

YNWA_AZ at SemEval-2026 Task 1: Bridging the Semantic-Visual Gap: Multimodal Humor Generation

Mohammad Erfan Zare¹, Tahere Abbasi², Hadi Veisi¹, Sayin Ala¹, Hanieh Naderi¹

¹University of Tehran, Iran

²University of Shiraz, Iran

{erfan.zare, h.veisi, sayin.ala81, hanieh.naderi}@ut.ac.ir
tahere7abbasii@gmail.com

Abstract

Developing Computational Humor systems at a multilingual and multimodal scale requires transcending simple text generation paradigms to focus on intent and context understanding. In this study, we address two key limitations in Foundation Models: *Association Failure* in textual tasks, which prevents the formation of coherent semantic links between incongruous concepts, and *Temporal Blindness* in video processing, which disrupts narrative comprehension. To tackle these challenges, we propose a unified architecture comprising an *Intent-Aware RAG* system for mitigating linguistic gaps across English, Spanish, and Chinese, and a *Cascaded Visual Perception* pipeline for modeling the narrative structure of video content. A key innovation of this work is the utilization of small language models (TinyLlama) as a *Semantic Denoise Filter*, converting noisy visual signals into structured, coherent textual representations. Experimental results demonstrate that this modular architecture reduces cultural-semantic gaps in certain languages and produces outputs that generally align better with human humor preferences, though highly nuanced languages still present a challenge.

1 Introduction

Humor generation is a complex cognitive challenge that requires models to move beyond semantic fidelity and engage in “creative semantic deviation.” We address the SemEval-2026 Task 1 on Multilingual and Multimodal Humor Generation (Castro et al., 2026), which targets these complexities across two subtasks. Subtask A evaluates text-based humor generation with specific constraints (keywords and news headlines) across English, Spanish, and Chinese. Subtask B evaluates multimodal humor by requiring models to generate comedic

captions for GIF images. Developing robust systems for these tasks is crucial to overcoming foundational AI limitations, specifically *Association Failure* (the inability to bridge incongruous concepts logically) and *Tone Shift Failure* (the inability to distinguish satire from toxic content).

To tackle these issues without the high computational cost of direct fine-tuning, we propose a unified *Retrieval-Augmented Generation (RAG)* architecture (Lewis et al., 2020) that treats data as “Enriched Semantic Structures.” For linguistic challenges, our *Intent-Aware RAG* dynamically retrieves structural blueprints based on user intent. To address *Temporal Blindness* in multimodal tasks—where models treat videos as disjointed static frames—we introduce a *Cascaded Visual Perception* pipeline. This pipeline leverages BLIP (Li et al., 2022) for frame extraction and TinyLlama (Zhang et al., 2024) as a *Semantic Denoise Filter* to compress scattered visuals into a coherent narrative.

Participating in this task revealed that while hardware-efficient, sub-10B parameter models can perform well when guided by targeted retrieval, notable cross-cultural hurdles remain. Our system achieved competitive results, tying for 1st place in Task B2 and securing 2nd place in English Task A. However, quantitative and qualitative analyses demonstrate that while our local RAG pipeline performs comparably to foundation models in English and multimodal structures, it struggles autonomously with the deep homophonic wordplay (*Xieyin*) inherent to Chinese humor when evaluated by human judges. Our complete implementation and reproducible configurations are publicly

available.¹

2 Related Work

Early research in computational humor primarily framed the problem as a classification or ranking task. A series of SemEval shared tasks progressively pushed the community beyond binary detection, reconceptualizing humor as a continuous spectrum (Potash et al., 2017), exploring its structural relationship with irony (Mikhalkova et al., 2018), and applying the Incongruity Theory to micro-edited texts (Hossain et al., 2020). Furthermore, pre-trained language models were shown to possess demographic blind spots, struggling to differentiate genuine humor from offensive content (Meaney et al., 2021). While these foundational studies mapped the theoretical boundaries of the field, they remained strictly focused on *detecting* humor rather than *generating* it.

Similar single-stage limitations exist in multimodal humor generation. Most vision-language approaches rely either on simple captioning pipelines or monolithic end-to-end models. Standard Vision-Language Models (VLMs) typically provide literal descriptions of visual frames, missing the pragmatic incongruity required for visual humor (Hwang and Shwartz, 2023). Moreover, while end-to-end models handle static images effectively, recent benchmarks prove they struggle with temporal reasoning in video sequences, treating frames as isolated images—a phenomenon we term *Temporal Blindness* (Li et al., 2024).

3 System Overview: Task A (Textual Humor)

Task A requires generating text-based jokes in English, Spanish, and Chinese that satisfy specific constraints: either incorporating two designated words or relating to a given news headline. To address the challenges of Association and Tone Shift failures in textual humor, we developed an Intent-Aware RAG architecture. Figure 1 illustrates the overall system flow, which begins with data segregation based on operational function, followed by language-

specific preprocessing, embedding-based retrieval, and finally generation via Llama-3.

3.1 Stage I: Data Preparation and Representation

We split our data based on its role: one part for generation and another for contrastive evaluation. To keep the humor culturally natural, we used real public datasets rather than machine-translated text.

For the *Keyword-to-Joke* scenario, we built a **Positive Corpus** using SemEval-2017 Task 7 (Miller et al., 2017) and Short-Jokes (Amoudgl, 2018) for English, the HAHA Corpus (Chiruzzo et al., 2019) for Spanish, and the Chinese-Humor dataset (Tseng and Tang, 2018) for Chinese. To improve retrieval precision without the high computational cost of fine-tuning, we avoided standard flat indexing. Instead, we structured this corpus into **anchor-positive pairs**. By mapping text anchors directly to their comedic structures, we reduced the vector distance in the embedding space, making the retrieval step much more accurate.

For the *News-to-Satire* task and the evaluation phase, we needed a baseline of non-humorous text. We created a **Discriminator Corpus** using real news headlines as **Negative Samples**. This gave our multi-agent “Judge” module clear contrastive examples, teaching it that formal news reporting is not a joke.

Finally, we applied specific fixes for each language’s limitations. In English, we used a safety filter to separate actual puns from toxic text. In Spanish, we removed redundant jokes using a cosine similarity threshold of 0.95. For Chinese, we used Jieba (Sun, 2015) for segmentation but intentionally kept the stop-words, as these particles are essential for comedic timing.

3.2 Stage II: Intent-Aware Retrieval

The retrieval engine operates by first classifying the incoming input into one of two intent categories: *Keyword-to-Joke* or *News-to-Satire*. This intent signal governs which sub-index of the vector store is queried. Embedding-based dense retrieval is performed using language-specific models: BAAI/bge-m3 (Chen et al., 2024) for English and Spanish,

¹<https://github.com/mr-zare/SemEval-2026-task1>

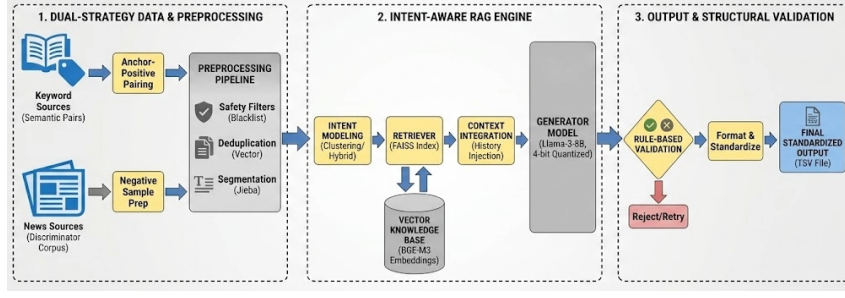


Figure 1: Architectural Flowchart for Task A (Textual Humor Generation).

and BAAI/bge-large-zh-v1.5 (BAAI, 2024) for Chinese. The top- K retrieved structures are passed as enriched context to the generation stage.

3.3 Stage III: Structured Generation

The enriched semantic context retrieved in Stage II is forwarded to Llama-3-8B (Dubey et al., 2024) for final joke generation. Rather than prompting the model with raw input, the generator receives a structured prompt that encodes the retrieved incongruity patterns alongside culturally-conditioned directives. The model is instructed to adopt an *Idea-First* methodology: it must leverage the retrieved context to forge hidden semantic associations before committing to a surface form.

4 System Overview: Task B (Multimodal Humor)

Task B requires generating humorous captions for GIF images in English, with two variants: B1 generates captions using only the GIF, while B2 completes a given text prompt using both the GIF and the prompt. In this task, the challenge transcends mere “text generation”; it constitutes an effort to decipher a silent visual narrative. Figure 2 depicts the complete architectural flow.

4.1 Stage I: Cascaded Visual Perception and Denoising

In this stage, we convert raw GIF videos into structured text. Since processing every single frame is computationally redundant, we uniformly sample exactly six evenly spaced keyframes per GIF. We caption each frame independently using the Salesforce/blip-image-captioning-large model (Li et al., 2022). For the decoding

step, we applied standard beam search with $\text{num_beams}=3$, $\text{max_new_tokens}=40$, and $\text{top_p}=0.9$ to maintain concise outputs and limit hallucination.

However, raw frame-by-frame captions naturally result in a disjointed list of events. To resolve this, we pass the combined captions through TinyLlama-1.1B (Zhang et al., 2024). The primary justification for selecting TinyLlama is hardware constraints: Llama-3 already consumes most of our memory, requiring a lightweight model to summarize the scattered captions into a single coherent paragraph without exceeding our 16GB VRAM limit. We call this *Temporal Compression*, transforming a looping video into a simple text prompt.

4.2 Stage II: Context-Aware Generation

The denoised narrative produced in Stage I is fed as context to the main language model (Llama-3-8B).

Task B1 (Roast Mode): We employ Strict Persona Prompting to shift the model from a descriptive state to an interpretive mode, engaging in “humorous judgment” of the visual reality.

Task B2 (Multimodal Completion): Integrating user text with the GIF narrative is achieved through Hybrid Query Construction. We concatenate the TinyLlama narrative with the input prompt and use our RAG pipeline to retrieve the single most relevant humor scenario ($k = 1$) from our curated English corpus. Each entry in this reference database consists of two primary fields: the *setup* (the contextual anchor) and the *punchline* (the humorous resolution). This allows the generator to ground its output in proven patterns of semantic incongruity, ensuring the final completion

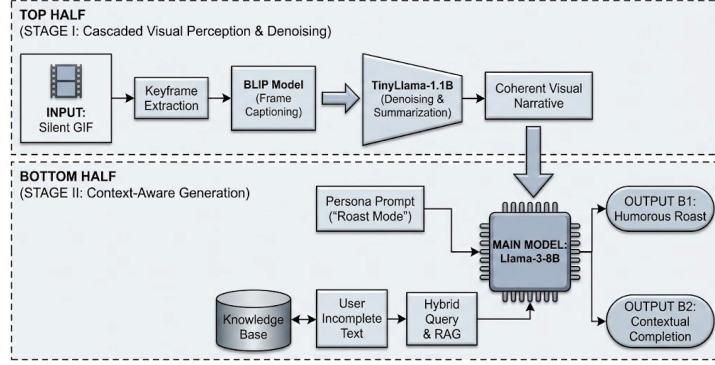


Figure 2: Architectural Flowchart for Task B (Multimodal Humor). The Cascaded Perception Pipeline (Stage I) converts noisy visual signals from BLIP into a clean narrative using TinyLlama.

remains both context-faithful and comedically effective.

5 Cross-Cultural Prompt Engineering and Multi-Agent Quality Assurance

A fundamental vulnerability in cross-lingual generation tasks is the assumption of semantic equivalence; translating a humorous premise directly from English to Chinese or Spanish typically results in catastrophic semantic degradation. To circumvent this, we abandoned zero-shot translation in favor of Culturally-Conditioned Persona Prompting coupled with a Multi-Agent Evaluation Framework.

5.1 The Generator: Constraint-Based Association

The generation prompt forces the LLM to adopt an Idea-First methodology. The model must utilize the retrieved context to forge hidden semantic associations. Prompts are localized (strategies detailed in Appendix A.4):

- English: Optimized for situational irony and concise setup-punchline delivery.
- Chinese: Strictly instructed to leverage homophonic wordplay (Xieyin / 谐音梗).
- Spanish: Calibrated to produce Doble Sentido (double entendre) and Chispa.

5.2 Multi-Agent Pipeline

As described by Zheng et al. (Zheng et al., 2023), using LLMs as judges can provide scalable evaluation. To prevent “Humor Hallucination”, we implemented a dual-agent auditing architecture:

- The Gatekeeper: A zero-tolerance classifier detecting hate speech and taboos.
- The Cultural Critic: A secondary evaluator that scores candidates across four language-specific dimensions (Humor, Irony, Relevance, Structure), detailed in Appendix A.4.

6 Experimental Setup

6.1 Data, Splits, and Hyperparameters

Across all tasks, we utilized the official SemEval-2026 training and development splits strictly for system tuning, prompt refinement, and RAG index construction. The test set was reserved exclusively for final generation. To accommodate the strict hardware constraint of a single 16GB VRAM GPU, the Llama-3-8B generator was deployed using 4-bit quantization via the BitsAndBytes library. Generation temperature was dynamically adjusted: set to 0.7 for creative humor generation, and 0.1 for the Multi-Agent evaluators.

6.2 Retrieval Engine Selection

Hardware constraints necessitated highly efficient embedding models. Based on our benchmarking, we selected BAAI/bge-m3 (Chen et al., 2024) for English and Spanish, and the monolingual BAAI/bge-large-zh-v1.5 (BAAI, 2024) for Chinese. A detailed comparative analysis is provided in Appendix A.3.

7 Results

7.1 Official Shared Task Leaderboard

The final standing of our system in the official human evaluation phase is presented in Table 1.

Table 1: Official Human Evaluation Leaderboard. The table reports the overall Rating and the 95% Confidence Interval (CI).

Task	Rank	System	Rating	95% CI
A-En	1	baseline	1081	[1045, 1110]
	2	(Ours)	1029	[1001, 1053]
A-Es	1	baseline	1140	[1098, 1182]
	8	(Ours)	985	[941, 1012]
A-Zh	1	baseline	1053	[1015, 1090]
	16	(Ours)	888	[845, 933]
B1	1	baseline	1124	[1084, 1164]
	3	(Ours)	1047	[1012, 1079]
B2	1	baseline	1022	[980, 1055]
	1	(Ours)	1035	[1006, 1069]

Our architecture demonstrated competitive performance across multiple tracks. In the multimodal domain, we secured a tied 1st place rank in Task B2 and 3rd place in Task B1. For textual humor (Task A), we achieved 2nd place in English.

When analyzing the 95% confidence intervals (CIs), a clear statistical divide emerges. For English text (Task A-En) and multimodal completion (Task B2), our model’s CIs strongly overlap with the baseline. This overlap suggests that our hardware-efficient, local RAG architecture effectively matches the comedic quality of massive foundation models in these specific domains.

Conversely, for Spanish, Chinese, and Task B1, there is no CI overlap between our system and the baseline. This lack of overlap indicates a statistically significant performance gap. As discussed further in our error analysis (Appendix A.8), the drop in Spanish (Rank 8) and Chinese (Rank 16) highlights the extreme difficulty of zero-shot cross-cultural humor alignment, proving that internal LLM-based evaluators cannot yet perfectly proxy human comedic judgment in highly nuanced languages.

7.2 Ablation Study: The Impact of RAG and Semantic Denoising

It is imperative to note that our baselines purposefully exclude large proprietary models (e.g., GPT-4) from the core architecture. Therefore, our ablation focuses on the architectural impact of our pipeline components applied to a standard open-weight model.

To empirically validate our architecture, we evaluated the baseline model in a zero-shot configuration (without RAG context and intermediate denoising). Table 2 illustrates the significant degradation when these components are removed.

Table 2: Ablation Study Results. Bypassing RAG and TinyLlama degrades context.

Task	Setup	Humor	Faith.
A	Zero-shot	2.15	1.85
	Ours (RAG)	4.38	4.97
B	Raw BLIP	2.41	3.10
	Cascaded	4.12	4.65

8 Conclusion

This research demonstrates that “Humor Generation” is not merely a linguistic issue but requires a precise architecture for information flow management. By employing Intent-Aware RAG, we helped capture the user intent behind keywords, and through the Cascaded Perception Pipeline (BLIP + TinyLlama), we mitigated the temporal blindness of text models regarding video content.

Future Work

While the generation module exhibited robust performance, the visual perception module showed limitations in processing highly complex GIF sequences. Future iterations will focus on optimizing the BLIP decoding parameters, specifically increasing the `repetition_penalty`. Furthermore, to address the cultural gaps identified in the error analysis, future work will explore integrating language-specific adapter modules (LoRA) rather than relying solely on RAG for cross-lingual tasks.

Limitations

While our Cascaded RAG framework demonstrates competitive performance within strict hardware constraints, it exhibits distinct limitations. First, our visual perception pipeline occasionally struggles with highly complex, rapid-motion GIF sequences, leading to generic or repetitive intermediate narratives that degrade the final humor output.

More critically, a significant methodological limitation was revealed regarding automated evaluation. Despite utilizing Culturally-Conditioned Persona Prompting, our models (specifically the sub-10B parameter Llama-3) exhibited a notable performance gap in human evaluations for highly nuanced, non-Latin languages like Chinese (Rank 16), compared to English (Rank 2). While our internal multi-agent evaluator (LLM-as-a-judge) indicated structural adherence for Chinese outputs, this did not fully align with human judgments. This mismatch highlights that internal LLM scores in cross-lingual humor generation should be treated strictly as diagnostic signals for formatting, rather than true reflections of comedic quality.

References

- Santiago Castro, Luis Chiruzzo, Santiago Góngora, Salar Rahili, Naihao Deng, Ignacio Sastre, Victoria Amoroso, Guillermo Rey, Aiala Rosá, Guillermo Moncecchi, J. A. Meaney, Juan José Prada, and Rada Mihalcea. 2026. SemEval-2026 Task 1: MWAHAHA, Models Write Automatic Humor And Humans Annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.
- A. Amoudgl. 2018. Short jokes dataset. *Kaggle*.
- BAAI. 2024. FlagEmbedding: Bge-large-zh-v1.5. *HuggingFace*.
- ByteDance. 2016. Toutiao news dataset. *Public Dataset*.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Luis Chiruzzo, Santiago Castro, Mathias Etcheverry, Diego Garat, Juan-Manuel Prada, and Aiala Rosá. 2019. Overview of HABA at IberLEF 2019: Humor analysis based on human annotation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*.
- Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Michael Günther, Jack Schneller, and Aleksandra Piktus. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.
- Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. Semeval-2020 Task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758.
- EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14337–14352.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.
- Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. *Frontiers in Artificial Intelligence*, 6:1162586.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, volume 33, pages 9459–9474.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023a. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Xianming Li and Jing Li. 2023b. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yina Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Xin Ping, et al. 2024. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119.
- Thomas Mesnard et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Elena Mikhalkova, Yuri Karyakin, Alexander Voronov, Dmitry Grigoriev, and Artem Leoznov. 2018. PunFields at SemEval-2018 task 3: detecting irony by tools of humor analysis. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 541–545.
- Tristan Miller, Christian F Hempelmann, and Iryna Gurevych. 2017. Semeval-2017 task 7: Detection and interpretation of english puns. In *Proceedings of SemEval-2017*, pages 58–68.
- MixedBread.ai. 2024. Mxbai-embed-large-v1. *HuggingFace*.
- NetEase Youdao. 2024. Bcembedding: Bilingual and crosslingual embedding for rag. *GitHub*.
- Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 Task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57.
- Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.
- Junyi Sun. 2015. Jieba: Chinese word segmentation module. *GitHub repository*.
- Sam Tseng and Y Tang. 2018. Construction of the chinese humor dataset. In *ROCLING*.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, volume 33, pages 5776–5788.
- Liang Xu et al. 2020. Clue: A chinese language understanding evaluation benchmark. In *COLING*.
- Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.
- Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.
- Lianmin Zheng et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Supplementary Materials

This appendix provides technical details regarding prompt configurations, qualitative outputs, and retrieval performance to support the reproducibility of our results.

A.1 Internal Generation Assessment (Task A)

We conducted an internal evaluation using our multi-agent framework to assess the model’s generation quality prior to official submission. Results are reported in Table 3.

An observation from the Chinese language results is noteworthy. While the automated evaluator assigned relatively high scores to the local RAG pipeline (internal Humor and Irony scores of 3.56 and 4.08, respectively), these figures differed significantly from the official human rankings. This suggests that the RAG pipeline primarily improved structural consistency and formatting in Chinese. However, the lack of alignment with human judgment indicates that these improvements did not translate into perceived comedic quality. Therefore, we view the LLM-based scores as diagnostic indicators of structural adherence rather than evidence of cultural alignment or performance superiority.

A.2 Multimodal Quantitative Performance (Task B)

The performance of the cascaded pipeline in visual scenarios is summarized in Table 4.

As shown in Table 4, the cascaded architecture (BLIP + TinyLlama) displays a stable performance in multimodal generation. While the massive baseline maintains an advantage in absolute humor ratings, our pipeline achieves comparable results in structural Fit (4.18 for B1 and 4.98 for B2) and visual Faithfulness. These observations indicate that summarizing visual data into text effectively reduces the narrative disconnect often found in frame-by-frame processing.

A.3 Retrieval Engine Benchmarking

Table 5 compares different embedding models across the three languages using NDCG and other standard metrics.

The data indicates a performance difference between multilingual and monolingual models. While BGE-M3 provides a stable retrieval

baseline for English and Spanish, it appears less sensitive to Chinese phonological details. The monolingual BGE-ZH model addresses this, resulting in the highest recorded NDCG (0.485).

A.4 Detailed Prompt Engineering Architecture

To ensure reproducibility, Table 6 details the prompt architecture. The system uses a constant Base Template, with language-specific directives and cultural metrics integrated according to the target language.

A.5 Extended Data Preprocessing Details

Table 7 summarizes the dataset sources and the filtering rates during the preparation phase.

A.6 Qualitative Analysis: Visual Narrative and Error Examples

Figures 3 and 4 illustrate examples from the multimodal pipeline, highlighting both its successes and cases where noisy visual signals may affect the generation.

Table 3: Internal assessment comparing the Gemini 2.5 Flash baseline with our local RAG pipeline.

Lang	Setup	Humor	Irony	Faith.	Struct.	Overall
EN	Baseline (Gemini 2.5 Flash)	4.13	4.86	4.88	4.29	4.54
	Ours (Local RAG)	4.00	4.82	5.00	4.80	4.73
ES	Baseline (Gemini 2.5 Flash)	4.39	4.95	4.79	4.58	4.68
	Ours (Local RAG)	4.02	4.98	4.93	4.17	4.31
ZH	Baseline (Gemini 2.5 Flash)	2.56	2.22	4.91	4.00	4.19
	Ours (Local RAG)	3.56	4.08	4.33	4.50	3.96

Table 4: Performance metrics for Tasks B1 and B2. The cascaded pipeline shows comparable results to the baseline in structural fit and visual faithfulness.

Task	Setup	Fit	Humor	Faith.	Score
B1	Baseline	4.00	4.67	4.90	4.59
	Ours	4.18	3.97	4.92	4.20
B2	Baseline	4.97	3.99	4.98	4.73
	Ours	4.98	3.79	4.94	4.37

Table 5: Comparison of embedding models (Top-K=10) for cross-lingual retrieval.

Lang	Model	R@10	MRR	NDCG
EN	MiniLM-L6	0.385	0.192	0.245
	mxbai-large	0.512	0.284	0.351
	gte-large	0.548	0.315	0.382
	bge-m3	0.621	0.387	0.458
ES	gemma-300m	0.475	0.245	0.315
	jina-v2	0.520	0.290	0.360
	gte-multi	0.560	0.330	0.405
	bge-m3	0.615	0.375	0.445
ZH	jina-v2	0.490	0.260	0.320
	bce-base	0.550	0.310	0.385
	bge-m3	0.590	0.355	0.430
	bge-zh	0.645	0.412	0.485

Table 6: Prompting framework showing the Base Template and language-specific components.

Constant Base Template (Generator & Critic)		
<p>GENERATOR: You are a humor expert. Generate high-quality, safe, and context-faithful humor grounded in the retrieved context. Constraints: Output strictly in {Target_Language}. Use Idea-First Generation. Connect concepts using RAG-Based. Adapt humor to {Target_Culture} by {Cultural_Injection}.</p> <p>CRITIC: Evaluate the joke based on 4 metrics (1 to 5): (1) Humor, (2) Irony, (3) Relevance, and (4) Structure. Output strictly valid JSON.</p>		
Language	Cultural_Injection (Generator)	Critic Focus
English	Focus on situational irony and concise setup-punchline delivery.	Clarity of punchline and effective use of satire.
Chinese	Leverage homophonic wordplay (Xieyin / 谐音梗). Mimic Xiangsheng tone.	机智与谐音: Homophones / lexical ambiguity and cultural idioms.
Spanish	Utilize Hispanic irony. Avoid rigid or literal translations.	Gracia and Chispa: Liveliness and Doble Sentido (double meaning).

Table 7: Dataset provenance, filtering results, and language-specific techniques.

Lang	Positive Source	Negative Source	Yield	Preprocessing
EN	SemEval-2017, Short-Jokes	Real Headlines	45K \rightarrow 15K	Contextual Filtering to separate puns from toxic text.
ES	HAHA Corpus	Political News	38K \rightarrow 24K	Deduplication via cosine similarity (0.95).
ZH	Chinese-Humor	Toutiao TNEWS	/ 32K \rightarrow 16K	Jieba segmentation with stop-word retention for timing.



(a) Case 1: Situational Irony

Stage I: "A man with his mouth open, his tongue hanging out."

Stage II: "Trying to remember a song while your tongue has other plans."

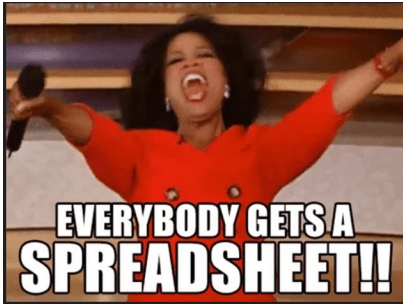


(b) Case 2: Daily Life Context

Stage I: "A man in a suit drinking from a water bottle."

Stage II: "Just a functioning adult trying to stay hydrated at work."

Figure 3: Sample qualitative results for Task B1.



(a) Case 3: Incongruity

Stage I: "A person typing on a computer with two men sitting at a table."

Stage II: "When your spreadsheet recipe results in a system error for dinner."



(b) Case 4: Dramatic Effect

Stage I: "A tennis ball moving in slow motion during a match."

Stage II: "Slow motion tennis ball: the visual definition of a dramatic moment."

Figure 4: Sample qualitative results for Task B2.

A.7 Cross-Lingual Output Examples

Table 8 presents selected outputs generated by the Intent-Aware RAG system for Task A. These examples show how the system incorporates retrieved context to address the semantic constraints. While the English tasks show closer alignment with target wit, maintaining this quality across all languages remains a complex objective.

A.8 Error Analysis and Failure Cases

To understand the performance gap in human evaluation for non-Latin languages, we conducted a qualitative error analysis. Table 9 provides representative failure cases. These examples suggest that while the model follows grammatical rules, it may miss deep cultural nuances or phonological details, highlighting the challenges for sub-10B parameter models in cross-lingual humor.

Table 8: Sample outputs from Task A. Non-English examples include translations (EN) and Pinyin (PY).

Task	Input / Context	Generated Output
Task A (EN)	Instagram closing teen accounts ahead of social media ban.	Why are Australian teens getting banned from Instagram? Because the ban is what's banned on Instagram!
Task A (ES)	The ocean as the last climate defense.	La COP30 de Belém es como la última pizza: si no nos cuidamos, la comida se vuelve debate.
Task A (ZH)	Press conference on fiscal allocation law.	某缩写直播演讲? 拜登被河蟹了! 说什么'更多的自由会影响社会稳定', 结果院版财划法修正草案来了!

Table 9: Representative failure cases in cross-lingual humor generation.

Failure Type	Input Constraint / Context	Generated Output (Failed)
Chinese (Phonetic Failure)	Keywords: 时尚 (Fashion), 高级 (High-end) <i>Context:</i> Instruction to use a traditional homophonic pun (<i>Xieyin</i>).	为什么时尚圈的人都会很高级? 因为他们的生活方式是一把『时尚』! (Why are people in the fashion industry always so high-end? Because their lifestyle is a 'shǎo' - a play on the phrase '时尚' and the word 'shǎo', meaning 'simple') Analysis: The model attempts a pun but incorrectly assumes that “时尚” (<i>shíshàng</i>) and “少” (<i>shǎo</i>) are homophones. This lack of phonological grounding prevents the joke from functioning correctly for a native speaker.
Spanish (Literalism)	Headline: Government announces new measures to control inflation.	<i>El gobierno controla la inflación como un hombre controla un globo.</i> Analysis: The model uses a simile that reads like a literal translation of an English joke. It fails to capture the cultural rhythm or 'Chispa' necessary for Spanish satire, resulting in a low humor score.