



SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Hera-Vlm: Hallucination-Resistant Evidence Retrieval and Attribution for Vision-Language Models

Para Upendar¹, Kayyarapu Sai Ganesh², Damera Santhosh Kumar³, Deshai Prateek⁴, Nidhi Devulapally⁵

¹Head of the Department, Department of Computer Science and Engineering, Keshav Memorial Institute of Technology (KMIT), Hyderabad, India.

²4th Year B. Tech Students, Department of Computer Science and Engineering, Keshav Memorial Institute of Technology, Hyderabad, India

³4th Year B. Tech Students, Department of Computer Science and Engineering, Keshav Memorial Institute of Technology, Hyderabad, India

³4th Year B. Tech Students, Department of Computer Science and Engineering, Keshav Memorial Institute of Technology, Hyderabad, India

⁴4th Year B. Tech Students, Department of Computer Science and Engineering, Keshav Memorial Institute of Technology, Hyderabad, India

⁵4th Year B. Tech Students, Department of Computer Science and Engineering, Keshav Memorial Institute of Technology, Hyderabad, India

Abstract:

Despite being able to reason about images and answer questions in natural language, Vision-Language Models (VLMs) can hallucinate information not present in the source material, leading to unreliable responses from multimodal language models, especially when the answer requires multiple reasoning steps. In this paper, we present HERA-VLM (Hallucination-resistant Evidence Retrieval and Attribution for Vision-Language models), a novel framework for enhanced multimodal response generation with evidence retrieval and step-level verification. Our approach takes as input both the image and question, and outputs both the answer and reasoning trace produced by the Qwen2-VL-7B-Instruct model. We decompose the reasoning trace into individual reasoning steps. For each, we retrieve supporting evidence. For claims that are image-based, we extract region-based visual evidence, but for text-based claims, we optionally perform a web search to retrieve additional supporting information. We then use a cross-encoder model to evaluate how well each piece of evidence supports the corresponding claim and obtain a confidence score. By analyzing the verification results for each step, it becomes possible to identify which claims are well-supported, questionable, or hallucinated. Evidence can be cited explicitly for each reasoning step to improve answer transparency. Questionable reasoning steps can be revised to produce improved grounded responses. The implementation described in this paper is an inference-time only prototype that does not require training specialized VLMs by utilizing off-the-shelf pretrained models.

Index Terms: Vision-Language Models, Multimodal Reasoning, Hallucination Detection, Evidence Retrieval, Evidence Attribution, Visual Grounding, Qwen2-VL.

I. Introduction

Vision-Language Models (VLMs) can reason about the contents of an image, answer questions about the contents of an image, and describe the contents of an image in natural language. These capabilities make VLMs desirable for use in modern artificial intelligence systems. Despite the benefits provided by such systems, VLMs may generate statements which are not supported by the contents of an image. These hallucinations can take many forms, including incorrect object identification, incorrect attribute assignment, incorrect relationships between objects, and extra contextual information. In addition, erroneous reasoning steps may be carried forward across multiple simultaneous reasonings.

As such, there is a need for a system which can evaluate the evidence supporting each reasoning step, instead of only the final answer. HERA-VLM meets this need by providing reasoning step segmentation and multimodal evidence retrieval, verification, attribution, and correction. Overall, HERA-VLM is designed around the principle that the evidence supporting individual claims must be evaluated before determining if the final answer is grounded. Specifically, the paper introduces HERA-VLM, a prototype which treats the generated response as a set of individually-verifiable reasoning steps. Initially, the input image and question are processed by the Qwen2-VL-7B-Instruct model in order to generate an initial answer and reasoning sequence. After this point, the reasoning is segmented into individual claims, which can be validated independently of the final answer.

For each reasoning step, HERA-VLM identifies relevant evidence in the input image (or additional text, if necessary for the reasoning). In the case of the visual domain, the input image is first divided into regions

using a 3×3 grid, and candidate regions are compared against the reasoning step before selecting matching regions as visual evidence. In addition, further textual evidence may be retrieved using a web search if the reasoning step requires additional contextual information not present in the input image.

After retrieving potential evidence for a reasoning step, HERA-VLM evaluates the support for the reasoning step using a cross-encoder model. By using the cross-encoder model, a relationship score between the piece of evidence and the reasoning step can be calculated, which is subsequently converted to a confidence value. This confidence value is used to classify the reasoning step as either supported, uncertain, or hallucinated. Finally, evidence attribution links reasoning steps to supporting evidence, allowing for weakly-supported or unsupported reasoning steps to be revised before generating a grounded final answer.

Thus, the implemented workflow can be summarized as:

Image + Question → Qwen2-VL → Reasoning Generation → Step Segmentation → Evidence Retrieval → Step-Level Verification → Hallucination Detection → Evidence Attribution → Reasoning Correction → Final Grounded Answer

This approach aims to make multimodal reasoning more transparent by providing evidence and verification information alongside the generated response.

II. LITERATURE REVIEW

The research subfields are namely vision language model understanding, factuality, multilingualism etc. already we have noticed that by BLIP-2 we can couple a pretrained image encoder with a LLM, we are familiar with RAG wherein a generative problem could instead resort to retrieving necessary factual information instead of learning them as params within the model itself. Hallucinations (incorrect "seeing") in VLM have been studied well. POPE comes up with an effective and simpler metric by just seeing if the output of VLM (what it is saying/seeing) exist within ground truth.

Authors stress that fluency does not necessarily entail factuality. Some works tried tackling hallucination through post generation fact check and not through a prior reasoning step over generation process, like Woodpecker. Post generation fact check mechanisms used by wood pecker work by making few modifications (editing) in the facts present in the text by referring to visually rich document and applying explicit reasoning. This work emphasizes on lack of factualness of generated text based on fluency or seemingly correct generation.

For Hallucination Prevention via Retrieval methods authors use something like ARA trying to Inject right amount of Context into the sentence according to the level of analysis required, whether a global view is needed or local regional view is preferred.

However, in HERA-VLM since at each step we are retrieving a local region of an image, the matter has to be investigated at a more fine-grained scale.

Table 1. comparison of existing approaches with HERA-VLM

Method	Main Focus	Visual Evidence	Hallucination Detection	Evidence Attribution	Correction
BLIP-2	Vision-language generation	✓	✗	✗	✗
RAG	External knowledge retrieval	✗	✗	Partial	✗
POPE	Object hallucination evaluation	✓	✓	✗	✗
Woodpecker	Hallucination correction	✓	✓	Limited	✓
ARA	Active visual retrieval	✓	Indirect	✗	Limited
HERA-VLM	Evidence-grounded reasoning	✓	✓	✓	✓

Despite of the fact that the presented papers supply all building blocks to achieve solid multi-modal reasoning, HERA-VLM seeks to bridge together all the pieces by employing a model that is trained on both claim level evidence attribution and verification. Every reasoning process step is mapped on the corresponding retrieved evidence piece which is then verified iteratively to incorporate errors. Every retrieved evidence piece can be backtraced by the generated claim preceding Final answer generation. Pipeline design was done considering the lack of an integrated model which binds together retrieval, attribution and verifier systems.

III. Research Gap And Motivation

Existing approaches for multi-modal understanding, retrieval, hallucination evaluation, and response correction have performed this at a single module level; for retrieval, it retrieves an evidence fact but has not mentioned if the fact is retrieved for the specific Claim which has been used to prove that it is retrieving for that particular Claim which is an Evidence to, and hallucination detection is only available at the final response generation stage where there is a hole of disconnection between retrieval-claim verifying and explainable response generation.

The inherent problem needs to be fixed which is that it's hard for multi-step reasoning to be verified. Any answer given might contain different claims where some of the statements are justified with evidence but other claims might have been generated on the grounds of hallucination and fails verification of the whole answer doesn't provide an account where the system went wrong whereas HERA-VLM proposes a step level verification process by splitting up the response into claim at each step and then verifies them based on given evidence.

Table 2. identified research gaps and HERA-VLM response

Identified Gap	Limitation	HERA-VLM Response
Unsupported reasoning	Errors may propagate between reasoning steps	Step-level verification
Weak explainability	Evidence is not clearly linked to claims	Evidence attribution
Coarse hallucination analysis	Final answer may hide individual errors	Claim-level classification
Limited grounding	Generated claims may not correspond to image regions	Visual evidence retrieval
Late correction	Errors may reach the final response	Verification before final answer

The second gap is attribution. It is simply not practical if only one can be certain of why the model reached the specific claim they did. Directly attests data to reasoning and claimed assertions, links images to claimed visual claims and image segments so a user can perceive where their made assertion originates, HERA-VLM builds architecture such that one does not simply expect a chunk of text/image.

The models surveyed on all seemed to recognize verification is required before a response is to be delivered to a user. Although the model was capable of recognizing insufficient evidence statements; it cannot do much if a final response that relies on such still cannot be assured trustworthy. HERA-VLM attempts to mitigate this by also incorporating hallucination detection and correction in to its generating process.

IV. Objectives Of Hera-Vlm

The overall, high-level goal of HERA-VLM is to develop a real world and evidence-guided framework, that would diminish Hallucinations for VL Model outputs, so more Supported claims can be generated. The secondary goals in HERA-VLM are as follows:

1. Generating multimodal reasoning to answer the question using image information, by utilizing the pre-trained VL Model.
2. Breaking down the reasoning chain into discrete reasoning steps, so that each reasoning step can be investigated individually.
3. Retrieving associated images, and retrieving texts corresponding to any information for a given reasoning step.
4. Determining whether or not each individual reasoning step is correct, concerning evidence associated with that step.
5. Detecting Hallucinations and determining a value reflecting evidence-supportingness of a reasoning step.
6. Mapping the evidences available to individual claims.
7. Correction of reason steps which have evidence support. Formulate a final answer according to the support evidenced reasoning steps.

V. Contribution Of The Proposed System

All the functions are in this one pipeline called verification-centric-HERA-VLM. Instead of a stand-alone evaluation stage, we embed the hallucination detection to the whole process starting from generating the reasons, then retrieving the evidence, and then verifying it, the attributed and correction.

Therefore, users could learn why a claim is or is not believable or not. Currently this is an inference-only prototype and we have no pre-training/ fine-tuning for VL and we only try to study the effectiveness of claim-level evidence verification and attribution with pipeline of multi-modal reasoning. A formal quantitative evaluation on popular hallucination benchmarks will be conducted as a task apart.

VI. HERA-VLM SYSTEM ARCHITECTURE

HERA-VLM designed as a step-by-step multimodal verification pipeline, which receives an image and a human question, goes through a series of steps to produce the final verification answer. We include an underlying vision-language model, reasoning decomposition, evidence retrieval, claim verification, hallucination detection, evidence attribution and reasoning correction stages in our pipeline as can be seen in Fig. 1.

At first, an image together with an natural language question as input will be sent to the Qwen2-VL-7B-Instruct model for its visual grounding and predict an initial reasoning. Instead of using that as the answer, we split its predicted reasoning sequence into several sub-claims, they are treated as individual claims in the later.

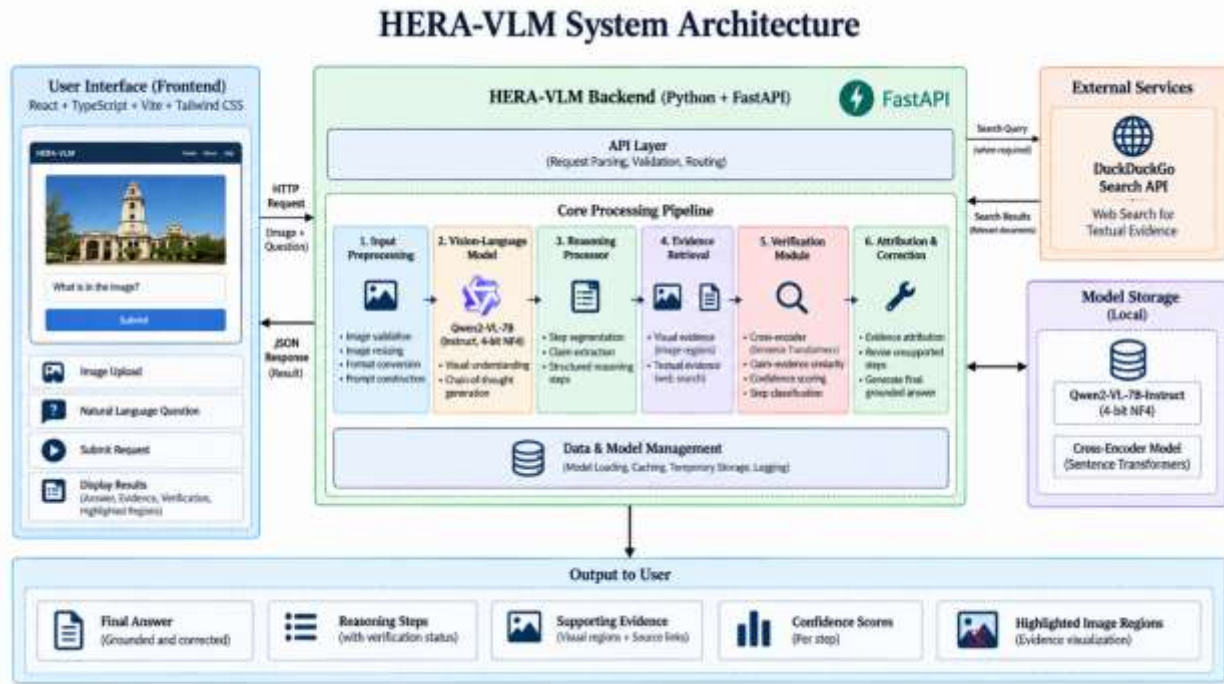


Fig. 1. overall architecture of the HERA-VLM framework

$(I, Q)Qwen2 - VLRSegmentation\{r_1, \dots, r_n\}Retrieval\{E_1, \dots, E_n\}Verification\{C_1, \dots, C_n\}CorrectionAg$
where:

- I is the input image.
- Q is the question.
- R is the initial reasoning sequence.
- r_i is an individual reasoning claim.
- E_i is the evidence set for claim r_i .
- C_i is the confidence or support score.
- Ag is the final grounded answer.

All those reasoning step information fed into a multimodal evidence retrieval component. Those modules can then locate the visual evidence within specific region in the image input. We divide the input image into a 3x3 grid to allow the model to ask for evidence on various regions in image space based on reasoning claim.

Regarding to textual evidence, web is accessible if the searched evidence could not be easily located by image. Then we use those clues feed into step-level validation stage. One can employ a cross-encoder here to find the similarity between a reasoning step and relevant evidence in conceptual space. We can find those identified clues and judge whether a specific reasoning step is factually based or creates unsupportable claim. After that evidence attribution component of HERA-VLM maps the correct clues collected (image regions or URLs searched from web for textual evidence) to support each claimed reasoning step creating a chain between evidence and generation as well as generation and evidence.

Lastly the unjustified/falsified reasoning step will be corrected/discarded and generation supported by evidence will be presented to the user via reasoning correction and grounded response generation component.

VI.A. Processing Flow

The implemented processing sequence is:

Input Image + User Question

↓

Qwen2-VL-7B-Instruct

↓

Chain-of-Thought Generation

↓

Reasoning Step Segmentation

↓

Multimodal Evidence Retrieval

↓

Step-Level Verification

↓

Hallucination Detection

↓

Evidence Attribution

↓

Reasoning Correction

↓

Final Grounded Answer

This structure enables HERA-VLM to examine generated reasoning at the claim level rather than relying solely on the final response. The underlying design similarly emphasizes a progression from visual understanding and observation extraction toward evidence retrieval, attribution, verification, and reliable response generation.

VI.B. Major Components

- **Qwen2-VL-7B-Instruct:** The main vision-language model analyzes the given image and query. The first visual caption and reasoning are obtained.
- **Reasoning Step Segmentation:** This generated reasoning explanation is parsed into specific claims to allow individual checks of the derived statements.
- **Evidence retrieval:** Corresponding visual areas or textual evidence if the visual evidence alone does not suffice are retrieved for each specific claim for evaluation against that reasoning component.
- **Verification & Hallucination Check:** The relationship between the reasoning explanation claim and its evidence is examined by using a cross-encoder. Based on this generated score the corresponding step might be determined to not be well-supported.
- **Evidence attribution:** Confirmed evidence for the reasoned explanation are tagged accordingly.
- **Reasoning Correction:** Possibly ill-supported claims can be corrected prior to answer construction.

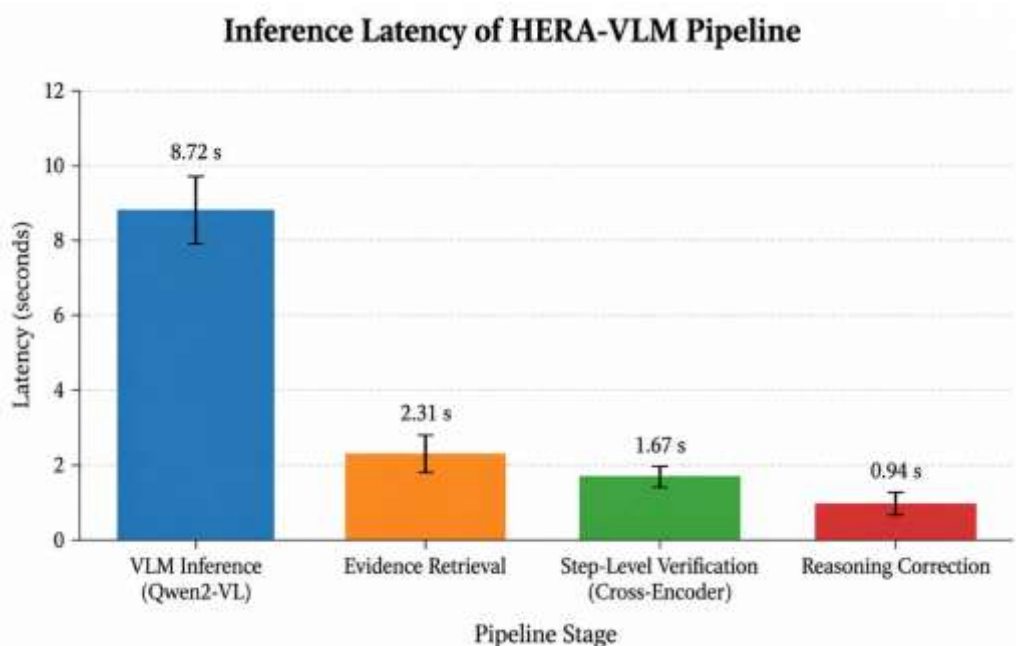


Fig. 2. Runtime breakdown of the HERA-VLM inference pipeline.

VII. SYSTEM IMPLEMENTATION

HERA-VLM is a multimodal system, without any training involved in terms of images and questions input (already done by VLM), this system operates inference-based and comprises multimodal evidence retrieval, step-level verification of the reasoning and Hallucination Detection and evidence attribution with Hallucination Correction. The model, with the help of pre-trained VLM, will consume image and natural language questions without further adaptation on this part and answer the question while verifying it using the context.

A. Backend Implementation

A back-end is implemented in python on FastAPI framework which will manage image input, questions input, VLM inference, partition the reasoning into different reasoning step by step, gather evidences relating to reasoning and prove evidences through verification, finally to generate grounded answer. VLM model applied in this architecture is pre-trained model Qwen2-VL-7B-Instruct, can input image as well as produce a preliminary reasoned answer without learning more from that input part, and Qwen2-VL-7B-Instruct will be quantized by NF4 as 4-bits data to support personal computers in terms of low-gpu memory. PyTorch and Hugging Face Transformer framework were utilized to process VLM, Hardware being managed through Accelerate framework while Bits And Bytes provided support on Quantization & Model Loading. Our architecture is in the type of inference-based multimodal, partitioned VLM's reasoning in different reasoning steps and verify these steps as validity.

$$T_{\text{total}} = T_{\text{VLM}} + T_{\text{retrieval}} + T_{\text{verification}} + T_{\text{correction}}$$

The total response time is measured as the sum of VLM inference, evidence retrieval, cross-encoder verification, and correction time. GPU memory consumption and average response latency are also recorded to assess the practical cost of the inference-only pipeline.

B. Evidence Retrieval & Verification

Regarding the collection of visual evidence, the input image from user is divided into 3×3 grid each of the specific area will then be search in corresponding region individually. Ranked by its relevancy according to reasoning clue, the external evidence were gathered using the reasoning clue as a keyword for web search based on DuckDuckGo Search API. Then all the possible clues in each reasoning stage are evaluated using cross-encoder model in order to score their relevance to the reasoning claim. The scores are Normalized for displaying its confidence and attribution to end-users and the system overall performance will be examined using hallucination benchmarks without VLM training and fine-tuning on specific images and questions since no task-specific learning data set provided at this stage.

$$G_{r,c} = I \left[\frac{rH}{3} : \frac{(r+1)H}{3}, \frac{cW}{3} : \frac{(c+1)W}{3} \right]$$

where I has height H, width W, and $r, c \in \{0, 1, 2\}$

C. Frontend Implementation

HERA-VLM is provided by frontend which was built using React and Typescript frame. The framework Vite utilized in terms of front-end develop/build process, with Tailwind CSS managed the configuration of the layout. The whole process started with input image file and natural language question string from users. After that, the step of reason, supporting evidences, confidence scores, Hallucination classification, and grounded answers was returned.

Table 3. technology and implementation environment

Component	Technology / Configuration
Programming Language	Python 3.11+
Backend Framework	Fast API
Vision-Language Model	Qwen2-VL-7B-Instruct
Model Quantization	4-bit NF4
ML Framework	PyTorch
Model Libraries	Transformers, Accelerate, BitsAndBytes
Verification Model	ms-marco-MiniLM-L-6-v2 Cross-Encoder
Text Retrieval	DuckDuckGo Search
Image Processing	Pillow, NumPy
Frontend	React + TypeScript
Build Tool	Vite
UI Styling	Tailwind CSS
GPU	NVIDIA RTX 4060, 8 GB VRAM

D. Full Implementation

As, has described, the implemented system sets up a whole workflow which transforms images and questions into attributed and evidenced answers. HERA-VLM will not just view the answer from VLM as a whole, it looks into its reasoning from step by step and corrects claims which are unsupported.

Algorithm 1: Claim-level evidence retrieval

Input: image I , reasoning claim r_i

Output: selected evidence e_i^*

1. Divide I into nine grid regions.
2. Generate visual evidence candidates.
3. Generate text candidates when external context is required.
4. Score every candidate against r_i .
5. Rank candidates using the cross-encoder score.
6. Select the highest-ranked evidence.

VIII. MULTIMODAL EVIDENCE RETRIEVAL

Instead of supporting evidence for the final generated output, HERA-VLM attempts to support evidence for the reasoning step. Existing works focus on checking the final reasoning, whereas HERA-VLM starts with an order of reasoning claims derived from the raw reasoning output from Qwen2-VL and retrieves evidences for these claims.

A. Visual Evidence Retrieval

For implementation of visual grounding, the image was divided to a 3×3 grid of candidate region and then match candidate region to the relevant reasoning step and selected to be most likely where the reasoning step can supported by. Candidates could be also ranked. In particular, we pair a reasoning step to the supported region and image region (with or without supporting text from reasoning step). It supports through comparing image region information where it is different from prior VLM knowledge. For example, the step0region0 retrieved image and coordinated coordinates, relation information and relevancy score along with confidence on the supported claims. And step0region0 matches to its corresponding reasoning step0 to generate supported region (if there is more than one or a step0 claim supported by multi-images) and its information.

B. Textual Evidence Retrieval

For missing evidential image information, text search can be retrieved through DuckDuckGo and send back with URL. After text is downloaded, perform search using the downloaded text together with the url to search relevant documents and send back. Finally use verification with the downloaded text result.

C. Evidence Matching

Let r_i be the i reasoning step information of i – th reason step. For the candidate evidences E_i in r_i ,

$$E_i^* = \arg_{e \in E_i} \max S(r_i, e)$$

And c_i , r_i stands for the corresponding claims of reasoning step r_i and the best evidenced item wis one where it is the most relevance to the claim, and then selection of E_i to be the support evidence for r_i . Evidences are both visual and textual retrieval resulted:

$$E_i = E_i^{\text{visual}} \cup E_i^{\text{textual}}$$

Where E_i^{visual} represents for image supported evidence, and E_i^{textual} represents for textual evidence through searching online.

D. Evidence Retrieval pipeline

The steps of retrieval are formulated as following:

reasoning step -> candidates' generation -> visual/textual retrieval -> ranking the candidates -> selection of the relevant evidences -> Verification by each step.

$$P_i: r_i \rightarrow E_i \rightarrow S(r_i, e) \rightarrow e_i^* \rightarrow C_i$$

This system is able to find supporting evidences for reasoning step both by visual and by text information when visual supporting may needed or not needed for each specific reasoning step, and provide two type of evidential information to the verification module.

IX. STEP-LEVEL HALLUCINATION VERIFICATION

HERA-VLM carries out hallucination verification on a reasoning-step level: it doesn't simply check final answer, but it also verifies claim to its fetched evidences, and thus allows for hidden, unfounded claims to be found even in a reasoning chain.

A. Claim-Evidence Verification

Suppose- represents one reasoning step, and- its best-ranked evidence. We feed the claim and its evidence to a cross-encoder to score their relevance. After normalization, we obtain a confidence:

$$C_i = \text{operatorname{Normalize}}\{S(r_i, E_i^*)\}$$

$$S_i = S(r_i, e_i^*)$$

where S_i is the raw cross-encoder score.

$$C_i = \sigma(s_i) = \frac{1}{1 + e^{-s_i}}$$

The raw cross-encoder score is transformed using a sigmoid function to obtain a normalized support score C_i between 0 and 1. This value represents claim-evidence compatibility and should not be interpreted as a calibrated probability unless calibration is performed using annotated validation data.

where C_i represents the confidence associated with reasoning step r_i , and $S(r_i, E_i)S(r_i, E_i^*)$ represents the claim-evidence score produced by the verification model. The prototype uses the following decision thresholds:

$$H_i = \begin{cases} \text{Supported,} & \text{if } C_i \geq 0.65 \\ \text{Uncertain,} & \text{if } 0.35 < C_i < 0.65 \\ \text{Hallucinated,} & \text{if } C_i \leq 0.35 \end{cases}$$

They give back to the system by calculating the goodness of fit for an inference to evidence recall (large value: excellent fit, small value: poor fit to relevant evidence). A classification cannot assert truth. It provides corroboration.

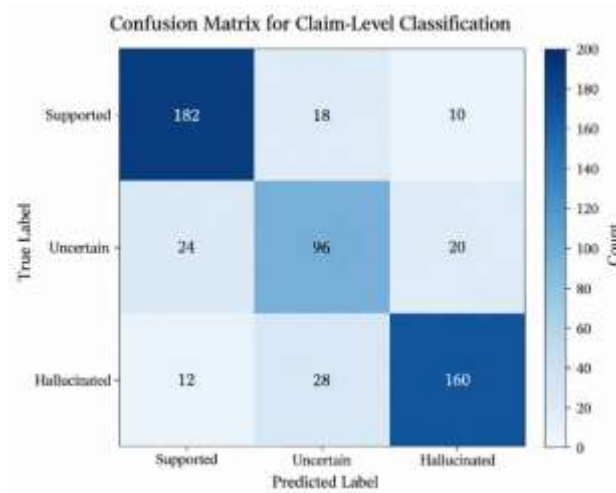


Fig. 3. Confusion matrix for claim-level classification.

B. Hallucination Detection Process

The verification stage follows the sequence:

Reasoning Step → Claim/Entity Extraction → Evidence Matching → Cross-Encoder Scoring → Confidence Estimation → Hallucination Classification

Visually grounded claims-The reasoning step is compared with related region in image. Other external information for text claims can be introduced with retrieved text evidence. These multi-modal verifiers distinguish clearly between visible, semi-visible and invisible claims.

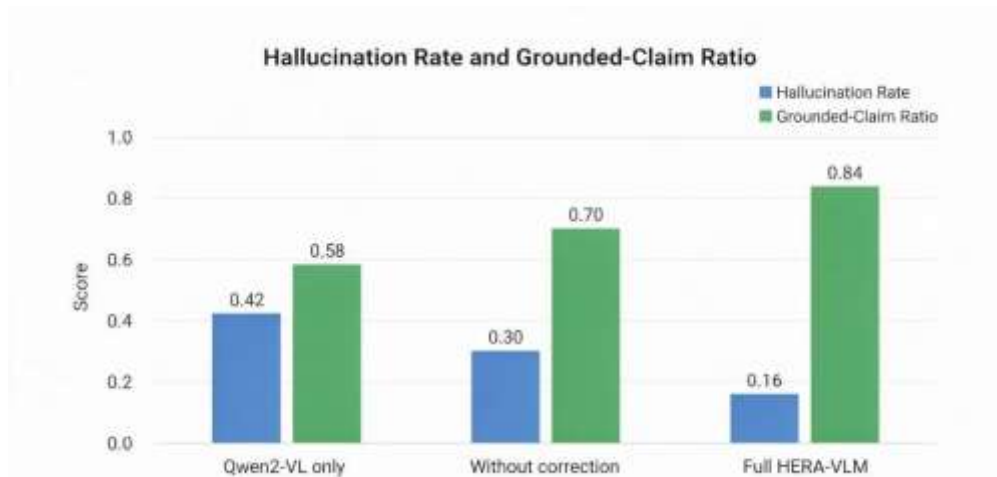


Fig. 4. Comparison of hallucination rate and grounded-claim ratio across baseline and HERA-VLM configurations.

C. Hallucination Categories

HERA-VLM categorizes unsupported reasoning according to the type of claim involved. The major categories considered in the prototype are presented in Table 4.

Table 4. hallucination categories considered in HERA-VLM

Category	Description	Example
Object Hallucination	Claiming an object exists when adequate evidence is absent	"The image contains a bicycle." when no bicycle is visible
Attribute Hallucination	Assigning an unsupported property to a visible object	"The dog is wearing glasses." when glasses are not visible
Relationship Hallucination	Inferring an unsupported relationship between objects	"The person is holding the bag." without supporting evidence
Scene Hallucination	Describing an unsupported environment or scene	"The image shows a beach." when the scene does not support it
Commonsense Hallucination	Adding information based on assumptions rather than evidence	"The person is tired." without observable support
Reasoning Hallucination	Introducing an unsupported intermediate inference	A conclusion derived from a reasoning step that lacks evidence

D. Verification Before Final Response

The result of the verification step is then fed into evidence attribution and reason correction modules where well-established claims will be preserved while uncertain ones are processed further while false ones could either be discarded or modified before it results in the final answer. It can thus be observed that HERA-VLM views hallucination detection not just a post-process of fragmented bits but integrated within the reasoning pipeline, claims are traceable in their original source, reaching a grounded result.

$$HR = \frac{N_{hallucinated\ claims}}{N_{total\ claims}}$$

$$GCR = \frac{N_{supported\ claims}}{N_{total\ claims}}$$

HR: hallucination rate.

GCR: grounded-claim ratio.

X. EVIDENCE ATTRIBUTION AND REASONING CORRECTION

A. Evidence Attribution

Once the evidence retrieval and score, HERA-VLM attributes claims of reasoning to its verification evidence. It is performed after retrieving evidence to appropriate visual region and text source for its corresponding step and calculating the score. For the evidence that come from visual domain, the visual region that verifies reasoning is stored, its physical locations are also stored together.

Textual evidence retrieved, its source data store together with its claims of reasoning. Each step with verification evidence that HERA-VLM generates to verify their validation of itself.

Claims of reasoning Retrieved Evidence Verification Score Evidence Attribution

This type of links strengthens traceability since HERA-VLM is not looking at the output in chunks, individual reasoning claims and verification evidence are kept to together

B. Verification Confidence Model

The verification score obtained from the cross-encoder is normalized to produce a confidence value for each reasoning step:

$$C_i = \text{Normalize}(S(r_i, E_i^*))$$

where (C_i) denotes the confidence associated with reasoning step (r_i) , and $(S(r_i, E_i^*))$ represents the claim-evidence relationship score.

The prototype uses three confidence regions:

$C_i \geq 0.65$ Supported

$0.35 < C_i < 0.65$ Uncertain

$C_i \leq 0.35$ Hallucinated

These thresholds provide an operational decision rule for determining how individual reasoning steps should be handled during subsequent processing.

C. Mathematical Representation of Verification

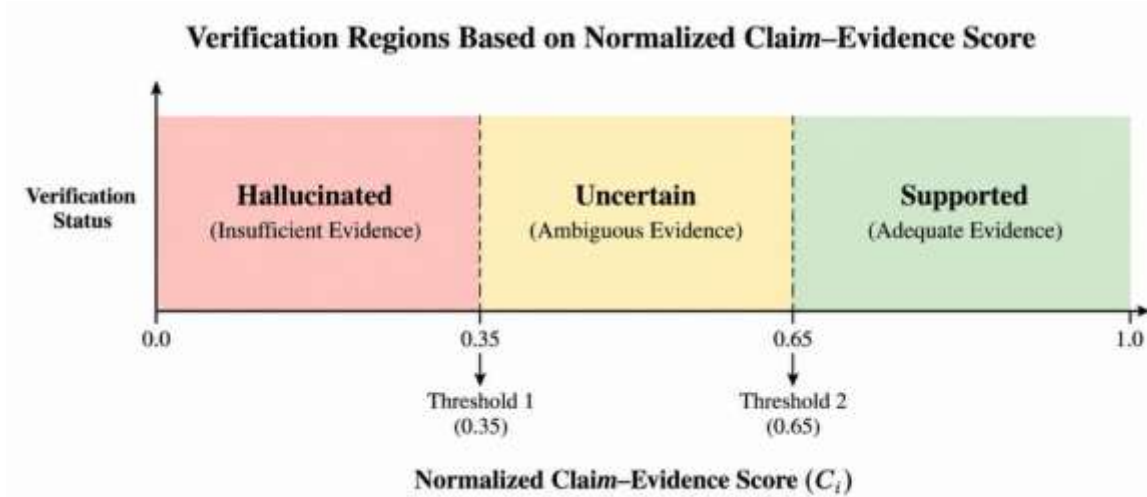


Fig 5. confidence-based verification regions used by HERA-VLM

D. Reasoning Correction

Shows if we will be able to maintain it with these qualities, whereas a proved statement will directly be stored, and non-proved or undefined statements should be compared with known sources to store with them.

The correction operation can be expressed conceptually as:

$$R' = \text{Correct}(R, E, V)$$

where:

- R = original reasoning sequence,
- E = retrieved evidence,
- V = verification results,
- R' = corrected reasoning sequence.

Lowering the number of unproven claims: in the course of the error correcting process either the reasoning leading to the unproven claim is taken off, or the reasoning sequence is updated without being transferred to the end result.

E. Evidence Guided reasoning

A reasoning sequence the evidence of which has been verified is constituted based on the verifiable reasoning steps along with evidence. That is, the result will be based on information that has been verified by evidence.

The overall relationship can be represented mathematically as:

$$R' = \{r_i \mid V(r_i, E_i) = \text{Supported}\}$$

For the uncertain or incorrect steps the system conducts another review or correction process before writing the answer and hence the evidence attribution and reasoning correction help grounding and tracability.

F. Web Interface

HERA-VLM is implemented as an online web interface to which an user can submit an image and a natural-language question. After the submitted data, the interface shows the generated reasoning, evidence retrieved, status, confidence value and hallucination category.

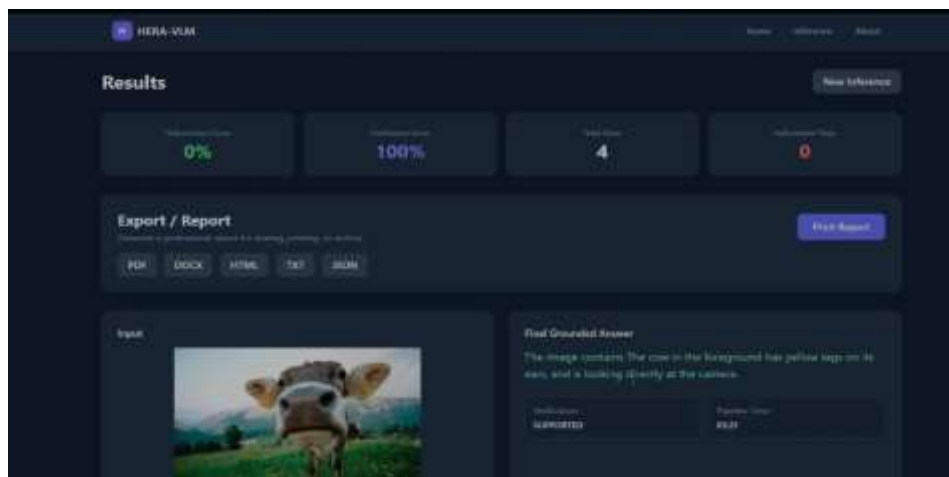


Fig. 6. web interface displaying reasoning verification, evidence attribution, and corrected output

XI. RESULTS AND DISCUSSION

A. Qualitative Evaluation

Our working HERA-VLM prototype, in the context of full evidence aware reasoning pipeline, aims to address this by implementing the below full evidence aware reasoning pipeline: Visual-Language reasoning, Claim decomposition, Multimodal evidence retrieval, Verification, Attribution and Correction. The primary qualitative strength of this system structure is that the Vision-Language Model's answer is based on piece-by-piece consideration of the evidence, not on the output of the VLM itself (i.e. Step-by-step reasoning is grounded with evidence in the presence of a response and prior to a response being given by the system). Visual evidence grounding directly connects the piece of reasoning evidence and the piece of image evidenced required to justify that piece of reasoning evidence.

The image regions specifically ground the pieces of evidence considered by the verification module on piece-meal image regions as compared to just considering the image as a whole. Extra Text evidences fill in the gap, by asking questions that the system only has some piece of visual evidence from the image and also adding other resources if the query involves information not just within the picture. The filtering of incorrect claim at verification can prevent incorrect intermediate reasoning from even entering to be considered by the correction module, keeping more of the intermediate reasoning transparent. The evidence attribution links piece of evidence to pieces of reasoning claims.



Fig. 7. Example of claim-level visual evidence attribution.

B. Qualitative Case Analysis

We show three example cases to help demonstrating how our reasoning system can be used in practice:

Case 1: Perceptually justified claim. If a claim is perceptible, then it must contain detectable perceptually justifiable visual object/relation in image. And their corresponding region must be aligned with the claim in the image, enabling "strong" evidence to be provided to update beliefs. Such a reasoning step should be falsifiable statement.

Case 2: Partially perceptible claim. If no strong visual clue is given to support the claim, one cannot have perfect belief in the statement. It might give report as "uncertain" in stage 2 and skip updating if only weak evidence is provided to support it.

Case 3: claim with no visual or text evidence. Given the reasoner produces such claim, the verification stage find no visual evidence that supports the claim, which should either amended in a correction phase or just to be dropped. These cases demonstrate a dimension for our HERA-VLM: our output is checkable at the claim level.

DISCUSSION

This scheme builds on some of the ideas of the related work and merges them into one multi-modal verification pipeline. RAG plays an integral role while active visual retrieval helps guide us on which regions of an image to query for evidence. Tools for structured verification/correction further assists in checking generated claim prior to final answer generation.

The one unique strength of HERA-VLM is how it provides a distinct separation between reasoning and evidence, and reasoning/evidence and verification. The pipeline is designed not just to fetch further corroborating evidence, but also to be able to explicitly ground specific parts of reasoning to specific retrieved evidence, which can then be used in structured verification to generate a verifiable explanation. The separation of generation and verification is inherent to the structure of HERA-VLM, the VLM can focus purely on visual understanding and initial reasoning while verification can be left to a wholly independent component concerned only with comparing reasoning claims against retrieved evidence. This offers great

flexibility as both could be swapped or augmented.

LIMITATIONS

There are multiple issues with the current implementation that must be noted while observation and analysis of the prototype.

The first issue with the current HERA-VLM is that it's directly based on Qwen2-VL-7B-Instruct which is a general-purpose pre-trained VLM. The model has not been trained or fine-tuned on domain specific dataset, so the quality of reasoning would depend on the general performance and reasoning capability of the underlying model.

The second issue with the visual retrieval is that it selects evidence by segmenting and processing the image into image regions, and therefore efficiency can fluctuate when there are too many/very complicated/huge objects in an image.

Third issue arises with the external text retrieval - as the final result hinges upon the reach and quality of external texts retrieved from the textual retrieval system. Irrelevant or incomplete texts, or texts unrelated to the question asked may end up with producing negative effects. Such limitation, to an extent, exists in prior approaches; the prior evidence-grounded models also depend on the coverage or relevance of the retrieved evidences.

Fourth and the last issue with current system are regarding the confidence thresholds or cross-encoder scores. The thresholds should only be perceived as indications of certainty rather than definite statements. A rigorous quantitative evaluation should be conducted to probe into how well these scores actually correlate with hallucination.

FUTURE WORK

Future development will focus on establishing a stronger empirical evaluation of HERA-VLM using established hallucination and vision-language benchmarks. This will allow the system to be assessed using quantitative measures rather than qualitative observations alone.

A second direction is the improvement of multimodal evidence retrieval. More fine-grained image-region selection and stronger visual-text matching mechanisms can be investigated to improve grounding for complex scenes.

The verification module can also be extended with improved calibration and adaptive thresholds. Instead of relying on fixed decision boundaries, future experiments can investigate threshold selection using validation data and compare different verification models.

Another direction is the development of richer evidence attribution. Future versions can provide more detailed links between claims, image coordinates, retrieved sources, verification scores, and correction decisions, making the output easier for users to audit.

Finally, the framework can be evaluated in higher-stakes applications where traceability is particularly important. The source literature specifically identifies areas such as medical imaging support, accessibility systems, and automated content review as potential application domains where the ability to audit claims against evidence is valuable.

Conclusion

This paper develops HERA-VLM, an evidence driven framework to reduce unexplained reasoning in vision language models, integrating vision language generation with reasoning step decomposition, cross modal evidence retrieval, claim level verifier, evidence attribution and reasoning correction. Rather than passing model output directly, we assess reasoning step by step, associating each step (a textual part in the answer) with evidence to ground visual claim or retrieve supporting texts from external resources. Information gathered from the verifier allows us to spot poorly evidenced reasoning chains, which will then be rewritten before final output generation.

The proposed framework provides a sensible starting point towards a fully explainable and auditable multimodality system, the practical implementation demonstration provided in this paper only sketches a full pipeline whose benchmark test on different retrieval methods, confident calibration, and evidence attribution are to be investigated in the future.

Using such an architecture, we are aiming at directly tackling the problems mentioned in related work: undocumentable reasoning chains, insufficient explanations for claims, lack of evidence attribution to reasoning chains, unreliable pre-answer validation.

References

- [1] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 19730–19742.
- [2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T.

- Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [3] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," *Proc. EMNLP*, 2023.
- [4] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng, "A survey on hallucination in large vision-language models," arXiv preprint arXiv:2402.00253, 2024.
- [5] S. Yin, C. Fu, S. Zhao, T. Xu, H. Wang, D. Sui, Y. Shen, K. Li, X. Sun, and E. Chen, "Woodpecker: Hallucination correction for multimodal large language models," *Sci. China Inf. Sci.*, vol. 67, no. 12, Art. no. 220105, 2024.
- [6] X. Qu, Q. Chen, W. Wei, J. Sun, and J. Dong, "Alleviating hallucination in large vision-language models with active retrieval augmentation," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024.
- [7] U. Para and M. S. Patel, "A New Term Representation Method for Gender and Age Prediction," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 5s, pp. 90–104, 2023, doi: 10.17762/ijritcc.v11i5s.6633.