NU_Internship team at ImageEval 2025: From Zero-Shot to Ensembles: Enhancing Grounded Arabic Image Captioning

Rana Gaber[1*] **Seif Eldin Amgad**[1*] **Ahmed Sherif Nasri**[2*]
Mohamed Ibrahim Ragab[3] **Ensaf H. Mohamed**[3]
[1] Faculty of Computers and Data Science, Alexandria University
[2] Faculty of Engineering, Ain Shams University
[3] CIS, School of Information Technology and Computer Science, Nile University
cds.{ranaahmed30309,seifamgad24237}@alexu.edu.eg
23p0270@eng.asu.edu.eg
{MoRagab, EnMohamed}@nu.edu.eg

Abstract

Arabic image captioning remains underexplored in vision–language research due to limited resources and the linguistic complexity of Arabic. In the ImageEval 2025 Shared Task, we evaluated three models, AIN, BLIP-Arabic-Flickr-8k, and Qwen 2.5, across zero-shot, fine-tuning, retrieval-augmented, and ensemble setups. Our official submission, fine-tuned BLIP with retrieval augmentation, ranked 5th overall based on both cosine similarity and LLM-as-a-judge scores. Post-submission experiments showed that ensemble captioning yielded the strongest captions across metrics. These findings demonstrate that even modest fine-tuning combined with retrieval augmentation can substantially improve Arabic captioning quality, which is significant in light of the limited resources for the language.

1 Introduction

Image captioning, the automatic generation of textual descriptions for visual content, has advanced significantly with the advent of large-scale vision–language models. While state-of-the-art systems achieve impressive performance in English and other high-resource languages, Arabic image captioning remains a challenging task due to its morphological complexity, limited annotated datasets, under-representation in multimodal benchmarks, lack of large-scale pretrained models, and tokenization compatibility issues. The difficulty is amplified in domain-specific contexts, where captions must reflect cultural, historical, and linguistic nuances accurately.

This study, conducted as part of ImageEval 2025 (Bashiti et al., 2025), focuses on a culturally and historically sensitive setting: generating Arabic captions for images related to the Palestinian Nakba. Producing accurate captions in this context requires not only linguistic fluency but also captions that remain faithful to historical narratives and avoid introducing misleading or invented details. Existing models often struggle with such specialized demands, leading us to assess which approaches perform best in this setting.

We present a comparative analysis of **AIN** (Heakl et al., 2025), **BLIP** (Li et al., 2022), and **Qwen 2.5 VL** (Bai et al., 2025) on the ImageEval 2025 Image Captioning Shared Task dataset. Each model is evaluated under four configurations: zero-shot with a RAG post-generation layer, fine-tuning, fine-tuning combined with RAG, and an LLM-based stacking ensemble for image captioning. The RAG component aims to improve domain relevance and factual grounding of the generated captions, while the stacking ensemble is designed to minimize errors by fusing captions produced by the top-performing models.

2 Background

2.1 Related Work

Prior work has explored transformer-based models, such as VIOLET (Mohamed et al., 2023), which employs a two-stage decoder for improved Arabic captioning. Additionally, multitask encoder–decoder approaches have been leveraged to enhance performance by leveraging action classification and pre-trained embeddings (Za'ter and Talafha, 2022).

2.2 Dataset

The dataset used in this study comprises 3,471 manually captioned images, primarily depicting events and scenes related to the Israeli–Palestinian conflict. It is divided into a training set of 2,718 images and a test set of 753 images. The training set was made available to participants for model development, while the test set was released later for automatic caption generation.

Each image is uniquely identified by its file-

[*] Equal contribution.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 419–431
November 8-9, 2025 ©2025 Association for Computational Linguistics

name (serving as its ID) and paired with a corresponding caption in a separate annotation file. The annotation file contains two columns—the textual description and the associated image filename—allowing direct mapping between captions and images.

The dataset is hosted on Hugging Face and distributed as part of the Image Captioning Shared Task 2025.

To expand our training data, we used Gemini-2.5-flash (Comanici et al., 2025) to paraphrase each caption twice, creating two additional captions per image. This allowed us to train with multi-reference captions. A custom Python script was developed to interact with the Gemini API, producing alternative Arabic captions that maintained the exact meaning of the originals. We tried various prompts on a small subset of the data, and the one that best preserved the original meaning was selected for generating the full dataset. This prompt, which ensured the quality and semantic consistency of the generated captions, is provided in Appendix A.1. Figure 1 shows an example image from the dataset with its human-written caption.

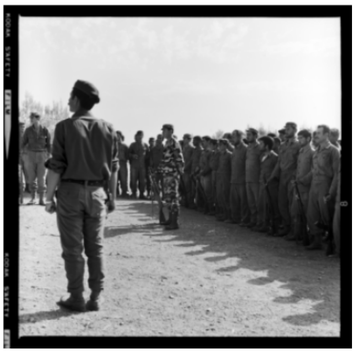

Figure 1: صوره لتدريب جنود جيش التحرير الفلسطيني في احد معسكرات التدريب.

3 System Overview

In our study, we conducted experiments utilizing three distinct models. The first is AIN (Arabic Inclusive Large Multimodal Model), which was developed by MBZUAI and trained on 3.6 million multimodal samples for English and Arabic captioning. The second is BLIP, a vision–language model extensively pre-trained on diverse web image–text datasets; we employed a publicly available variant from Hugging Face that had been fine-tuned on the Arabic Flickr8k dataset. The third is Qwen 2.5-VL, provided by the task organizers, which had already been fine-tuned and served as a benchmark for comparison in our analyses.

To enhance domain adaptability, we integrated a post-generation refinement layer inspired by (Ramos et al., 2023), adapting it to our task. A vector store was constructed from all Palestinian Nakba–related captions in the augmented training set, enabling the retrieval of examples with high semantic or lexical similarity to each generated caption. The retrieved examples, together with the original output, were provided to **Gemini-2.5-flash**, which revised the caption to align with the tone, style, and terminology of the retrieved material. The goal was to improve historical accuracy and stylistic consistency, while reducing obvious mistakes or hallucinations.

As the concluding phase in enhancing the quality of generated captions, we implemented an LLM-based stacking ensemble inspired by (Bianco et al., 2023). This approach involved providing captions generated by the models with the highest BLEU and cosine similarity scores to a meta-learner, utilizing the prompt detailed in Appendix A.3. This methodology facilitated the amalgamation of the most promising candidate captions, resulting in outputs that synthesized the strengths of multiple models while effectively mitigating their prevalent errors. The complete ensemble pipeline is illustrated in Figure 2.

We opted for the Gemini-2.5-flash model for both refinement and ensemble methodologies, owing to its generation quality. Additionally, its complementary tier rendered it a pragmatic choice for conducting iterative experimentation.

Building on the two methodologies described above, we designed four experiments aimed at systematically assessing and enhancing caption quality, namely:

- Zero-shot captioning with post-generation RAG
- Fine-tuned captioning
- Fine-tuned captioning with post-generation RAG
- LLM-based stacking ensemble

These approaches enabled a comparison of model performance across fine-tuning, retrieval-based contextual enhancement, combined methods, and ensemble learning.

Results reported for each configuration are post-submission results, obtained by submitting model outputs to the **CodaLab** evaluation server of the Image Captioning Shared Task 2025. Generated

GitHub Repository

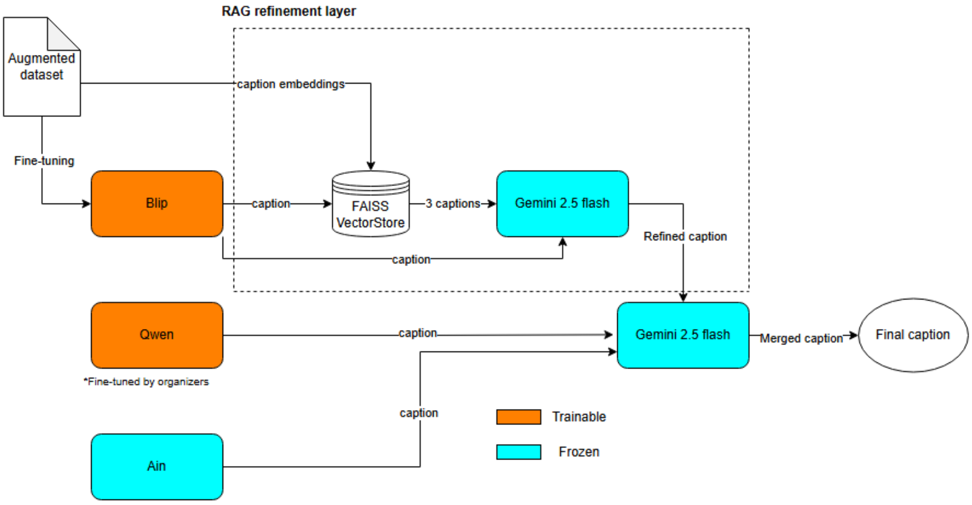

Figure 2: Pipeline diagram of the ensemble system. It integrates fine-tuned BLIP augmented with a RAG refinement layer, fine-tuned Qwen, and zero-shot Ain, with final caption fusion performed by Gemini.

captions were compared to ground-truth references using established metrics: BLEU (1–4) (Papineni et al., 2002), mean cosine similarity, and LLM-as-a-judge, (Wei et al., 2024), which, in this case, is **GPT-4o** (OpenAI et al., 2024). However, all official results are included in section 5.2.

4 Experiment Setup

All models under evaluation were fine-tuned using LORA (Hu et al., 2021) on the augmented Training dataset. All experiments were conducted using the Lightning.ai platform. For fine-tuning, AIN model was trained on NVIDIA L40S GPUs. while BLIP was trained on NVIDIA L4 GPUs. All training scripts were executed in a VS Code environment within Lightning.ai.

4.1 Data Preprocessing

Prior to training, all textual data was standardized using the Camel Tools library (Obeid et al., 2020) to ensure consistency and reduce orthographic variability in Arabic script. The preprocessing pipeline included Unicode and digit normalization, removal of diacritics, and orthographic unification (e.g., mapping آ, أ, إ to ا, ى to ي, and ة to ه). Elongation marks (Tatweel) were stripped, along with non-linguistic elements such as dates, numbers, and punctuation. Finally, whitespace was normalized by collapsing multiple spaces and trimming edges. These steps yielded clean, linguistically normalized input free of irrelevant tokens, leading to cleaner inputs and more dependable downstream results.

4.2 Model Fine-Tuning

The fine-tuning configuration is summarized in Table 1. **BLIP** was trained for three epochs with a batch size of 16 and a weight decay of 0.001. **AIN** was trained for ten epochs with a batch size of 64 and a higher weight decay to enhance its ability to generate high-quality Arabic captions. In contrast, **Qwen** was fine-tuned by the shared task organizers for fifteen epochs with a batch size of 16 using a cosine learning rate scheduler.

5 Results

5.1 Post-submission Results

This section reports the performance of the evaluated image captioning models under the four configurations described in Section 3, with zero-shot results included for comparison in Table 2. While traditional n-gram overlap metrics yielded relatively low scores, performance was higher on LLM-as-a-judge and cosine similarity, indicating that the generated captions were semantically related to the images but diverged from the ground truth in lexical choice.

Metric	BLIP	AIN-8B	Qwen-7B
BLEU-1 (mean)	3.58	3.5	**9.92**
BLEU-2 (mean)	1.57	1.17	3.23
BLEU-3 (mean)	0.95	0.64	1.90
BLEU-4 (mean)	0.78	0.44	1.33
Cosine Similarity (mean)	38.01	**59.69**	55.77
LLM-as-a-Judge	6.29	25.27	**27.11**

Table 2: Evaluation scores for zero-shot Model captioning

Model	Learning Rate	Batch Size	Epochs	Optimizer	Weight Decay	Loss Function
BLIP	2×10^{-4}	16	3	AdamW	0.001	Cross-Entropy
AIN	2×10^{-4}	64	10	AdamW	0.01	Cross-Entropy
Qwen 2.5	2×10^{-5}	16	15	AdamW	0	Cross-Entropy

Table 1: Fine-tuning hyperparameters for each evaluated model.

5.1.1 Zero-shot captioning with post-generation RAG

This experiment evaluated the effect of the RAG layer on zero-shot models without prior task-specific training. As shown in table 3, For BLIP, both cosine similarity and LLM-as-a-judge improved slightly, reaching **42.96** and **7.2**. For AIN and Qwen, only LLM-as-a-judge increased, with scores of **29.49** and **30.51**, while cosine similarity declined relative to the zero-shot baseline. Nonetheless, RAG improved BLEU_1 across all three models, with BLIP, AIN, and Qwen achieving **10.12**, **7.79**, and **10.28**, though the gain for Qwen was marginal. These findings suggest that RAG helped in some cases (like BLIP) but not in others, meaning the gains depend heavily on how each model integrates external context. Captioning examples are provided in Appendix B.

Metric	BLIP	AIN	Qwen-2.5 VL
BLEU-1 (mean)	10.12	7.79	**10.28**
BLEU-2 (mean)	3.73	3.25	4.42
BLEU-3 (mean)	2.31	2.0	2.75
BLEU-4 (mean)	1.86	1.41	1.89
Cosine Similarity (mean)	42.96	**55.15**	52.39
LLM-as-a-Judge	7.2	29.49	**30.51**

Table 3: Evaluation scores for zero-shot Model captioning with RAG.

5.1.2 Fine-Tuned model captioning

Fine-tuning improved model alignment with the Palestinian Nakba domain (Table 4), though the magnitude of improvement varied across models. BLIP demonstrated substantial gains, achieving a mean cosine similarity of **54.18** and an LLM-as-a-judge score of **22.99**. Qwen also improved, though less markedly, with a mean cosine similarity of **58.46** and an LLM-as-a-judge score of **30.82**. In contrast, AIN generalized poorly, reflecting weaker domain alignment, We suspect this may be because AIN was originally trained on broader multimodal data and struggled to adapt to the very specific Nakba-related captions, as both cosine similarity and LLM-as-a-judge scores declined. Captioning examples are documented in Appendix C.

Metric	BLIP	AIN	Qwen-2.5-VL
BLEU-1 (mean)	**21.40**	3.64	16.98
BLEU-2 (mean)	10.66	1.21	8.62
BLEU-3 (mean)	6.15	0.75	5.43
BLEU-4 (mean)	4.29	0.56	3.05
Cosine Similarity (mean)	54.18	52.92	**58.46**
LLM-as-a-Judge	22.99	15.66	**30.82**

Table 4: Evaluation scores for Fine-tuned Model captioning

5.1.3 Fine-tuned Captioning with Post-generation RAG

The integration of both domain adaptation techniques yielded a notable improvement in performance. As shown in table 5, BLIP showed only marginal gains over the raw fine-tuned model across n-gram overlap, cosine similarity, and LLM-as-a-judge scores, scoring **22.77**, **55.32** and **24.87** respectively. However, AIN exhibited a more nuanced increase in LLM-as-a-judge, accompanied by a slight decline in cosine similarity; however, its BLEU score increased significantly, rising to **8.25** compared to the raw fine-tuned model. By contrast, Qwen's performance declined slightly across both cosine similarity and LLM-as-a-judge metrics. Appendix D contains the captioning examples.

Metric	BLIP	AIN	Qwen-2.5-VL
BLEU-1 (mean)	**22.77**	8.25	14.23
BLEU-2 (mean)	11.29	3.2	7.44
BLEU-3 (mean)	6.34	2.02	5.09
BLEU-4 (mean)	4.39	1.41	3.63
Cosine Similarity (mean)	**55.32**	51.63	53.91
LLM-as-a-Judge	24.87	20.51	**26.52**

Table 5: Evaluation scores for Fine-tuned Model captioning with RAG

5.1.4 LLM-based stacking ensemble

To leverage complementary strengths across models, we fused captions generated by Zero-shot AIN, Fine-tuned Qwen, and Fine-tuned BLIP with RAG using the meta-learner described in Section 3. Although this ensemble did not achieve the highest BLEU score,only scoring a BLEU_1 score of **8.47**, it outperformed all non–zero-shot configurations in terms of semantic alignment and human-likeness, attaining the best cosine similarity **59.17** and LLM-as-a-judge **32.92** scores as shown in table 6. The captioning examples are presented in Appendix E.

Metric	Meta-Learner
BLEU-1 (mean)	8.47
BLEU-2 (mean)	4.08
BLEU-3 (mean)	2.29
BLEU-4 (mean)	1.5
Cosine Similarity (mean)	**59.17**
LLM-as-a-Judge	**32.92**

Table 6: Evaluation scores for LLM-based stacking ensemble (Meta-Learner)

Overall, the results highlight a clear performance hierarchy across the four approaches. RAG provided improvements in both zero-shot and fine-tuned settings, with its effect on zero-shot models being substantial, though still below the gains achieved through fine-tuning alone. When combined with fine-tuning, RAG yielded further gains across most models. Notably, while other models reached top performance in individual metrics, the ensemble consistently achieved near top results across most metrics, yielding the best overall performance on average.

5.2 Official Results

As mentioned in Section 1, our official results are based on the outputs of fine-tuned BLIP with RAG, which determined our ranking. The evaluation primarily relied on LLM-as-a-judge and cosine similarity metrics, yielding scores of **24.87** and **55.32**. Additionally, 5 percent of the test set was evaluated by humans using qualitative criteria, **cultural relevance**, **conciseness**, **completeness**, and **accuracy**, rated from 1 to 4, with definitions in Appendix F. Our model showed competitive performance, with conciseness achieving a score of **2.97** and cultural relevance **2.57**, while completeness and accuracy obtained scores of **2.13** and **2.23**, respectively, as shown in table 7.

Metric	Score
Cultural Relevance	2.57
Conciseness	2.97
Completeness	2.13
Accuracy	2.23

Table 7: Official human evaluation results for fine-tuned BLIP with RAG.

6 Conclusion

In conclusion, most reported results were obtained post-submission, whereas the official ranking relied exclusively on fine-tuned BLIP with a RAG layer, which achieved the highest BLEU score of **22.77**. The ensemble's meta-learner attained the top LLM-as-a-judge score of **32.92** and nearly matched zero-shot AIN in cosine similarity with **59.17**. The effect of RAG, however, varied across models: while it consistently acted as a refinement layer that enhanced outputs, its contribution was contingent on the strength of the underlying model.

7 Limitations

This study's limitations stem from computational and resource constraints. Conducted on the free tier of the Lightning.ai platform with only 15 GPU credits, our experiments were limited in scale and duration. This precluded exhaustive hyperparameter searches and constrained the number of training epochs for larger models. The post-generation RAG and ensemble layers, implemented with the rate-limited Gemini-2.5-flash API, required test set inferences to be batched across multiple days and reduced opportunities for extensive prompt engineering. Finally, while our LLM-based stacking ensemble achieved the best qualitative performance, its sequential inferences and reliance on an additional LLM meta-learner make it computationally expensive, resulting in high latency and memory demands. These factors limited its practicality for real-time, resource-constrained industrial deployment. In addition to these computational constraints, the ground-truth captions for the test data were hidden from participants, precluding the use of additional evaluation metrics that might have provided further insights into morphologically rich Arabic.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Ahlam Bashiti, Alaa Aljabari, Hadi Hamoud, Md. Rafiul Biswas, Bilal Shalash, Mustafa Jarrar, Fadi Zaraket, George Mikros, Ehsaneddin Asgari, and Wajdi Zaghouani. 2025. ImageEval 2025: The

First Arabic Image Captioning Shared Task. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Simone Bianco, Luigi Celona, Marco Donzella, and Paolo Napoletano. 2023. Improving image captioning descriptiveness by ranking and llm-based fusion. *Preprint*, arXiv:2306.11593.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Ahmed Heakl, Sara Ghaboura, Omkar Thawkar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman Khan. 2025. Ain: The arabic inclusive large multimodal model. *Preprint*, arXiv:2502.00094.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Inciarte, and Muhammad Abdul-Mageed. 2023. Violet: A vision-language model for Arabic image captioning with gemini decoder. In *Proceedings of ArabicNLP 2023*, pages 1–11, Singapore (Hybrid). Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681, Dubrovnik, Croatia. Association for Computational Linguistics.

Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. 2024. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *ArXiv*, abs/2408.13006.

Muhy Eddin Za'ter and Bashar Talafha. 2022. Benchmarking and improving arabic automatic image captioning through the use of multi-task learning paradigm. *Preprint*, arXiv:2202.05474.

A Prompt Templates

A.1 Prompt to Expand Training Dataset

هذه هي التسمية الأصلية للصورة: «orig» من فضلك، اكتب نوعين آخرين من التسمية باللغة العربية لوصف نفس الصورة، بصياغة مختلفة ولكن المعنى ذاته.

A.2 Prompt used to Train our models

Analyze the content of this image, which is potentially related to the Palestinian Nakba and Israeli occupation of Palestine, and provide a concise and meaningful caption in Arabic — about 15 to 50 words. The caption should reflect the scene's content and emotional context, and should be natural and culturally appropriate. Do not include any English, metadata, or titles — the caption must be in Arabic.

A.3 Prompt sent to the Meta learner in the LLM-based stacking ensemble method

You are an expert in image captioning. Your task is to merge three captions (one descriptive zero-shot caption and two domain-specific captions) into a single final caption.

Rules:

- Preserve and prioritize important domain-specific terminology and keywords from the domain captions, since they are valuable for matching hidden reference captions.
- Use the zero-shot caption to add descriptive details and improve fluency, but keep the result concise and natural.

- Conflict resolution rule:
  - If the domain captions conflict with each other or with the zero-shot caption, treat the zero-shot caption as the reference and adjust the domain information accordingly.
  - If the conflict is minor (e.g., synonyms or phrasing), prefer the domain wording.
- Avoid redundancy, contradictions, or overlong sentences.
- Keep the result concise (one short sentence if possible) and avoid redundancy.
- Output one final caption that balances accuracy, fluency, and semantic richness.

Captions to merge:

- Zero-shot caption: zero
- Domain-specific caption 1: domain1
- Domain-specific caption 2: domain2

Output: return a single caption that integrates the strengths of all three inputs, maximizes semantic overlap and cosine similarity with likely hidden reference captions. caption:

B Zero-shot with RAG Captioning Examples

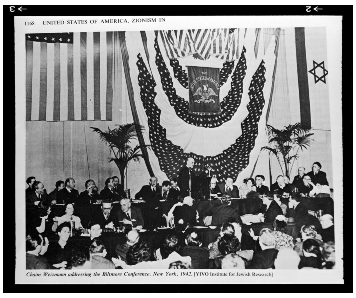
1168 UNITED STATES OF AMERICA, ZIONISM IN
Chaim Weizmann addressing the Biltmore Conference, New York, 1942. [YIVO Institute for Jewish Research]

Figure 3: Zero-shot with RAG on Example 1

AIN: الصورة تظهر مؤتمر بيلتمور في نيويورك عام 1942. يظهر فيها حاييم وايزمان يلقي خطابا امام عدد من الشخصيات الجالسة حول طاولات مستديرة. تبرز في الخلفية اعلام امريكية وعلم يحمل نجمة داود. الصورة من مكتبة ييفو للابحاث اليهودية.

Blip: صورة لمجموعة من الشباب الفلسطينيين يرتدون ملابس ملونة ويحملون الأعلام في مهرجان الشباب العربي بالجزائر

Qwen: تظهر الصورة اجتماعا في نيويورك عام 1942، حيث يتحدث الزعيم الصهيوني حاييم وايزمان عن القضية اليهودية وواقع الشعب الفلسطيني، وذلك في سياق المشروع الصهيوني. يظهر في الخلفية العلم الامريكي وعلم الحركة الصهيونية.

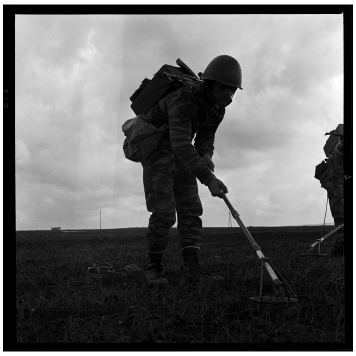

Figure 4: Zero-shot with RAG on Example 2

AIN: تظهر الصورة تدريب جنود من جيش التحرير الفلسطيني على سلاح مضاد للطيران عيار ملم في احد معسكرات التدريب.

Blip: صوره لرجل يحمل معدات عسكريه

Qwen: الصورة تظهر جندي الاحتلال وهو يقوم بتمشيط ارض خالية بحثا عن الالغام، ضمن جهود التعامل مع اثار الصراع في المنطقة

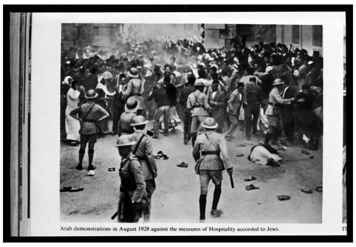
Arab demonstrations in August 1928 against the measures of Hospitality accorded to Jews.

Figure 5: Zero-shot with RAG on Example 3

AIN: صورة من عام 1928 لقمع قوات الأمن لمظاهرة عربية في المنطقة، احتجاجا على السياسات البريطانية الداعمة للاستيطان اليهودي. يظهر فيها رجال بزي عسكري مسلحين يواجهون المتظاهرين، حيث أسفرت المواجهات العنيفة عن سقوط أشخاص على الأرض.

Blip: سكان رفح يحيون الدبابات المصرية عند وصولها معبر رفح خلال حرب

Qwen: مظاهرات عربية في اغسطس 1928 احتجاجا علي سياسات الاحتلال البريطاني التي استهدفت حقوق ومقدسات المسلمين، وسط تواجد مكثف للقوات العسكرية التي قمعت المظاهرات.

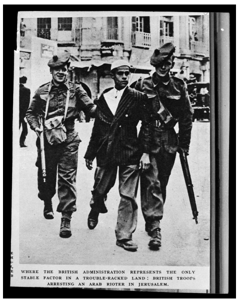
WHERE THE BRITISH ADMINISTRATION REPRESENTS THE ONLY STABLE FACTOR IN A TROUBLE-RACKED LAND: BRITISH TROOPS ARRESTING AN ARAB RIOTER IN JERUSALEM.

Figure 6: Zero-shot with RAG on Example 4

AIN: تظهر الصورة مشهدا من مدينة القدس عام 1948، حيث يقوم جنود الاحتلال البريطاني باعتقال متظاهر عربي. يظهر في الصورة ثلاثة جنود بريطانيين يرتدون الزي العسكري ويحملون البنادق، بينما يقتاد الرجل العربي الذي يرتدي ملابس تقليدية. يعكس هذا المشهد قمع الاحتلال البريطاني للمقدسيين خلال فترة الانتداب.

Blip: صوره لجنود الجيش المصري يسيرون في الشارع

Qwen: تظهر الصورة اعتقال جنود بريطانيين لرجل عراقي في القدس، وذلك خلال اضطرابات.

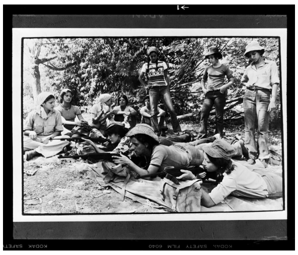

Figure 7: Zero-shot with RAG on Example 5

AIN: تظهر الصورة تدريب مجموعة من المقاتلين، بينهم نساء، على استخدام الأسلحة في أحد معسكرات الثورة الفلسطينية، حيث يتخذون وضعيات تدريب مختلفة على الأرض.

Blip: صوره لناس يستلقون على الارض

Qwen: تظهر الصورة تجمعا لمجموعة من الاشخاص يجلسون علي الارض. تاريخ الصورة ومصورها مجهولان.

C Fine-tuned Captioning Examples

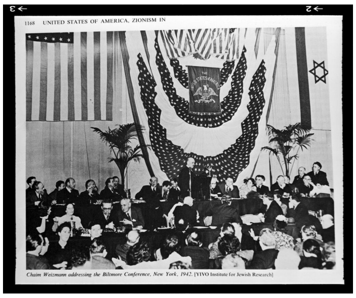
1168 UNITED STATES OF AMERICA, ZIONISM IN
Chaim Weizmann addressing the Biltmore Conference, New York, 1942. [YIVO Institute for Jewish Research]

Figure 8: Fine-tuned models on Example 1

AIN: تلتقط الصورة لحظة تاريخية في قاعة احتفالية في فلسطين عام 1942. في المقدمة، يجلس الحضور في ملابس رسمية في مدرج، يراقبون الحدث الذي يجري في المسرح. مركز الصورة هو الحدث الرئيسي الذي يجري في المسرح.

Blip: صورة لاحتفال عسكري لقوات الثورة الفلسطينية في بيروت

Qwen: صورة لاجتماع في نيويورك عام 1942

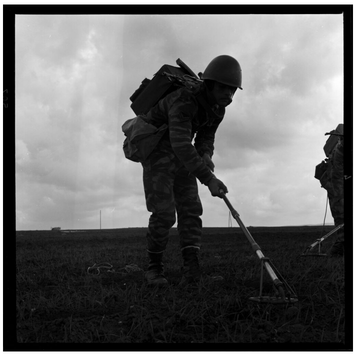

Figure 9: Fine-tuned models on Example 2

AIN: تلتقط الصورة مشهدا في منطقة عسكرية في منطقة القوقاز في عام 1992 خلال حرب الشيشان. في المقدمة، يظهر جندي يرتدي زي عسكري ويحمل مسدس مثبت في حزامه. يحمل جهازا يشبه جهاز الاستشعار في يديه، والذي يحتمل انه يستخدم في عمليات البحث عن المتفجرات.

Blip: صورة لتدريب مقاتلين الثورة الفلسطينية في احد معسكرات التدريب

Qwen: الصورة تظهر جنديًا يبحث عن الألغام في منطقة مستهدفة من قبل الاحتلال الإسرائيلي خلال حرب حزيران/ يونيو 1967.

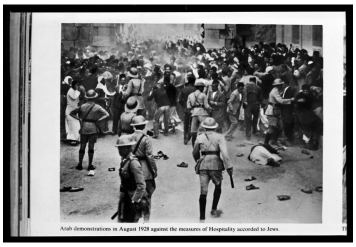
Arab demonstrations in August 1928 against the measures of Hospitality accorded to Jews.

Figure 10: Fine-tuned models on Example 3

AIN: تلتقط الصورة مشاهد مروعة من احتجاجات عربية في أغسطس عام 1928، احتجاجا على التدابير البريطانية التي تدعم الاستيطان اليهودي في المنطقة. في المقدمة، يظهر جندي بريطاني يركل متظاهرا عربيا بينما ترقد امرأة على الأرض في حالة صعبة.

Blip: صورة لاحتفال عسكري لقوات الثورة الفلسطينية في بيروت

Qwen: الصورة تظهر احتجاجًا عسكريًا من قبل السلطات الاستعمارية

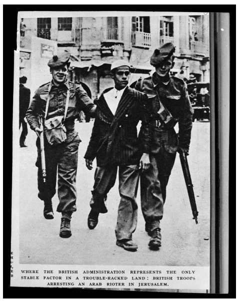
WHERE THE BRITISH ADMINISTRATION REPRESENTS THE ONLY STABLE FACTOR IN A TROUBLE-RACKED LAND: BRITISH TROOPS ARRESTING AN ARAB RIOTER IN JERUSALEM.

Figure 11: Fine-tuned models on Example 4

AIN: تلتقط الصورة مشهدا مثيرا في القدس عام 1948. في المقدمة، يقوم جنديان بريطانيان باحتجاز رجل عربي يرتدي ملابس تقليدية، أحدهما يحمل بندقية والآخر مسدسا. يظهر في الخلفية مشاة من المارة.

Blip: صورة لاحتفال عسكري لقوات الثورة الفلسطينية في بيروت

Qwen: الصورة تظهر جنديين بريطانيين يعتقمان شخصًا في القدس

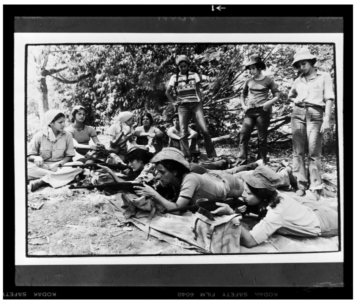

Figure 12: Fine-tuned models on Example 5

AIN: تلتقط الصورة مشهدا في الغابة حيث تشارك مجموعة من النساء في تدريبات عسكرية تكتيكية، يتخذن وضعيات مختلفة للرماية.

Blip: صورة لاحتفال عسكري لقوات الثورة الفلسطينية في بيروت

Qwen: الصورة تظهر مجموعة من الفتيات اللاتي يتدربن على إطلاق النار، ويظهر في الصورة 10 فتاة، حيث يظهر في الصورة 5 فتيات يحملن البنادق ويستعدن لإطلاق النار، بينما يجلس الباقي في الخلفية، ويظهر في الصورة شجرة خلف الفتيات.

Image	AIN	Blip	Qwen
Example 1	A picture of the Biltmore Conference in New York in 1942. Chaim Weizmann appears giving a speech to a number of figures sitting at round tables, with American flags and a flag with the Star of David in the background. The picture is from the YIVO Institute for Jewish Research.	A picture of a group of Palestinian youth in colorful clothes carrying flags at the Arab Youth Festival in Algeria.	A picture showing a meeting in New York in 1942, where the Zionist leader Chaim Weizmann speaks about the Jewish cause and the reality of the Palestinian people within the Zionist project, with the American flag and the Zionist movement's flag appearing in the background.
Example 2	A picture of Palestine Liberation Army soldiers training on an anti-aircraft weapon of a certain caliber in a training camp.	A picture of a man carrying military equipment.	A picture showing an occupation soldier combing an empty field in search of mines, within efforts to deal with the effects of conflict in the region.
Example 3	A picture from 1928 showing security forces suppressing an Arab demonstration in the region, protesting British policies supporting Jewish settlement. Armed men in military uniforms appear confronting demonstrators, where violent clashes occurred leading to the fall of individuals to the ground.	Rafah residents receiving Egyptian tanks entering Rafah crossing during the war.	Arab demonstrations in August 1928 rejecting the policies of the British occupation that targeted the rights and sanctities of Muslims, amid an intense military presence that suppressed the protests.
Example 4	A picture showing a scene from Jerusalem in 1948, where British occupation soldiers arrest an Arab protester. Three British soldiers in military uniforms carrying rifles appear, escorting an Arab man in traditional clothes, reflecting the suppression of Jerusalemites by the British during the Mandate period.	A picture of Egyptian army soldiers walking in the street.	A picture showing British soldiers arresting an Iraqi man in Jerusalem during the events.
Example 5	The photo shows a group of fighters, including women, being trained to use weapons in a Palestinian revolution camp, where they take different training positions on the ground.	A picture of people lying on the ground.	A picture showing a gathering of people sitting on the ground. The date and the photographer are unknown.

Table 8: English translations of captions generated by the models under the zero-shot with RAG configuration.

D Fine-tuned with RAG Captioning Examples

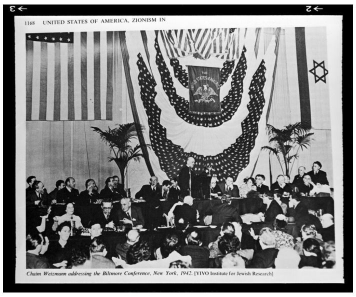
1160 UNITED STATES OF AMERICA. ZIONISM IN
Chaim Weizmann addressing the Biltmore Conference, New York, 1942. [YIVO Institute for Jewish Research]

Figure 13: Fine-tuned models with RAG on Example 1

AIN: تظهر الصورة جانبا من حدث تاريخي أقيم في قاعة احتفالات بفلسطين عام 1942. يجلس في المقدمة عدد من الحضور بملابس رسمية في مدرج، يتابعون الحدث الرئيسي الذي يعرض على المسرح. تبرز شخصية بارزة بين الحضور في الصف الأمامي.

Blip: صوره من احتفال عسكري لقوات الثوره الفلسطينيه في بيروت

Qwen: صورة توثق اجتماعاً في نيويورك عام 1942.

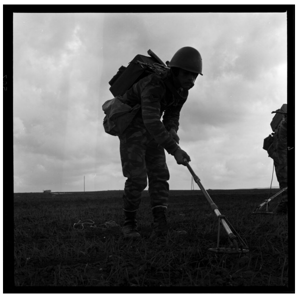

Figure 14: Fine-tuned models with RAG on Example 2

AIN: صورة لمشهد في منطقة عسكرية بالقوقاز عام 1992، تظهر جنديا يرتدي زيا عسكريا ويحمل مسدسا مثبتا في حزامه، كما يحمل جهازا يشبه جهاز استشعار يعتقد أنه يستخدم في عمليات البحث عن المتفجرات.

Blip: صوره لتدريب جنود جيش التحرير الفلسطيني في احد معسكرات التدريب

Qwen: صورة لجندي من قوات الاحتلال الإسرائيلي يبحث عن الألغام في منطقة خلال حرب حزيران/ يونيو 1967.

Image	AIN	Blip	Qwen
Example 1	The picture captures a historical moment in a ceremonial hall in Palestine in 1942. In the foreground, the attendees sit in formal clothing in a tiered seating, observing the event taking place on the stage. The center of the picture is the main event taking place on the stage.	A picture of a military celebration of the Palestinian revolution forces in Beirut.	A picture of a meeting in New York in 1942.
Example 2	The picture captures a scene in a military area in the Caucasus region in 1992 during the Chechen war. In the foreground, a soldier appears wearing a military uniform and carrying a pistol fixed in his belt. He is holding a device resembling a sensor in his hands, which is likely used in operations for detecting explosives.	A picture of the training of the Palestinian revolution fighters in one of the training camps.	The picture shows a soldier searching for mines in an area targeted by the Israeli occupation during the June 1967 war.
Example 3	The picture captures horrific scenes from Arab protests in August 1928, protesting the British measures supporting Jewish settlement in the region. In the foreground, a British soldier is kicking an Arab protester while a woman lies on the ground in a difficult state.	A picture of a military celebration of the Palestinian revolution forces in Beirut.	The picture shows a military protest by the colonial authorities.
Example 4	The picture captures an exciting scene in Jerusalem in 1948. In the foreground, two British soldiers are detaining an Arab man wearing traditional clothing, one of them holding a rifle and the other a pistol. In the background, pedestrians appear.	A picture of a military celebration of the Palestinian revolution forces in Beirut.	The picture shows two British soldiers arresting a person in Jerusalem.
Example 5	The picture captures a scene in the forest where a group of women are participating in tactical military training, taking different shooting positions.	A picture of a military celebration of the Palestinian revolution forces in Beirut.	The picture shows a group of girls who are training in shooting, and the picture shows 10 girls, where in the picture 5 girls carry rifles and are preparing to shoot, while the rest are sitting in the background, and in the picture there is a tree behind the girls.

Table 9: English translations of captions generated by the models under the fine-tuning configuration.

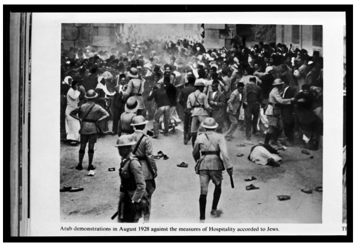
Arab demonstrations in August 1928 against the measures of Hospitality accorded to Jews.

Figure 15: Fine-tuned models with RAG on Example 3

AIN: تلتقط الصورة مشاهد مروعة من احتجاجات عربية في أغسطس عام 1928، ضد سياسات الانتداب البريطاني في المنطقة. يظهر جندي بريطاني وهو يركل متظاهرا في الشارع، بينما ترقد امرأة على الأرض في حالة إعياء.

Blip: صوره من احتفال عسكري لقوات الثوره الفلسطينيه في بيروت

Qwen: تُظهر الصورة احتجاجًا عسكريًا تنفذه السلطات الاستعمارية الصهيونية.

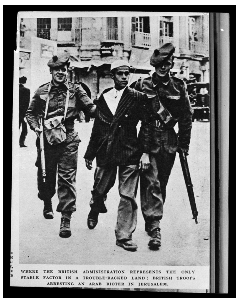
WHERE THE BRITISH ADMINISTRATION REPRESENTS THE ONLY STABLE FACTOR IN A TROUBLE-RACKED LAND: BRITISH TROOPS ARRESTING AN ARAB RIOTER IN JERUSALEM.

Figure 16: Fine-tuned models with RAG on Example 4

AIN: تظهر الصورة مشهدا من القدس عام 1948، خلال فترة الانتداب البريطاني. في المقدمة، يظهر جنديان بريطانيان بالزي الرسمي يقومان باحتجاز رجل وعربة، أحدهما يحمل بندقية والآخر مسدسا.

Blip: صوره من احتفال عسكري لقوات الثوره الفلسطينيه في بيروت

Qwen: صورة تُظهر جنديين بريطانيين يعتقلان شخصًا في القدس أثناء حرب.

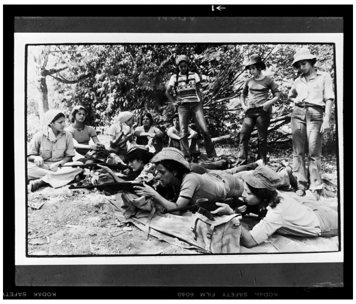

Figure 17: Fine-tuned models with RAG on Example 5

AIN: تظهر الصورة تدريب مقاتلات شابات ومقاتلين من قوات الثورة الفلسطينية على الرماية والتدريبات التكتيكية في أحد معسكرات التدريب بلبنان.

Blip: صوره من احتفال عسكري لقوات الثوره الفلسطينيه في بيروت

Qwen: صورة لمجموعة من عشر فتيات يتدربن على إطلاق النار، حيث تحمل خمس منهن بنادق ويستعددن للإطلاق، بينما تجلس الأخريات في الخلفية، وتظهر شجرة خلف الفتيات.

E LLM-based stacking Ensemble Captioning Examples

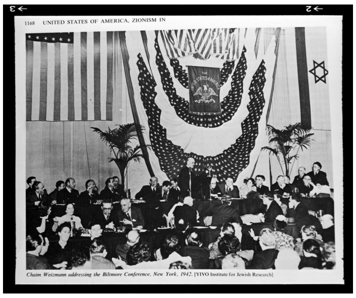
1168 UNITED STATES OF AMERICA, ZIONISM IN
Chaim Weizmann addressing the Biltmore Conference, New York, 1942. [YIVO Institute for Jewish Research]

Figure 18: LLM-based stacking Ensemble on Example 1

Meta Learner: صورة بالابيض والاسود لمؤتمر بيلتمور في نيويورك عام 1942، يظهر فيها شيمون وايزمان يلقي خطاباً أمام مجموعة من الرجال، مع وجود علم أمريكي كبير وعلم يحمل نجمة داود في الخلفية.

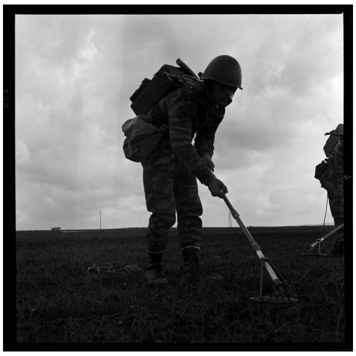

Figure 19: LLM-based stacking Ensemble on Example 2

Meta Learner: تظهر الصورة جنديين من جيش التحرير الفلسطيني يرتديان ملابس تمويهية، أحدهما يحمل جهازا يشبه المجرفة ويبحث عن الألغام في حقل عشبي ضمن منطقة مستهدفة من الاحتلال الإسرائيلي خلال حرب حزيران/ يونيو 1967.

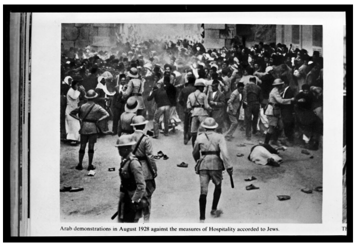
Arab demonstrations in August 1928 against the measures of Hospitality accorded to Jews.

Figure 20: LLM-based stacking Ensemble on Example 3

Meta Learner: تُظهر الصورة احتجاجًا عسكريًا متوترًا من عام 1928، حيث تواجه السلطات الاستعمارية المتظاهرين العرب اعتراضًا على إجراءات الضيافة الممنوحة للمستوطنين اليهود، مع وجود رجال يرتدون زي الجيش ويحملون أسلحة وأشخاص ملقون على الأرض.

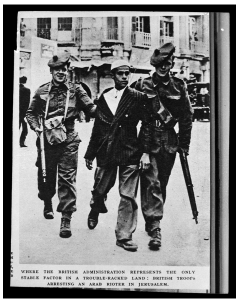
WHERE THE BRITISH ADMINISTRATION REPRESENTS THE ONLY STABLE FACTOR IN A TROUBLE-RACKED LAND: BRITISH TROOPS ARRESTING AN ARAB RIOTER IN JERUSALEM.

Figure 21: LLM-based stacking Ensemble on Example 4

Meta Learner: تظهر الصورة جنوداً بريطانيين يرتدون زياً عسكرياً ويعتقلون متظاهراً عربياً يرتدي ملابس تقليدية في القدس عام 1948.

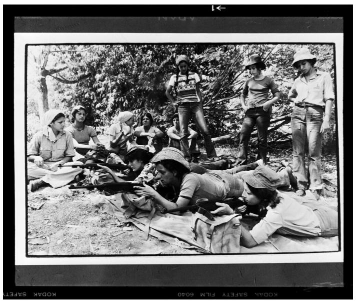

Figure 22: LLM-based stacking Ensemble on Example 5

Meta Learner: تظهر الصورة مجموعة من الفتيات، يرتدين ملابس غير رسمية، يشاركن في تدريب عسكري على إطلاق النار لقوات الثورة الفلسطينية في بيروت، حيث تحمل خمس منهن بنادق ويستعدن لإطلاق النار بينما تجلس البقية في الخلفية، مع شجرة خلفهن.

Image	AIN	Blip	Qwen
Example 1	The picture shows a side of a historical event held in a ceremonial hall in Palestine in 1942. In the foreground, several attendees in formal clothing sit in a tiered seating, following the main event presented on the stage. A prominent figure stands out among the attendees in the front row.	A picture from a military celebration of the Palestinian revolution forces in Beirut.	A picture documenting a meeting in New York in 1942.
Example 2	A picture of a scene in a military area in the Caucasus in 1992, showing a soldier wearing a military uniform and carrying a pistol fixed to his belt, also carrying a device resembling a sensor believed to be used in operations for detecting explosives.	A picture of the training of soldiers of the Palestinian Liberation Army in one of the training camps.	A picture of a soldier from the Israeli occupation forces searching for mines in an area during the June 1967 war.
Example 3	The picture captures horrific scenes from Arab protests in August 1928, against the policies of the British mandate in the region. A British soldier is shown kicking a protester in the street, while a woman lies on the ground in a state of exhaustion.	A picture from a military celebration of the Palestinian revolution forces in Beirut.	The picture shows a military protest carried out by the Zionist colonial authorities.
Example 4	The picture shows a scene from Jerusalem in 1948, during the British mandate. In the foreground, two British soldiers in official uniform are shown detaining a man and a cart, one of them holding a rifle and the other a pistol.	A picture from a military celebration of the Palestinian revolution forces in Beirut.	A picture showing two British soldiers arresting a person in Jerusalem during a war.
Example 5	The picture shows training of young female fighters and fighters from the Palestinian revolution forces in shooting and tactical training in one of the training camps in Lebanon.	A picture from a military celebration of the Palestinian revolution forces in Beirut.	A picture of a group of ten girls training on shooting, where five of them are carrying rifles and preparing to shoot, while the others sit in the background, and a tree appears behind the girls.

Table 10: English translations of captions generated by the models under the fine-tuning with RAG configuration.

F Human Assessment metric Definitions

Cultural Relevance – Measures whether the description reflects cultural specificity and provides contextual information related to the scene.

Conciseness – Assesses whether the description conveys information directly and succinctly, without unnecessary repetition or dispersion of details.

Completeness – Evaluates the extent to which the description covers all aspects of the image, including events, entities, and relevant elements.

Accuracy – Measures whether the description contains correct information, free from factual or conceptual errors.

Image	Meta Learner
Example 1	A black-and-white picture of the Biltmore Conference in New York in 1942, in which Chaim Weizmann appears giving a speech in front of a group of men, with a large American flag and a flag with the Star of David in the background.
Example 2	The picture shows two soldiers from the Palestinian Liberation Army wearing camouflage clothing, one of them holding a device resembling a shovel and searching for mines in a grassy field within an area targeted by the Israeli occupation during the June 1967 war.
Example 3	The picture shows a tense military protest from the year 1928, where the colonial authorities confronted the Arab demonstrators objecting to the hospitality measures granted to the Jewish settlers, with men wearing military uniforms carrying weapons and people lying on the ground.
Example 4	The picture shows British soldiers wearing military uniforms arresting an Arab demonstrator wearing traditional clothing in Jerusalem in 1948.
Example 5	The picture shows a group of girls, wearing informal clothing, participating in military training on shooting for the Palestinian revolution forces in Beirut, where five of them are carrying rifles and preparing to shoot while the rest are sitting in the background, with a tree behind them.

Table 11: English translations of captions generated by the models under the LLM stacking ensemble configuration.