards unsupervised knowledge extraction. 2021.

Ben Zion Vatashsky and Shimon Ullman. Understand, compose and respond - answering visual questions by a composition of abstract procedures. *arXiv preprint*, arXiv:1810.10656v1, 2018.

Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural-symbolic models for interpretable visual question answering. pp. 6428–6437, 2019.

Lara Verheyen, Jérôme Botoko Ekila, Jens Nevens, Paul Van Eecke, and Katrien Beuls. Neuro-symbolic procedural semantics for reasoning-intensive visual dialogue tasks. pp. 2419–2426, 2023.

P. Vickers, Nikolaos Aletras, Emilio Monti, and Loïc Barrault. In factuality: Efficient integration of relevant facts for visual question answering. pp. 468–475, 2021.

Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge. *arXiv preprint*, arXiv:2203.14499v1, 2022.

Mohammad Saif Wajid, Hugo Terashima-Marín, Peyman Najafirad, and M. A. Wajid. Deep learning and knowledge graph for image/video captioning: A review of datasets, evaluation metrics, and methods. *Engineering Reports*, 6, 2023.

Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, J. B. Tenenbaum, and Chuang Gan. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13384–13394, 2024a.

Bin Wang, Fuyong Xu, Peiyu Liu, and Zhenfang Zhu. Hypermr: Hyperbolic hypergraph multi-hop reasoning for knowledge-based visual question answering. pp. 8505–8515, 2024b.

Jianfeng Wang, Anda Zhang, Huifang Du, Haofen Wang, and Wenqiang Zhang. *Knowledge-Enhanced Visual Question Answering with Multi-modal Joint Guidance*. 2022a.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models, 2024c. URL `https://arxiv.org/abs/2406.14852`.

Peng Wang, Qi Wu, Chunhua Shen, A. Dick, and A. Hengel. Explicit knowledge-based reasoning for visual question answering. *ArXiv*, abs/1511.02570, 2015.

Peng Wang, Qi Wu, Chunhua Shen, A. Dick, and A. van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2413–2427, 2016.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Visually-augmented language modeling. *ArXiv*, abs/2205.10178, 2022b.

Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, and Alan Yuille. Compositional 4d dynamic scenes understanding with physics priors for video question answering. *arXiv preprint*, arXiv:2406.00622v1, 2024d.

Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. *arXiv preprint*, arXiv:2205.11501v2, 2022c.

Zhu Wang, Sourav Medya, and Sathya N. Ravi. Differentiable outlier detection enable robust deep multimodal analysis. *arXiv preprint*, arXiv:2302.05608v1, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL `https://arxiv.org/abs/2201.11903`.

Jiahui Wei, Zhixin Li, Jianwei Zhu, and Huifang Ma. Enhance understanding and reasoning ability for image captioning. *Applied Intelligence*, 53:2706–2722, 2022.

Haoyang Wen, Honglei Zhuang, Hamed Zamani, Alexander Hauptmann, and Michael Bendersky. Multimodal reranking for knowledge-intensive visual question answering. *ArXiv*, abs/2407.12277, 2024.

Zhang Wen and Yuxin Peng. Multi-level knowledge injecting for visual commonsense reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:1042–1054, 2021.

Weixi Weng, Jieming Zhu, Hao Zhang, Xiaojun Meng, Rui Zhang, and Chun Yuan. Learning to compress contexts for efficient knowledge-based visual question answering. 2024.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use, 2025. URL `https://arxiv.org/abs/2505.19255`.

Qi Wu, Chunhua Shen, Anton van den Hengel, Peng Wang, and Anthony Dick. Image captioning and visual question answering based on attributes and external knowledge. *arXiv preprint*, arXiv:1603.02814v2, 2016.

Sen Wu, Guoshuai Zhao, and Xueming Qian. Resolving zero-shot and fact-based visual question answering via enhanced fact retrieval. *IEEE Transactions on Multimedia*, 26:1790–1800, 2024a.

Shirley Wu, Shiyu Zhao, Qian Huang, Kexin Huang, Michihiro Yasunaga, V. Ioannidis, Karthik Subbian, J. Leskovec, and James Zou. Avatar: Optimizing llm agents for tool-assisted knowledge retrieval. *ArXiv*, abs/2406.11200, 2024b.

Xiaoqian Wu, Yong-Lu Li, Jianhua Sun, and Cewu Lu. Symbol-llm: Leverage language models for symbolic system in visual human activity reasoning. *arXiv preprint*, arXiv:2311.17365v1, 2023.

Zeyu Xi, Ge Shi, Xuefen Li, Junchi Yan, Zun Li, Lifang Wu, Zilin Liu, and Liang Wang. Knowledge guided entity-aware video captioning and a basketball benchmark. 2024.

Shouguan Xiao and Weiping Fu. Visual relationship detection with multimodal fusion and reasoning. *Sensors (Basel, Switzerland)*, 22, 2022.

Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu. Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning. pp. 3022–3029, 2021.

Jialiang Xu, Michael Moor, and Jure Leskovec. Reverse image retrieval cues parametric memory in multimodal llms. *arXiv preprint*, arXiv:2405.18740v1, 2024a.

Ruinian Xu, Hongyi Chen, Yunzhi Lin, and Patricio A. Vela. Sgl: Symbolic goal learning in a hybrid, modular framework for human instruction following. *arXiv preprint*, arXiv:2202.12912v1, 2022.

Wenjia Xu, Zijian Yu, Yixu Wang, Jiuniu Wang, and Mugen Peng. Rs-agent: Automating remote sensing tasks through intelligent agents. *arXiv preprint*, arXiv:2406.07089v1, 2024b.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Y. Fung, and Heng Ji. Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *ArXiv*, abs/2402.11943, 2024.

Dizhan Xue, Shengsheng Qian, and Changsheng Xu. Integrating neural-symbolic reasoning with variational causal inference network for explanatory visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2024.

Yibin Yan and Weidi Xie. Echosight: Advancing visual-language models with wiki knowledge. *ArXiv*, abs/2407.12735, 2024.

Jianwei Yang, Jiayuan Mao, Jiajun Wu, Devi Parikh, David Cox, J. Tenenbaum, and Chuang Gan. Object-centric diagnosis of visual reasoning. *ArXiv*, abs/2012.11587, 2020.

Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. pp. 5356–5362, 2019.

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, 2022. URL `https://arxiv.org/abs/2109.05014`.

Zhenyu Yang, Lei Wu, Peian Wen, and Peng Chen. Visual question answering reasoning with external knowledge based on bimodal graph neural network. *Electronic Research Archive*, 2023.

Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). *arXiv preprint*, arXiv:2401.08392v4, 2024.

Keren Ye, Mingda Zhang, and Adriana Kovashka. Breaking shortcuts by masking for robust visual reasoning. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3519–3529, 2021.

Kexin Yi, Jiajun Wu, Chuang Gan, A. Torralba, Pushmeet Kohli, and J. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. pp. 1039–1050, 2018.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint*, arXiv:1910.01442v2, 2019.

Chengxiang Yin, Zhengping Che, Kun Wu, Zhiyuan Xu, and Jian Tang. Multi-clue reasoning with memory augmentation for knowledge-based visual question answering. *ArXiv*, abs/2312.12723, 2023.

D. Yu, Jianlong Fu, Xinmei Tian, and Tao Mei. Multi-source multi-level attention networks for visual question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15:1 – 20, 2019.

J. Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. Cross-modal knowledge reasoning for knowledge-based visual question answering. *ArXiv*, abs/2009.00145, 2020.

Weijiang Yu, Haofan Wang, G. Li, Nong Xiao, and Bernard Ghanem. Knowledge-aware global reasoning for situation recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:8621–8633, 2023.

Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi R. Fung, Hao Peng, and Heng Ji. Craft: Customizing llms by creating and retrieving from specialized toolsets. *arXiv preprint*, arXiv:2309.17428v2, 2023a.

Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023b.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *arXiv preprint*, arXiv:1811.10830v2, 2018.

Huayi Zhan, Peixi Xiong, Xin Wang, Xin Wang, and Lan Yang. Visual question answering by pattern matching and reasoning. *Neurocomputing*, 467:323–336, 2021.

Chunbai Zhang, Chao Wang, Yang Zhou, and Yan Peng. Vikser: Visual knowledge-driven self-reinforcing reasoning framework. *arXiv preprint*, arXiv:2502.00711v1, 2025a.

Gengyuan Zhang, Yurui Zhang, Kerui Zhang, and Volker Tresp. Can vision-language models be a good guesser? exploring vlms for times and location reasoning. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 625–634, 2023a.

Jiawei Zhang. Graph-toolformer: To empower llms with graph reasoning ability via prompt augmented by chatgpt. *arXiv preprint*, arXiv:2304.11116v3, 2023.

Liyang Zhang, Shuaicheng Liu, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao. Rich visual knowledge-based augmentation network for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 32:4362–4373, 2020.

Shunyu Zhang, X. Jiang, Zequn Yang, T. Wan, and Zengchang Qin. Reasoning with multi-structure commonsense knowledge in visual dialog. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4599–4608, 2022a.

Xiaohan Zhang, Yan Ding, S. Amiri, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang. Grounding classical task planners via vision-language models. *ArXiv*, abs/2304.08587, 2023b.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14, 2023c.

Yichi Zhang, Jianing Yang, Jiayi Pan, Shane Storks, Nikhil Devraj, Ziqiao Ma, Keunwoo Peter Yu, Yuwei Bao, and Joyce Chai. Danli: Deliberative agent for following natural language instructions. *arXiv preprint*, arXiv:2210.12485v1, 2022b.

Yichi Zhang, Jianing Yang, Jianing Yang, Keunwoo Peter Yu, Yinpei Dai, Jiayi Pan, N. Devraj, Ziqiao Ma, and J. Chai. Seagull: An embodied agent for instruction following through situated dialog. 2023d.

Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1356–1365, 2021.

Yifeng Zhang, Ming Jiang, and Qi Zhao. Query and attention augmentation for knowledge-based explainable reasoning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15555–15564, 2022c.

Yifeng Zhang, Shi Chen, and Qi Zhao. Toward multi-granularity decision-making: Explicit visual reasoning with hierarchical knowledge. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2573–2583, 2023e.

Yifeng Zhang, Ming Jiang, and Qi Zhao. Grace: Graph-based contextual debiasing for fair visual question answering. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:272430309.

Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Y. Liu. Interpretable detection of out-of-context misinformation with neural-symbolic-enhanced large multimodal model. 2023f.

Zefan Zhang, Yi Ji, and Chunping Liu. Knowledge-aware causal inference network for visual dialog. *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 2023g.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities, 2025b. URL https://arxiv.org/abs/2410.17385.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *ArXiv*, abs/2309.07915, 2023a.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint*, arXiv:2303.10868v3, 2023b.

Yibiao Zhao. A quest for visual commonsense: Scene understanding by functional and physical reasoning. 2015.

Kaizhi Zheng, Jeshwanth Bheemanpally, Bhrigu Garg, Seongsil Heo Dhananjay, Sonawane Winson, Chen Shree, Vignesh S Xin, and Eric Wang. Sage: A multimodal knowledge graph-based conversational agent for complex task guidance. URL https://api.semanticscholar.org/CorpusID:266186657.

Wenfeng Zheng, Lirong Yin, Xiaobing Chen, Zhiyang Ma, Shan Liu, and Bo Yang. Knowledge base graph embedding module design for visual question answering model. *Pattern Recognit.*, 120:108153, 2021.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint*, arXiv:2307.13854v4, 2023.

He Zhu, Ren Togo, Takahiro Ogawa, and M. Haseyama. Multimodal natural language explanation generation for visual question answering based on multiple reference data. *Electronics*, 2023.

Yi Zhu, Xiwen Liang, Bingqian Lin, Qixiang Ye, Jianbin Jiao, Liang Lin, and Xiaodan Liang. Configurable graph reasoning for visual relationship detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33:117–129, 2020a.

Zihao Zhu. From shallow to deep: Compositional reasoning over graphs for visual question answering. *arXiv preprint*, arXiv:2206.12533v1, 2022.

Zihao Zhu, J. Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *ArXiv*, abs/2006.09073, 2020b.

M. Ziaeefard and F. Lécué. Towards knowledge-augmented visual question answering. pp. 1863–1873, 2020.