

UniTor at EVWSD-ITA: Zero-Shot Visual Word Sense Disambiguation via Visual Question-Answering

Claudio D. Hromei¹, Antonio Scaiella², Danilo Croce¹ and Roberto Basili¹

¹Department of Enterprise Engineering, University of Rome Tor Vergata
Via del Politecnico 1, 00133, Rome, Italy

²Reveal S.r.l. - Via Kenia 21, 00144, Rome, Italy

Abstract

Visual Word Sense Disambiguation (VWSD) requires selecting, among ten candidates, the image that matches the intended sense of an ambiguous word given minimal context. We present the UniTor system for EVWSD-ITA at EVALITA 2026, a fully zero-shot generate-and-verify pipeline that combines: (i) LLM-based generation of discriminative positive/negative visual questions and a holistic target description, (ii) VQA-based verification over each candidate image with token-level confidence aggregation to score local visual evidence, and (iii) CLIP-based global matching against both the original query and the synthetic description. The final ranking is obtained by fusing local verification and global similarity signals. Experiments on the official benchmark show that our approach substantially improves over a CLIP-only baseline and is particularly effective at filtering hard negatives such as co-hyponyms and alternative senses.

Keywords

Visual WSD, Multi Modality, LLMs, Zero-Shot VQA

1. Introduction

The interaction between Natural Language and Visual Perception represents one of the long-standing goals of Artificial Intelligence [1, 2]. While humans effortlessly map linguistic concepts to visual objects, a process known as *symbol grounding* [3], machines often struggle when language becomes ambiguous. A word like “bank” can refer to a financial institution or a riverside; similarly, in Italian, the word “gomma” can denote an eraser, a car tyre, or simply a material. Visual Word Sense Disambiguation (VWSD) addresses this challenge: given a target word and a limited linguistic context, the system must identify the correct visual depiction among a set of candidates [4].

This paper describes the participation of the **UniTor** team in the **EVWSD-ITA** task at EVALITA 2026 [5]. The task extends the VWSD formulation (as introduced at SemEval2023 [6]) by introducing a high degree of visual complexity. Systems are provided with a target word, a hypernym, and a gloss fragment derived from BabelNet [7], and must retrieve the correct image from a pool of 10 candidates. Crucially, the candidate set includes not only unrelated images but also *hard negatives*: images representing different senses of the same word (polysemy) or visually similar objects from the same category (co-hyponyms).

Consider the example in Figure 1. The input query is “attrezzo gomma cancelleria” (en: “*tool rubber stationery*”). A standard dual-encoder model like CLIP [8], which relies on global visual-semantic similarity, might correctly discard a “tyre” (polysemous distinction) but fail to distinguish the “eraser” from a “pencil” or “ruler” (co-hyponyms), as they all share the visual embedding space of “stationery objects”. Most state-of-the-art approaches tackle this by fine-tuning massive encoders on task-specific data [9, 4]. However, this requires significant computational resources and often lacks interpretability.

We propose a different direction: a fully **Zero-Shot**, training-free architecture that re-frames disambiguation as a *Visual Constraint Verification* process. Inspired by recent advancements in Multimodal

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 - 27, Bari, IT
✉ hromei@ing.uniroma2.it (C. D. Hromei); scaiella@revealsrl.it (A. Scaiella); croce@info.uniroma2.it (D. Croce); basili@info.uniroma2.it (R. Basili)

ORCID 0009-0000-8204-5023 (C. D. Hromei); 0000-0001-9111-1950 (D. Croce); 0000-0001-5140-0694 (R. Basili)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

gomma attrezzo cancelleria

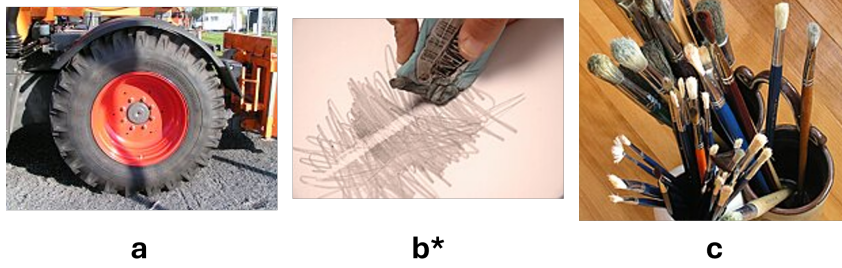


Figure 1: Running Example of the EVWSD-ITA Task. The input query is the scrambled triple “attrezzo gomma cancelleria” (en: “tool, rubber, stationery”). For simplicity, we report here only 3 images instead of 10 as in the task. The system must distinguish the *Target Image* (it:gomma, en:rubber, i.e. an eraser) from a *Polysemous Distractor* (a tyre) and a *Hard Negative Co-hyponym* (a pencil). Standard VLM embeddings often confuse the Target with the Hard Negative due to visual similarity. The Gold Standard image is noted with *.

Large Language Models (MLLMs) [10, 11, 12], our pipeline synergises the reasoning capabilities of LLMs with the discriminative power of Visual Question Answering (VQA). Instead of simply embedding the query, we use an LLM (GPT-5.1) to “explode” the ambiguous input into a set of discriminative requirements, formulated as Positive and Negative questions. A VQA model, specifically Qwen3-VL-8b [13], then acts as an auditor, verifying these constraints against each candidate image.

As an example, a simple way of distinguishing the *eraser* (image b from Figure 1) from the *tyre* (image a from Figure 1) would be asking a question that addresses its usage, such as “*L’oggetto è un grande anello nero con battistrada montato su una ruota?*” (en: “Is the object a large black ring with treads mounted on a wheel?”). On the other hand, in order to discriminate a hard negative, such as the *brushes* (image c from Figure 1), a very precise question, such as “*L’interazione primaria con questo oggetto prevede la rimozione di tratti grafici esistenti?*” (en: “Does the primary interaction with this object involve removing existing graphic features?”) will address the key difference between the two.

Our contribution is threefold: (1) We introduce a novel pipeline that disentangles linguistic reasoning from visual perception; (2) We demonstrate that generated “Negative Questions” are crucial for filtering hard negatives; (3) We show that this approach achieves competitive performance without any parameter updates. Specifically, we address the following Research Questions (RQs):

1. **RQ1:** Can the EVWSD problem be effectively modelled as a *Visual Constraint Verification* task, where the final ranking depends on the distinct satisfaction of linguistic requirements (“accounting”) rather than just embedding similarity?
2. **RQ2:** Can LLM-generated questions effectively filter hard negatives (e.g., co-hyponyms) in a zero-shot setting, surpassing standard global matching?
3. **RQ3:** Does the injection of a synthetic holistic description (generated by the LLM) provide a more robust context for global matching compared to the raw query?
4. **RQ4:** How does the ensemble of local feature verification (VQA) and global semantic matching (CLIP) impact the final ranking?

In the remainder of this paper, Section 2 reviews the State of the Art in Visual Word Sense Disambiguation, focusing on previous shared tasks and the limitations of current multimodal encoders. Section 3 details our proposed methodology, describing the five-stage pipeline from semantic analysis to multi-view fusion. Section 4 presents the experimental setup, reports the results achieved on the EVWSD-ITA dataset, and discusses the ablation study concerning the structured query parsing. Finally, Section 5 draws conclusions and outlines future research directions.

2. State of the Art

The task of Word Sense Disambiguation (WSD) has long been a cornerstone of Natural Language Processing, aiming to identify the correct meaning of a polysemous word within a given context. With the advent of large-scale multimodal knowledge bases such as **BabelNet** [7], this challenge has naturally extended to the visual domain. In BabelNet, concepts are organised into *synsets* (sets of synonyms sharing a common meaning), which are often associated with visual depictions. In this landscape, Visual Word Sense Disambiguation (VWSD) requires a system to map a linguistic query to the image that best represents the corresponding synset, distinguishing it not only from completely unrelated concepts but, more critically, from *co-hyponyms*, distinct concepts that share the same hypernym (ancestor) and possess similar visual features.

The complexity of this task was formally benchmarked in the **SemEval-2023 Task 1** on Visual Word Sense Disambiguation [4]. The shared task highlighted that the most successful approaches relied heavily on massive ensembles of Vision-Language Models (VLMs) or on fine-tuning state-of-the-art encoders like CLIP [8] on task-specific data. While effective, these strategies require significant computational resources and extensive training data, often overfitting to the specific visual domains of the training set (e.g., ImageNet-like pictures).

A significant limitation of standard dual-encoder models, such as CLIP, is their tendency to function as “visual bag-of-words” models. Research has shown that while these models excel at global semantic matching, they struggle with compositional understanding and syntax [14]. For instance, CLIP might align the query “gomma” (eraser) with an image of a “pencil” simply because both frequently appear in the context of “stationery”, failing to capture the fine-grained visual constraints that distinguish the two objects. This lack of granular reasoning makes them prone to errors when facing hard negatives, such as co-hyponyms.

To mitigate these limitations, recent works in Multimodal Learning have explored the use of Large Language Models (LLMs) to enhance visual reasoning. Techniques such as *Chain-of-Thought* (CoT) prompting [15] have been adapted to generate intermediate reasoning steps or to expand short queries into detailed visual descriptions [16, 17, 18, 19]. Similarly, Visual Question Answering (VQA) [20] has been proposed not just as a standalone task, but as a proxy to probe specific visual attributes in zero-shot classification scenarios.

Our approach distinguishes itself from the existing literature by integrating these components into a unified **Zero-Shot** pipeline that treats the input query not as a flat text string, but as a composite semantic object. Unlike previous methods that directly map the noisy query to the image space, we leverage the LLM to disentangle the semantic signals (lemma, hypernym, and gloss) and generate explicit verification constraints. By combining the generative power of LLMs with the discriminative precision of VQA, we address the lack of fine-grained reasoning in standard VLMs without the need for resource-intensive fine-tuning.

3. Visual Word Sense Disambiguation through Visual Question Answering

Our approach to the EVWSD-ITA task is grounded in the intuition that visual disambiguation requires a verify-then-match strategy. Rather than relying solely on global embedding similarities, we decompose the complex linguistic input into granular visual constraints. The proposed architecture is a five-stage pipeline that synergises the generative reasoning of Large Language Models (LLMs) with the discriminative verification of Vision-Language Models (VLMs). Figure 2 illustrates the five-stage pipeline adopted in our approach. The process begins with the **Input**, consisting of the scrambled linguistic triple and the set of candidate images. In **Phase 1**, the Large Language Model analyses the query to produce a structured JSON object containing a *Holistic Description* and a set of *Positive and Negative Questions*. The workflow then splits into two parallel branches:

- The **Local Verification Branch** (Phases 2 & 3), where the VQA model answers the generated

questions against the candidate images, producing the confidence score S_{VQA} via geometric mean aggregation.

- The **Global Matching Branch** (Phase 4), where the CLIP encoder computes the similarity of the image against both the original query ($S_{Original}$) and the generated holistic description ($S_{Synthetic}$).

Finally, in **Phase 5**, these three distinct signals are fused to compute the final score used to rank the images.

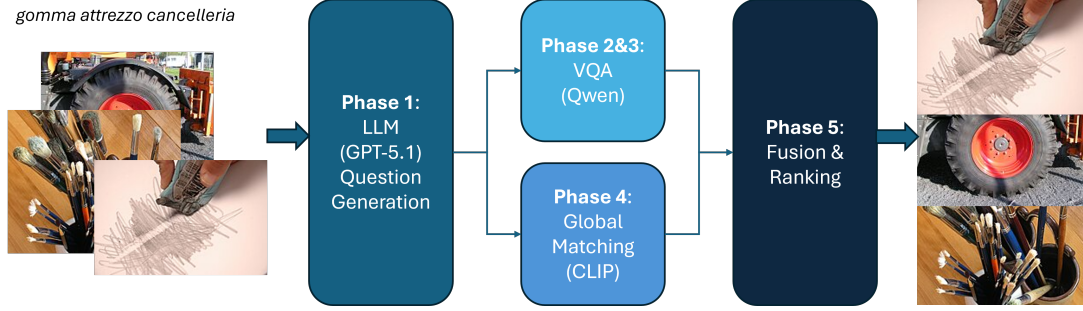


Figure 2: Overall workflow of the UniTor system. The pipeline processes the input query through an LLM (Phase 1) to generate discriminative questions and a holistic description. These outputs condition the visual verification performed by the VQA model (Phase 2 & 3) and the global matching performed by CLIP (Phase 4). Finally, the local and global scores are fused to produce the ranking (Phase 5).

3.1. Phase 1: Semantic Analysis and Weighted Question Generation

The first phase aims to disambiguate the input context via explicit semantic analysis to generate targeted visual checks. The raw input, a scrambled sequence of a lemma, a hypernym, and a gloss fragment (“gomma attrezzo cancelleria”), contains noisy signals that must be disentangled before visual matching can occur, i.e. we must first identify which is the main concept to focus on.

We employ **GPT-5.1** acting as a “*Computational Linguist*” to analyse this input fragment. Through a structured prompt, the model is instructed to perform a key reasoning step: identify the concrete *Visual Target* (the specific synset, i.e. “gomma”) implied by the gloss (i.e. “cancelleria”) and explicitly list potential *Semantic Distractions*, including specific co-hyponyms (e.g., distinguishing a “pencil” from an “eraser”) and alternative polysemous senses (e.g., distinguishing a “car tyre” from an “eraser”).

The output is a structured JSON object containing:

1. **Target Definition:** A synthetic, holistic description of the identified visual concept, which serves as a clean, expanded query for global matching. For the example in Figure 1: “*Piccolo oggetto solido in materiale gommoso o vinilico, impugnabile e sfregato sulla carta per rimuovere segni di matita o, in alcuni casi, di penna.*” (en: “A small, solid object made of rubber or vinyl that can be held in the hand and rubbed on paper to remove pencil marks or, in some cases, pen marks.”).
2. **Distractors List:** The identified entities to be excluded (e.g. *pencil*, *car tyre*, etc. in the case of Figure 1).
3. **Weighted Questions:** A set of 10 binary questions, divided into 5 *Positive Questions* (capturing the necessary visual features of the target) which will strengthen the visual signal of the correct image and 5 *Negative Questions* (specifically designed to rule out the visual features of the identified distractors) which will lower the visual signal of the wrong images.
4. **Importance Weights:** Inspired by [21], where the authors demonstrate that asking for a probability to the language model improves its generation variability, each question is associated with a generated weight $w \in [0.5, 1.0]$, representing its estimated discriminative power in the disambiguation process. This refinement enabled the production of a broader spectrum of questions, ensuring greater semantic variability compared to the simple version.

3.2. Phase 2: Visual Assessment

In the second phase, we verify whether candidate images satisfy a set of generated visual constraints. We employ **Qwen3-VL-8b**, a relatively small state-of-the-art Visual Question Answering (VQA) model, in a zero-shot setting. The model iterates through the set of $K = 10$ candidate images provided by the task. For each image, it answers all 10 generated questions. Crucially, we do not restrict the model to a binary text output; instead, we instruct it via prompt to answer exclusively “Yes” or “No” and access the underlying logits. We define the probability of the “Yes” token as $p(\text{Yes})$ and set $p(\text{No}) = 1 - p(\text{Yes})$. Although the model technically remains free to generate arbitrary sequences, in which case we would default to $p(\text{Yes}) = p(\text{No}) = 0.5$, this deviation was not observed in our experiments.

3.3. Phase 3: VQA Confidence Estimation

In order to determine a robust measure of the visual assessment above, we map the VQA output to a numerical score based on token probabilities. For each image I and each question q_j in the set of generated questions $Q(I)$, we query the VQA model and extract the probability of the first generated token. Let $p_{\text{YES}}^{(j)}$ be the probability associated with the “YES” tokens.

We define the local score s_j for the j -th question depending on its expected polarity (i.e., whether the image *should* or *should not* exhibit the feature):

$$s_j = \begin{cases} p_{\text{YES}}^{(j)} & \text{if } q_j \in Q(I)_{\text{positive}} \\ 1 - p_{\text{YES}}^{(j)} & \text{otherwise} \end{cases} \quad (1)$$

Finally, the global VQA score S_{VQA} for the image is computed as the geometric mean of the local scores. We explicitly chose the geometric mean (computed via log-space summation) over the arithmetic mean to penalise images that fail even a few critical questions (e.g., an image that looks like the target but fails a specific negative check against a co-hyponym). The score is calculated as follows, where $N = 10$ is the total number of questions:

$$S_{\text{VQA}} = \exp \left(\frac{1}{N} \sum_{j=1}^N \log(s_j) \right) \quad (2)$$

For the example in Figure 1, we show here a couple of questions with their answers and scores for each image:

1. **Image A, final score 0.03:**

- **Positive Question:** “L’oggetto è pensato per rimuovere segni da un foglio di carta?” (en: “Is the object designed to remove marks from a sheet of paper?”) → No (confidence 0.94 → score $1 - 0.94 = 0.06$)
- **Negative Question:** “L’oggetto è un grande anello nero con battistrada montato su una ruota?” (en: “Is the object a large black ring with treads mounted on a wheel?”) → Yes (confidence 0.97 → score $1 - 0.97 = 0.03$)

2. **Image B, final score 0.69:**

- **Positive Question:** “L’oggetto è pensato per rimuovere segni da un foglio di carta?” (en: “Is the object designed to remove marks from a sheet of paper?”) → Yes (confidence 0.99 → score 0.99)
- **Negative Question:** “L’oggetto è un grande anello nero con battistrada montato su una ruota?” (en: “Is the object a large black ring with treads mounted on a wheel?”) → No (confidence 0.77 → score 0.77)

3. **Image C, final score 0.47:**

- **Positive Question:** “L’oggetto è pensato per rimuovere segni da un foglio di carta?” (en: “Is the object designed to remove marks from a sheet of paper?”) → No (confidence 0.64 → score $1 - 0.64 = 0.36$)
- **Negative Question:** “L’oggetto è un grande anello nero con battistrada montato su una ruota?” (en: “Is the object a large black ring with treads mounted on a wheel?”) → No (confidence 0.89 → score 0.89)

3.4. Phase 4: Multimodal Semantic Matching

While the VQA module handles fine-grained details, we maintain a global perspective to capture the general “vibe” and semantic composition of the image. We use **CLIP** to compute the cosine similarity between the image embeddings and the text embeddings. To maximise robustness, we compute this similarity from two different linguistic views:

- $S_{Original}$: The similarity between the candidate image and the **Original Query** (the input sequence provided by the competition for the image I). This represents the noisy but grounded truth. As an example from Figure 1, CLIP produces the following scores with the original query: A 0.26, B 0.29, C 0.27, which implies a very poor similarity, even with the original image, but still a positive one.
- $S_{Synthetic}$: The similarity between the candidate image and the **Generated Holistic Description** produced in Phase 1. This represents a cleaner, richer, and syntactically well-formed description of the target concept, which usually contains different and more general words to describe the concept, even multi-faceted. For the example in Figure 1, CLIP produces the following scores with the synthetic description: A 0.15, B 0.39, C 0.33, which implies a better discrimination between the correct and wrong images.

3.5. Phase 5: Multi-Channel Fusion and Ranking

The final decision is produced by fusing the local verification score with the global semantic matching scores. The final score for each image can be computed as an unweighted sum of different combinations of the components, as determined empirically during the development phase (see Section 4):

$$Score_{Final} = S_{VQA} + S_{Original} + S_{Synthetic} \quad (3)$$

The candidate images are sorted by $Score_{Final}$ in descending order, and the top-ranked image is selected as the answer.

4. Results and Discussion

In this section, we present the experimental evaluation of the proposed architecture. We first detail the setup and the tuning process performed on the development set, then report the official results achieved in the EVWSD-ITA task, discuss an ablation study concerning the impact of structured query parsing, and finally present an error analysis over a borderline example.

4.1. Experimental Setup

Our pipeline relies on the following components:

- **Reasoning Engine:** We utilised **GPT-5.1**¹ for the semantic analysis and question generation phases.
- **VQA Model:** We employed **Qwen3-VL-8b**², a relatively small state-of-the-art vision-language model, to perform the zero-shot verification of the generated questions.

¹Accessed via OpenAI API in December 2025

²<https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct>

- **Vision-Language Encoder:** For global semantic matching, we used CLIP³ to compute cosine similarities between image and text embeddings.

The evaluation metrics adopted for the task are **HIT@1** (accuracy of the top-1 retrieved image) and **MRR** (Mean Reciprocal Rank), as detailed in the shared task paper.

4.2. Development and Tuning

Since we do not perform any training and an official validation set was not provided, we employed the entire training set as a development set to calibrate the system components and select the optimal configuration. We performed an extensive grid search to determine the best strategies for question selection and score fusion.

Question Selection. We investigated the impact of the number and type of questions (k) on the disambiguation performance. We compared using only Positive questions (varying $k \in [1, 5]$) against using the full battery of Positive and Negative questions. Experiments showed that Negative questions are essential for filtering hard negatives (co-hyponyms), and the best performance was achieved using the maximum set size (5 Positive + 5 Negative).

Fusion Strategy. We compared a Weighted Sum approach against a Max-Pooling strategy. We tested various coefficients for the linear combination:

$$Score = w_1 \cdot S_{VQA} + w_2 \cdot S_{Original} + w_3 \cdot S_{Synthetic} \quad (4)$$

Experimental results indicated that the system is robust enough that complex weighting schemes are unnecessary. The most stable configuration was an **unweighted sum** ($w_1 = w_2 = w_3 = 1$), which treats the evidence from the original query, the synthetic description, and the visual verification as equally important signals.

4.3. Main Results

We submitted three runs to the EVWSD-ITA challenge. The first two runs represent our primary generation pipeline, while the third explores an alternative structured approach, which is detailed in Section 4.4.

- **Run 1 (VQA+Or):** This run combines the VQA verification score with the global CLIP similarity of the original query. The final score is computed as:

$$Score_{Run1} = S_{VQA} + S_{Original} \quad (5)$$

- **Run 2 (VQA+Or+Desc):** This is our complete pipeline. It adds the global CLIP similarity of the *Synthetic Holistic Description* generated by the LLM. The final score corresponds to the full fusion formula:

$$Score_{Run2} = S_{VQA} + S_{Original} + S_{Synthetic} \quad (6)$$

Discussion. Table 1 summarises the performance of our Zero-Shot architecture on the EVWSD-ITA test set. The results provide strong empirical evidence answering our initial research questions. The standard **CLIP Baseline**, which relies solely on the cosine similarity between the candidate images and the raw input triple, achieves a HIT@1 of 45.05%. While this performs significantly better than random chance (10%), it highlights the inherent difficulty of the task: standard dual-encoders struggle to distinguish the target from Hard Negatives (co-hyponyms) that share a similar embedding space. This confirms the “bag-of-words” limitation discussed in Section 2. A dramatic improvement is observed when moving to **Run 0** (VQA only), which uses the aggregated confidence scores of the generated questions without any global CLIP signal. This configuration achieves 65.32% HIT@1, surpassing the baseline by over 20 percentage points. This result provides a definitive answer to **RQ1** and **RQ2**:

³<https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1>

Table 1

Official results on the EVWSD-ITA test set. **Run 2** represents our best-performing configuration. *Or* stands for the CLIP score for each image against the original input sequence and *Desc* for the CLIP score against the description generated by the LLM, as detailed in the previous Section.

Run ID	HIT@1	MRR
CLIP (Baseline)	45.05%	62.44%
Run 0 (VQA)	65.32%	76.82%
Run 1 (VQA+Or)	69.37%	81.00%
Run 2 (VQA+Or+Desc)	71.17%	81.82%
Run 3 (2stepVQA+Or+Desc)	66.67%	77.70%

modelling EVWSD as a *Visual Constraint Verification* task is far more effective than global embedding matching for fine-grained disambiguation. The VQA model, guided by LLM-generated positive and negative constraints, successfully filters out hard negatives that visually resemble the target but lack specific distinguishing features (e.g., specific shapes or usage contexts). Comparing **Run 0** with **Run 1** demonstrates the value of the ensemble approach (**RQ4**). Adding the original CLIP score ($S_{Original}$) to the VQA score improves performance from 65.32% to 69.37%. This suggests that while local feature verification (VQA) is crucial for precision, the global semantic signal from CLIP helps maintain a holistic view, acting as a safeguard when the generated questions might be too specific or ambiguous. The two modules are complementary: VQA handles the fine-grained details, while CLIP ensures global coherence. Finally, comparing **Run 1** and **Run 2** allows us to address **RQ3**. The inclusion of the $S_{Synthetic}$ score in Run 2 yields a further performance improvement, reaching the best result of **71.17% HIT@1** and **81.82% MRR**. This improvement of approximately 1.8% confirms that the raw query provided by the dataset is often too noisy or fragmented for optimal global matching. The LLM-generated description acts as a “semantic bridge”, providing a well-formed text that aligns better with the pre-training distribution of the visual encoder, giving a final boost to the accuracy of our system.

4.4. Ablation Study: The Cost of Structure

To investigate whether explicit linguistic structuring benefits the reasoning process, we submitted a third run based on a **Two-Step Generation** policy. In this configuration, we forced the LLM to first parse the noisy input string into a formal structure (separating Lemma, Hypernym, and Gloss) and then generate the visual questions based solely on this intermediate representation. The final score aggregation remains identical to Run 2 (Equation 6).

The results for **Run 3**, reported in Table 1, offer a crucial insight into prompt engineering for VWSD. With a HIT@1 of **66.67%**, this structured approach significantly outperforms the CLIP Baseline (+21.62%), confirming that the core VQA-based verification strategy is sound regardless of the generation nuance. However, compared to our best configuration (**Run 2**), Run 3 suffers a performance drop of **4.5 percentage points** in HIT@1 and over 4 points in MRR.

This substantial gap refutes the intuition that cleaner, structured inputs necessarily lead to better questions in this context. We attribute this failure to *error propagation*: the rigid “Parse → Generate” pipeline creates an information bottleneck. If the LLM introduces even a minor hallucination or misinterpretation during the initial parsing (e.g., mis-assigning a polysemous word from the gloss to the hypernym field), this error becomes hard-coded in the intermediate structure, irretrievably compromising the subsequent question generation. In contrast, the **Direct Generation** approach (Runs 1 and 2) allows the LLM to leverage its full attention mechanism over the raw, noisy context. This enables the model to resolve ambiguities *implicitly* and dynamically during the generation of questions, proving that for high-capacity models like GPT-5.1, an end-to-end reasoning process is more robust than a formally decomposed pipeline.

4.5. Error Analysis



Figure 3: Top-5 retrieval results for the query “sella borsa bisaccia”, showing a preference for a mounted usage context (17 . jpg) over the gold product image (1216 . jpg).

Table 2

Comparison of VQA responses between the predicted image (17 . jpg) and the gold image (1216 . jpg) for the query “sella borsa bisaccia”. The positive (first 5) questions favour the mounted context, while the negative (last 5) questions are not discriminative enough. The Score for each question is computed as $p(YES)$ for the positive ones and $1 - p(YES)$ for the negative ones, as detailed in Section 3, Phase 3.

Question	17 . jpg (Predicted)			1216 . jpg (Gold)		
	Answer	$p(YES)$	Score	Answer	$p(YES)$	Score
<i>Si vedono due borse morbide fissate ai lati di una sella?</i>	YES	0.6225	0.6225	NO	0.0180	0.0180
<i>Le borse sono collegate tra loro e poggiano simmetricamente sul dorso del mezzo?</i>	YES	0.7311	0.7311	NO	0.0373	0.0373
<i>Le borse sembrano progettate per contenere oggetti durante uno spostamento?</i>	YES	1.0000	1.0000	YES	0.9998	0.9998
<i>Il materiale delle borse appare robusto, come cuoio o tessuto spesso?</i>	YES	0.7773	0.7773	YES	0.9968	0.9968
<i>Le borse sono posizionate dietro o accanto al sedile principale del cavallo o moto?</i>	NO	0.1480	0.1480	NO	0.2689	0.2689
<i>Si vede solo un sedile rigido senza borse laterali agganciate?</i>	NO	0.0000	1.0000	NO	0.0000	1.0000
<i>L’oggetto è una borsa singola portata a spalla da una persona?</i>	NO	0.0000	1.0000	NO	0.0000	1.0000
<i>L’oggetto è uno zaino con spallacci indossato sulla schiena?</i>	NO	0.0000	1.0000	NO	0.0000	1.0000
<i>Il contenitore è un bauletto rigido montato centralmente sul portapacchi?</i>	NO	0.0000	1.0000	NO	0.0000	1.0000
<i>La borsa è appesa al manubrio senza appoggiarsi ai lati della sella?</i>	NO	0.0180	0.9820	NO	0.0000	1.0000

To illustrate the failure modes of the proposed VQA-based scoring strategy, we analyse a representative example in which the predicted top-ranked image differs from the gold annotation. The following discussion focuses on two recurring phenomena observed across multiple errors: the effect of ambiguous images providing partial but question-aligned visual evidence, and the impact of negative questions that are logically correct yet insufficiently discriminative. The example reported in Figure 3 and Table 2 serves as a concrete case study to highlight how these factors interact and influence the final ranking.

In the analysed case, the system is strongly influenced by the presence of partial visual evidence that directly matches the formulation of the VQA questions. The first two questions explicitly assume the visibility of saddlebags already mounted on a saddle, implicitly requiring the presence of the vehicle context. Image 17 . jpg satisfies these assumptions by depicting a bicycle with lateral bags mounted, and consequently receives affirmative answers with medium-to-high confidence. In contrast, the gold image 1216 . jpg, while correctly representing a saddle bag as an object, is photographed in isolation

and does not include the saddle or the vehicle. As a result, the same questions yield near-zero scores, penalising the gold image despite its higher semantic relevance to the query. This behaviour indicates that the system prioritises local visual alignment with the questions over the correct identification of the target object class.

In addition, the example highlights the role of negative questions that are correct but poorly aligned with the discrimination task. The last five questions in Table 2 exclude clearly distinct categories such as backpacks, shoulder bags, rigid top cases, or handlebar bags. Both images correctly receive negative answers with near-maximal confidence, leading to similar score contributions. While these questions are valid from a logical standpoint, they do not help distinguish between semantically close candidates, such as mounted lateral bags and saddle bags depicted as standalone products. Their contribution to the aggregated score is therefore inflationary rather than informative, amplifying the advantage already accumulated by the image favoured by context-dependent questions. This behaviour suggests that negative questions should be carefully selected or weighted according to their actual discriminative power, in order to avoid trivial exclusions dominating the decision process.

5. Conclusions

In this paper, we presented the UniTor system for the EVWSD-ITA task at EVALITA 2026. We proposed a fully Zero-Shot, training-free architecture that redefines Visual Word Sense Disambiguation as a process of *Visual Constraint Verification*. By leveraging the reasoning capabilities of Large Language Models to guide the discriminative power of Visual Question Answering models, our approach moves beyond the limitations of standard global embeddings.

Our experimental results allow us to draw three main conclusions regarding the initial research questions:

1. **Verification outperforms Matching:** The substantial performance gap between our VQA-based runs and the CLIP baseline confirms that decomposing a query into fine-grained Positive and Negative questions is essential for distinguishing the target from hard negatives, such as co-hyponyms.
2. **Context Reconstruction is beneficial:** Augmenting the noisy input with an LLM-generated holistic description provides a cleaner semantic signal, offering a robust bridge between the linguistic and visual modalities (**Run 2**).
3. **Implicit Reasoning is robust:** The ablation study on structured parsing (**Run 3**) revealed that enforcing a rigid “Parse → Generate” pipeline introduces information bottlenecks. Direct generation proves to be superior, suggesting that state-of-the-art LLMs manage ambiguity better when allowed to attend to the raw context implicitly.

Future work will focus on the efficiency of the pipeline and a thorough investigation of the selection of negative questions. While the current architecture achieves high accuracy, it relies on large, API-based models. We plan to investigate whether smaller, open-source Language Models can be fine-tuned to generate high-quality visual constraints, making the approach more sustainable and deployable in resource-constrained environments.

Acknowledgments

We acknowledge financial support from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and support from Project ECS 0000024 Rome Technopole - CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, funded by the European Union - NextGenerationEU.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

- [1] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE transactions on pattern analysis and machine intelligence* 41 (2018) 423–443.
- [2] A. Mogadala, M. Kalimuthu, D. Klakow, Trends in integration of vision and language research: A survey of tasks, datasets, and methods, *J. Artif. Intell. Res.* 71 (2021) 1183–1317. URL: <https://doi.org/10.1613/jair.1.11688>. doi:10.1613/JAIR.1.11688.
- [3] S. Harnad, The symbol grounding problem, *Physica D: Nonlinear Phenomena* 42 (1990) 335–346.
- [4] A. Raganato, I. Calixto, A. Ushio, J. Camacho-Collados, M. T. Pilehvar, SemEval-2023 task 1: Visual word sense disambiguation, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2227–2234. URL: <https://aclanthology.org/2023.semeval-1.308/>. doi:10.18653/v1/2023.semeval-1.308.
- [5] F. Cutugno, A. Miaschi, A. Palmero Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, EVALITA 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: *Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026)*, CEUR.org, Bari, Italy, 2026.
- [6] A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Association for Computational Linguistics, Toronto, Canada, 2023. URL: <https://aclanthology.org/2023.semeval-1.0/>.
- [7] R. Navigli, S. P. Ponzetto, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial intelligence* 193 (2012) 217–250.
- [8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, *CoRR abs/2103.00020* (2021). URL: <https://arxiv.org/abs/2103.00020>. arXiv:2103.00020.
- [9] X. Zhang, T. Zhen, J. Zhang, Y. Wang, S. Liu, Srcb at semeval-2023 task 1: Prompt based and cross-modal retrieval enhanced visual word sense disambiguation, in: *Proceedings of the 17th international workshop on semantic evaluation (SemEval-2023)*, 2023, pp. 439–446.
- [10] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, 2024. URL: <https://arxiv.org/abs/2306.13549>. arXiv:2306.13549.
- [11] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-vl technical report, 2025. URL: <https://arxiv.org/abs/2502.13923>. arXiv:2502.13923.
- [12] C. D. Hromei, A. Scaiella, D. Croce, R. Basili, Grounded semantic role labelling from synthetic multimodal data for situated robot commands, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Suzhou, China, 2025, pp. 23758–23781. URL: <https://aclanthology.org/2025.emnlp-main.1212/>. doi:10.18653/v1/2025.emnlp-main.1212.
- [13] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin,

- X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, K. Zhu, Qwen3-vl technical report, 2025. URL: <https://arxiv.org/abs/2511.21631>. arXiv:2511.21631.
- [14] M. Yuksekogunul, F. Bianchi, P. Kalluri, D. Jurafsky, J. Zou, When and why vision-language models behave like bags-of-words, and what to do about it?, arXiv preprint arXiv:2210.01936 (2022).
 - [15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* 35 (2022) 24824–24837.
 - [16] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, K. Keutzer, How much can clip benefit vision-and-language tasks?, arXiv preprint arXiv:2107.06383 (2021).
 - [17] A. Shtedritski, C. Rupprecht, A. Vedaldi, What does clip know about a red circle? visual prompt engineering for vlms, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11987–11997.
 - [18] C. Cuttano, G. Rosi, G. Trivigno, G. Averta, What does clip know about peeling a banana?, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2238–2247.
 - [19] R. Ramos, V. Stojnić, G. Kordopatis-Zilos, Y. Nakashima, G. Tolas, N. Garcia, Processing and acquisition traces in visual encoders: What does clip know about your camera?, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 17056–17066.
 - [20] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
 - [21] J. Zhang, S. Yu, D. Chong, A. Sicilia, M. R. Tomz, C. D. Manning, W. Shi, Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2025. URL: <https://arxiv.org/abs/2510.01171>. arXiv:2510.01171.

A. Appendix

In this Appendix, we report the prompts used in our pipeline, which are the prompt for GPT-5.1 for generating the questions (A.1), the VQA prompt for answering the generated questions (A.2) and the two prompts used in the Run3 of our ablation study (A.3).

A.1. Question Generator Prompt

This prompt instructs the GPT-5.1 model to adopt the persona of a computational linguist specialised in visual and lexical semantics. The primary objective is to generate a structured JSON output that profiles the semantic content of a given text segment (k words) for the purpose of assessing image-text consistency. The procedure operates in two phases in the same answer, conditioning the question generation step on the identified sense. First, the model performs a semantic analysis to identify the concrete visual target (synset) and its hypernym, while simultaneously detecting potential semantic ambiguities, including co-hyponyms and polysemous alternative senses. Second, the model constructs a JSON object containing the identified target, a functional description, and a set of ten weighted binary questions (five confirming, five discriminating). These questions are strictly grounded in visual properties (shape, texture, colour) to validate the presence of the concept. Each question is assigned a weight ($w \in [0.5, 1.0]$) indicating either its representativeness of the target or its power to discriminate the target from semantic distractors.

Sei un linguista computazionale esperto di semantica visiva e lessicale.

Il tuo obiettivo è creare un JSON che rappresenti la conoscenza semantica di un contesto testuale contenente k parole e che serva per valutare se un'immagine rappresenta correttamente il concetto espresso dal contesto.

FASE 1 – ANALISI SEMANTICA

1. Considera il contesto testuale come un frammento estratto da una rete semantica (tipo BabelNet o WordNet).
2. Identifica (ma non con questo ordine):
 - l'OGGETTO VISIVO CONCRETO (synset, target visivo);
 - la CATEGORIA GENERALE (iperonimo);
 - la sua descrizione (solitamente una parola estratta dalla sua glossa).
3. Elenca 3-6 concetti affini o potenzialmente confondibili (co-iponimi o sensi alternativi). Questi rappresentano le "distrazioni semantiche".

FASE 2 – GENERAZIONE DEL JSON

Genera un JSON nel formato seguente:

```
{
  "query": "<contesto di k parole>",
  "identified_target": "<synset, l'oggetto visivo concreto>",
  "description": "<breve spiegazione visiva e funzionale del target identificato>",
  "questions_yes": [
    {"testo": "<domanda>", "w": <valore tra 0.5 e 1.0>},
    {"testo": "<domanda>", "w": <valore tra 0.5 e 1.0>},
    ...
  ],
  "hypothesized_co_hyponyms": [
    "<termine1>",
    "<termine2>",
    "<termine3>",
    ...
  ],
  "senses_to_exclude": [
    "<senso alternativo 1>",
    "<senso alternativo 2>",
    "<senso alternativo 3>",
    ...
  ],
  "questions_no": [
    {"testo": "<domanda>", "w": <valore tra 0.5 e 1.0>},
    {"testo": "<domanda>", "w": <valore tra 0.5 e 1.0>},
    ...
  ]
}
```

ISTRUZIONI PER LE DOMANDE

- Scrivi tutto in ITALIANO corretto.
- Domande concise (al più 15 parole), rispondibili con "si" o "no".
- Basate su proprietà visive concrete (forma, materiale, colore, posizione, funzione, ecc. Tutto ciò che distingue il concetto target da qualunque altro senso possibile).
- Genera esattamente 5 domande "si" e 5 domande "no".
- Ogni domanda ha un campo `w`:
 - Per le **questions_yes**: indica quanto fortemente la domanda coglie l'essenza visiva del target identificato.
 - Per le **questions_no**: indica quanto fortemente la domanda è discriminante nel distinguere il target identificato dai sensi alternativi.

LINEA GUIDA SEMANTICA

- Le domande con `w` più alto devono essere le più rappresentative o discriminanti.
- Es: per le questions_yes, quella con `w`=1.0 deve esprimere la caratteristica più distintiva del concetto target.

- Es: per le questions_no, quella con `w`=1.0 deve cogliere la differenza chiave che fa capire che NON si tratta del concetto target.
- Se utile, immagina di addestrare un modello visivo-linguistico: le domande con `w` alto sono le più informative per capire il concetto.

ESEMPIO

Input: "attrezzo gomma da cancellare atto"

Output:

```
{
  "query": "attrezzo gomma da cancellare atto",
  "identified_target": "Gomma da cancellare",
  "description": "Piccolo blocco di materiale gommoso o vinilico utilizzato per rimuovere tratti di matita o inchiostro dalla carta tramite abrasione.",
  "questions_yes": [
    {
      "testo": "L'oggetto sembra progettato per essere sfregato contro una superficie cartacea?",
      "w": 1.0
    },
    {
      "testo": "L'oggetto si presenta come un piccolo blocco solido, spesso a forma di parallelepipedo?",
      "w": 0.9
    },
    {
      "testo": "L'oggetto ha una texture liscia e opaca, tipica della gomma o del vinile?",
      "w": 0.8
    },
    {
      "testo": "L'oggetto è tipicamente bianco, rosa o bicolore rosso e blu?",
      "w": 0.7
    },
    {
      "testo": "L'oggetto è contestualizzato vicino a strumenti di scrittura come matite o fogli?",
      "w": 0.6
    }
  ],
  "hypothesized_co_hyponyms": [
    "Temperamatite",
    "Bianchetto liquido",
    "Correttore a nastro",
    "Righello"
  ],
  "senses_to_exclude": [
    "Pneumatico (gomma dell'auto)",
    "Gomma da masticare (chewing gum)",
    "Elastico (anello di gomma)",
    "Caucciù (materiale grezzo)"
  ],
  "questions_no": [
    {
      "testo": "L'oggetto è un grande anello nero con battistrada montato su un cerchione?",
      "w": 1.0
    },
    {
      "testo": "L'oggetto è un piccolo confetto o lastrina confezionata per uso alimentare?",
      "w": 0.9
    },
    {
      "testo": "L'oggetto è un sottile anello flessibile usato per tenere insieme altri oggetti?",
      "w": 0.8
    },
    {
      "testo": "L'oggetto è un flaconcino contenente un liquido bianco con un applicatore?",
      "w": 0.7
    },
    {
      "testo": "L'oggetto presenta lame metalliche visibili per affilare il legno?",
      "w": 0.6
    }
  ]
}
```

```
]
}
```

Ora elabora il seguente concetto:
\$QUERY\$

Genera SOLO il JSON richiesto, senza testo aggiuntivo o spiegazioni.

A.2. Prompt for VQA

The VQA (Qwen3-VL-8b) inference prompt is a simple request to answer the question with just Yes or No, in order to verify the visual features and extract the associated probabilities as scores.

Stai svolgendo un compito di Visual Question Answering (VQA).

Ti fornisco un'immagine e una domanda. La domanda è stata generata a partire da tre indizi lessicali che definiscono un concetto target (lemma del concetto, lemma del suo iperonimo e un termine di contesto estratto dalla glossa). L'ordine dei tre indizi è arbitrario.

Indizi: \$INPUT QUERY\$

Istruzioni:

- Rispondi basandoti esclusivamente su ciò che è visibile nell'immagine.
- Se l'evidenza visiva è assente o ambigua, rispondi NO.
- Produci come output una sola parola: YES oppure NO.
- Scrivi YES/NO come primo token e non aggiungere altro testo.

Domanda: \$GENERATED QUESTIONS\$

A.3. Prompt for Run3: Structured Parsing and Question Generation

In this section, we report the prompts used for the third run of our submission, i.e., the one where we first ask GPT-5.1 to parse the input concept into a structured object, and then to generate the list of questions.

The prompt for parsing the concept into a structured form is the following:

Sei un esperto semantico specializzato in analisi lessicale e ontologie (p.e., BabelNet).

Il tuo compito è analizzare una stringa di testo che contiene tre componenti semantici mescolati in ordine casuale e senza separatori espliciti. Devi segmentare la stringa e ricostruire la struttura logica originale.

I tre componenti nascosti nella stringa sono:

1. "lemma_concetto": Il concetto specifico (Target). Può essere composto da una o più parole (es. "gomma da cancellare").
2. "lemma_iperonimo": La categoria generale a cui appartiene il concetto. Può essere composto da una o più parole (es. "attrezzo"). Relazione: [Concetto] È UN [Iperonimo].
3. "parola_glossa": Una singola parola estratta dalla definizione del dizionario (glossa) del concetto.

ISTRUZIONI DI RAGIONAMENTO:

- Identifica la relazione gerarchica: cerca la coppia di termini dove uno è il "padre" (iperonimo) e l'altro è il "figlio" (concetto).
- Riconosci le espressioni polirematiche: in italiano molti concetti sono formati da più parole (es. "carta di credito", "mulino ad acqua"). Non spezzarle.
- La parola rimanente, che non è né il concetto né l'iperonimo, è la parola della glossa.

ESEMPI RISOLTI (Few-Shot Learning):

Esempio 1:

Input: "strumento musicale tasti pianoforte"

Analisi: "pianoforte" è un tipo di "strumento musicale". "tasti" è una parola che descrive il pianoforte.

Output JSON:

```
{
  "lemma_concetto": "pianoforte",
  "lemma_iperonimo": "strumento musicale",
  "parola_glossa": "tasti"
}
```

Esempio 2 (Ordine diverso, concetto multi-parola):

Input: "felino gatto domestico mammifero"

Analisi: Qui c'è ambiguità, ma "gatto domestico" è il concetto più specifico. "mammifero" è la categoria più alta, ma spesso l'iperonimo diretto è "felino". Se la stringa fosse "gatto domestico felino mammifero", "felino" è l'iperonimo diretto. Assumiamo: concetto="gatto domestico", iperonimo="felino", glossa="mammifero".

Output JSON:

```
{
  "lemma_concetto": "gatto domestico",
  "lemma_iperonimo": "felino",
  "parola_glossa": "mammifero"
}
```

Esempio 3 (Iperonimo multi-parola, ordine mescolato):

Input: "pagamento carta di credito tessera magnetica"

Analisi: "carta di credito" è il concetto. "tessera magnetica" è la categoria (iperonimo). "pagamento" è l'uso (glossa).

Output JSON:

```
{
  "lemma_concetto": "carta di credito",
  "lemma_iperonimo": "tessera magnetica",
  "parola_glossa": "pagamento"
}
```

Esempio 4 (Il tuo caso specifico):

Input: "attrezzo gomma da cancellare atto"

Analisi: "gomma da cancellare" è l'oggetto specifico. "attrezzo" è la categoria generale. "atto" è la parola della descrizione.

Output JSON:

```
{
  "lemma_concetto": "gomma da cancellare",
  "lemma_iperonimo": "attrezzo",
  "parola_glossa": "atto"
}
```

TASK CORRENTE:

Analizza la stringa seguente e restituisci SOLO il JSON formattato.

Input: \$INPUT CONCEPT\$

The prompt to generate the questions from the structured concept is the following:

Sei un esperto di Vision-Language e Prompt Engineering per la challenge EVALITA VWSD.

Il tuo input è un oggetto JSON che definisce un concetto visivo preciso, già disambiguato.

Input JSON:

```
{
  "lemma_concetto": "<Target Visivo>",
  "lemma_iperonimo": "<Categoria/Padre>",
  "parola_glossa": "<Contesto/Disambiguatore>"
}
```

OBIETTIVO:

Generare una batteria di domande SI/NO in **italiano** da sottoporre a un modello CLIP/VLM per capire se un'immagine candidata rappresenta **ESATTAMENTE** il "lemma_concetto".

STRATEGIA AVANZATA:

Il modello visivo deve distinguere il Target dai "Hard Negatives" (immagini molto simili ma sbagliate).

1. Se il Target è polisemico (ha più significati), usa la "parola_glossa" per capire quale senso è

- quello giusto ed escludere gli altri.
- Se il Target ha "fratelli" nella stessa categoria (co-iponimi), devi generare domande che evidenzino le differenze sottili.
 - Esempio: Target="Ghepardo", Iperonimo="Felino".
 - Hard Negative: "Leopardo".
 - Domanda NO efficace: "Le macchie sul mantello sono a forma di rosetta con il centro vuoto?" (Vero per Leopardo, Falso per Ghepardo).

REGIONE DI GENERAZIONE:

- **Solo Visuale**** Le domande devono riguardare forme, colori, texture, layout, azioni visibili. Niente concetti astratti.
- **Niente "Target"****: NON usare la parola contenuta in "lemma_concetto" nel testo delle domande. Descrivilo.
- **Domande SI (Positive)****: Devono essere VERE solo per il Target.
- **Domande NO (Trappole)****: Devono essere FALSE per il Target, ma VERE per i suoi simili (co-iponimi) o per i suoi altri sensi. Devono servire a scartare le immagini "quasi giuste".
- **Ogni domanda ha un campo "w"*****:
 - Per le ****questions_yes****: indica quanto fortemente la domanda coglie l'essenza visiva del target identificato.
 - Per le ****questions_no****: indica quanto fortemente la domanda è discriminante nel distinguere il target identificato dai sensi alternativi.

OUTPUT FORMAT (JSON):

Restituisci solo il JSON seguente.

```
{
  "description": "Breve descrizione del concetto che hai ricevuto in input, con focus particolare sul lemma target e sul suo iperonimo",
  "questions_yes": [
    {"<Domanda 1: Focus su attributo univoco del target>", "w": <score per la domanda 1>},
    {"<Domanda 2: Focus su forma/struttura>", "w": <score per la domanda 2>},
    {"<Domanda 3: Focus su contesto tipico>", "w": <score per la domanda 3>},
    {"<Domanda 4: Focus su dettagli specifici>", "w": <score per la domanda 4>},
    {"<Domanda 5: Conferma visiva finale>", "w": <score per la domanda 5>}
  ],
  "hypothesized_co_hyponyms": [
    "<termine1>",
    "<termine2>",
    "<termine3>",
    ...
  ],
  "senses_to_exclude": [
    "<senso alternativo 1>",
    "<senso alternativo 2>",
    "<senso alternativo 3>",
    ...
  ],
  "questions_no": [
    {"<Domanda 1: Vera per un co-iponimo specifico, falsa per il target>", "w": <score per la domanda 1>},
    {"<Domanda 2: Vera per un altro senso della parola (se esiste), falsa per questo>", "w": <score per la domanda 2>},
    {"<Domanda 3: Vera per l'iperonimo in generale ma falsa per questo specifico sottotipo>", "w": <score per la domanda 3>},
    {"<Domanda 4: Trappola visiva su dettaglio errato>", "w": <score per la domanda 4>},
    {"<Domanda 5: Altra trappola per co-iponimo>", "w": <score per la domanda 5>}
  ]
}
```

ESEMPIO

Input:

```
{
  "lemma_concetto": "gomma da cancellare",
  "lemma_iperonimo": "attrezzo",
  "parola_glossa": "atto"
}
```

Output:

```
{
  "description": "Piccolo blocco di materiale gommoso o vinilico utilizzato per rimuovere tratti di matita o inchiostro dalla carta tramite abrasione.",
  "questions_yes": [
```

```

    {
      "testo": "L'oggetto sembra progettato per essere sfregato contro una superficie cartacea?",
      "w": 1.0
    },
    {
      "testo": "L'oggetto si presenta come un piccolo blocco solido, spesso a forma di
        parallelepipedo?",
      "w": 0.9
    },
    {
      "testo": "L'oggetto ha una texture liscia e opaca, tipica della gomma o del vinile?",
      "w": 0.8
    },
    {
      "testo": "L'oggetto è tipicamente bianco, rosa o bicolore rosso e blu?",
      "w": 0.7
    },
    {
      "testo": "L'oggetto è contestualizzato vicino a strumenti di scrittura come matite o fogli?",
      "w": 0.6
    }
  ],
  "hypothesized_co_hyponyms": [
    "Temperamatite",
    "Bianchetto liquido",
    "Correttore a nastro",
    "Righello"
  ],
  "senses_to_exclude": [
    "Pneumatico (gomma dell'auto)",
    "Gomma da masticare (chewing gum)",
    "Elastico (anello di gomma)",
    "Caucciù (materiale grezzo)"
  ],
  "questions_no": [
    {
      "testo": "L'oggetto è un grande anello nero con battistrada montato su un cerchione?",
      "w": 1.0
    },
    {
      "testo": "L'oggetto è un piccolo confetto o lastrina confezionata per uso alimentare?",
      "w": 0.9
    },
    {
      "testo": "L'oggetto è un sottile anello flessibile usato per tenere insieme altri oggetti?",
      "w": 0.8
    },
    {
      "testo": "L'oggetto è un flaconcino contenente un liquido bianco con un applicatore?",
      "w": 0.7
    },
    {
      "testo": "L'oggetto presenta lame metalliche visibili per affilare il legno?",
      "w": 0.6
    }
  ]
}

```

INPUT: \$PARSED CONTEXT\$