# Zero-Shot Reasoning with BLIP and SmolLM

Notebook for the MultimodalReasoning Lab at CLEF 2025

Elena Tosheva[1], Dimitar Dimitrov[1], Ivan Koychev[1] and Preslav Nakov[2]

[1]Sofia University "St. Kliment Ohridski", Bulgaria
[2] Mohamed bin Zayed University of Artificial Intelligence, UAE

**Abstract**

This article was developed as part of the ImageCLEF 2025 competition. We adapted the BLIP-Base image-captioning model for the Multimodal Reasoning task, integrating the SmolLM-360M model for question answering and training on the MBZUAI EXAMS-V dataset (16 724 training and 4 208 validation examples). We then conducted a prompt-ablation study using three different templates to evaluate their impact on answer-key accuracy, measured by case-insensitive substring matching against the correct option within the provided set of three to five answers. Finally, we analyzed the distributions of generated caption lengths.

**Keywords**

MultiModal, Image CLEF 2025, Image Captioning, MultiModal Reasoning

## 1. Introduction

Multimodal reasoning—the ability to jointly interpret and reason over visual and textual inputs—is a core challenge in AI, with important applications in education, accessibility, and cross-modal search. Tasks such as visual question answering and image-based multiple-choice problems require systems to understand both the semantic content of images and the structure of natural language prompts. The ImageCLEF 2025 Multimodal Reasoning task [1];[2], which is part of the Image CLEF Multimedia retrieval [3] requires selecting the correct answer from 3–5 provided options, given an image of a science question. Efficient deployment on limited hardware motivates using mid-scale models such as BLIP-Base [4] for captioning and SmolLM-360M [5] for reasoning. We made experiments varying prompt templates - none, "A photo of", and "Describe what you see:" — to optimize caption informativeness for multiple-choice prediction.

## 2. Related Work

Modern vision–language systems combine powerful image encoders with autoregressive text decoders to perform tasks such as image captioning, visual question answering, and multimodal reasoning. Early works such as CLIP demonstrated that contrastive pretraining on large-scale image–text pairs yields embeddings that transfer well to downstream classification and retrieval tasks. Building on this, BLIP introduced a dual objective of contrastive alignment and generative captioning, producing models such as `Salesforce/blip-image-captioning-base` and `-large` that achieve state-of-the-art results on COCO and other benchmarks [6].

At the same time, recent advances in compact causal language models (with less than 500 M parameters) show that mid-scale Transformers can deliver strong generative performance under tight compute budgets. SmolLM-360M [5] is one such model, featuring 24 layers, rotary positional embeddings, and optimized training for inference on a single 12 GB GPU. Prompt engineering has emerged as a simple yet effective way to steer generative models toward desired behaviors. In vision–language captioning, prepending a short instruction (e.g, "A picture of …") can influence both the style and content of the

*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*
*Corresponding author.
elena.tosheva@gmail.com (E. Tosheva); mitko.bg.ss@gmail.com (D. Dimitrov); koychev@fmi.uni-sofia.bg (I. Koychev); preslav.nakov@mbzuai.ac.ae (P. Nakov)

generated description. Understanding how prompt phrasing affects downstream tasks—such as extracting multiple-choice answers from generated captions, is critical for reliable deployment in real-world settings.

The ImageCLEF 2025 Multimodal Reasoning task challenges systems to select the correct answer from 3–5 provided options, given an image of a science exam question, covering topics from chemistry to physics, across multiple languages [1]. The publicly released MBZUAI EXAMS-V dataset [7] provides 16,724 training and 4,208 validation examples, each consisting of a question image, a balanced four-way answer key, and associated metadata. In our study, we leverage this dataset to evaluate how BLIP-based captioning and prompt variations impact the ability of an LLM-powered pipeline to recover the correct answer via simple substring matching.

## 3. System Overview

### 3.1. Dataset

We use the **MBZUAI EXAMS-V** dataset [7], which consists of 16,724 training and 4,208 validation science exam questions in image format. Each image comes with a 3–5-option multiple-choice question and associated metadata. Importantly, the dataset spans multiple languages, and in our experiments, we utilize all available languages to evaluate model robustness across multilingual contexts.

### 3.2. Captioning Pipeline

To generate image captions, we use the following encoder-decoder models:

- **BLIP-Base(Salesforce/blip-image-captioning-base)**
- **BLIP-Large (Salesforce/blip-image-captioning-large,CLIP-ViT-L/14 backbone)**

These encoders extract visual features from the exam images and decode them into textual descriptions. The captions are later used as inputs for question-answering via a language model.

### 3.3. Prompt Ablation

We assess how prompt phrasing affects caption content and downstream Accuracy. Each image is paired with one of three prompt templates:

- **None**: The image is passed without additional text.
- **"A picture of"**: encourages concise object-focused captions.
- **"Describe what you see:"**: encourages detailed, descriptive captions.

### 3.4. Model

To perform reasoning over our generated image captions, we employ a lightweight yet powerful language model:

- **SmalLM 360M** - a compact transformer-based language model optimized for low-resource inference. With only 360 million parameters and efficient deployment on hardware with as little as 12 GB of GPU memory, it enables practical experimentation without sacrificing performance. Despite its small size, **SmolLM-360M is currently the best-performing model in its category (sub-500M parameters)**. According to the Hugging Face benchmark [6], it outperforms other similarly sized models—including MobileLM-350M and Qwen2-500M—across a range of benchmarks that test general knowledge, commonsense reasoning, and reading comprehension.

  We use a **zero-shot** setup: SmolLM-360M is not fine-tuned. Given a caption produced by BLIP, we prompt the model as follows:

- **Prompt**:

```
[CAPTIONED QUESTION]
{caption}
Choose the correct answer from the following options: A, B, C, D, E.\n
 Answer:
```

  This zero-shot approach allows us to simulate realistic, low-resource deployment conditions while assessing how well the model can reason over image-derived text alone.

### 3.5. Evaluation

Answer-key accuracy is the percentage of validation samples whose generated caption contains the correct option letter (A–E) as a standalone token. Formally:

$$\text{accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\text{token}(k_i) \in \text{tokens}(p_i)] \times 100\%.$$

Here $N = 4208$ for the validation split. We also report the distribution of caption token lengths.

## 4. Results

All experiments are run on Google Colab Pro's T4 GPU.

### 4.1. Prompt-Ablation Results

Table 1 summarizes answer-key accuracy and average caption length for each prompt template on the validation set. Figure 1 shows the graphics distribution of the results below.

**Table 1**
Prompt Ablation and Model Comparison on Validation Set ($N = 4208$).

| Model | Prompt Template | Accuracy (%) | Avg. Length (caption tokens) |
|---|---|---|---|
| BLIP-Base | None | 22.01 | 11.37 |
| BLIP-Base | A photo of | 20.41 | 14.27 |
| BLIP-Base | Describe what you see: | 21.99 | 14.70 |
| BLIP-Large | None | 21.95 | 13.14 |
| BLIP-Large | A photo of | 21.95 | 12.96 |
| BLIP-Large | Describe what you see: | 22.01 | 12.65 |

Adding either prompt ("A photo of" or "Describe what you see:") increases caption length by 1–3 tokens compared to no prompt.

For BLIP-Base, "A photo of" actually hurts accuracy (20.41% vs. 22.01% with no prompt), while "Describe what you see:" matches the no-prompt accuracy.

For BLIP-Large, all three templates yield nearly identical accuracy ( 21.95–22.01%), with "Describe what you see:" giving the very slight edge (22.01%) and the shortest captions of the three.

### 4.2. Official Submission Results

Our system was officially evaluated as part of the ImageCLEF 2025 Multimodal Reasoning task [1], where we participated under the team name *elenat*. We submitted a single, zero-shot pipeline that used BLIP-based captioning and the compact SmolLM-360M model for reasoning.

Unlike many participating systems that focused on individual languages, we ran our model on the **entire multilingual test set**, which includes science exam questions in multiple languages such as English, Bulgarian, Arabic, and others. This multilingual setup allowed us to evaluate the generalization

capabilities of our lightweight models across a diverse range of inputs. Importantly, our official submission used the **bare image input without any additional prompt text** for captioning—i.e., we did not prepend instructions like "A photo of" or "Describe what you see:". This minimal setup demonstrates the capability of our pipeline to extract useful semantic information from images alone.

And here are the results that we achieved:

**Table 2**
Official Results of Team *elenat* on ImageCLEF 2025 Multimodal Reasoning Task ran on the test set.

| Category | Rank | Accuracy (%) |
|---|---|---|
| English | 11 | 25.20 |
| Bulgarian | 6 | 23.50 |
| Multilingual | 10 | 21.88 |

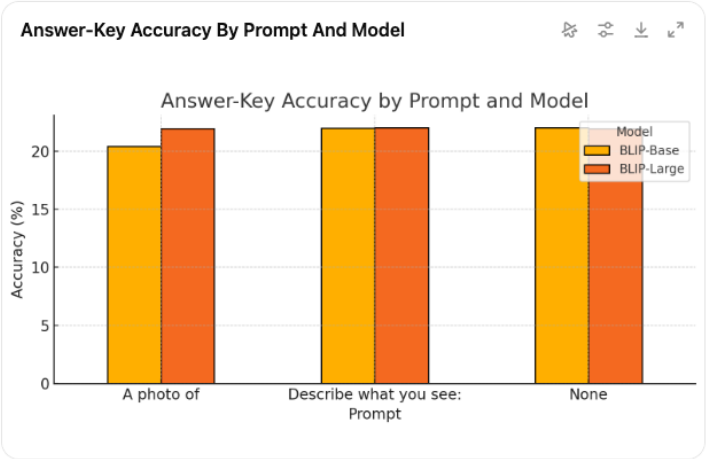


**Figure 1:** Accuracy by prompt and model.

- **English**: placed 11th with **25.20 %** accuracy
- **Bulgarian**: placed 6th with **23.50 %** accuracy
- **Multilingual**: placed 10th with **21.88 %** accuracy

In other words, our strongest relative showing was in the Bulgarian track (rank 6), even though the absolute highest accuracy was in the English track.

## 5. Discussion

Our experiments reveal two overarching trends:

- **Caption Conciseness Correlates with Accuracy** Across both BLIP-Base and BLIP-Large, the shortest outputs consistently yield the best match against the answer key. For BLIP-Base, omitting any leading prompt ("None") produces the briefest captions (11.4 tokens) and delivers the highest accuracy (22.0%). Likewise, for BLIP-Large, the "Describe what you see:" template—despite being wordier than no prompt—actually results in the most concise captions (12.7 tokens) of the three setups and achieves the top performance (22.0%).
- **Prompt Wording Matters—But Only Modestly** Swapping among "A photo of," "Describe what you see:," or no explicit prefix shifts accuracy by at most 1.6 points. In contrast, average caption lengths vary by as much as 3 tokens. This gap suggests that while prompt phrasing reliably inflates or trims verbosity, it only marginally influences the model's ability to generate an answer–key match. In other words, template choice can nudge performance but is not the dominant factor.

## 6. Conclusion

We present a prompt-ablation study for **BLIP-Base** on ImageCLEF 2025 [3], demonstrating that simple question prompt variations can affect multiple-choice accuracy. Encouraging the model to keep captions brief (either via no prompt or a very lean template) appears to help it mention the correct multiple-choice letter more reliably. Future work may include dynamic prompt optimization and multilingual adaptation.

## Declaration on Generative AI

During the preparation of this work, we used OpenAI GPT-4o to assist with grammar and spelling improvements. All suggestions were reviewed and edited by the authors, who take full responsibility for the final content of the publication.

## References


[1] Imageclef 2025 multimodal reasoning task, https://www.imageclef.org/2025/multimodalreasoning, 2025.

[2] D. Dimitrov, M. S. Hee, Z. Xie, R. Jyoti Das, M. Ahsan, S. Ahmad, N. Paev, I. Koychev, P. Nakov, Overview of imageclef 2025 – multimodal reasoning, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[3] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[4] Salesforce, Blip image captioning base model, https://huggingface.co/Salesforce/blip-image-captioning-base, 2023.

[5] Hugging Face, Smollm-360m model, https://huggingface.co/HuggingFaceTB/SmolLM-360M, 2024.

[6] Hugging Face Blog, Smollm: Blazingly fast and remarkably powerful, https://huggingface.co/blog/smollm, 2024.

[7] R. Das, S. Hristov, H. Li, D. Dimitrov, I. Koychev, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7768–7791. URL: https://aclanthology.org/2024.acl-long.420/. doi:10.18653/v1/2024.acl-long.420.