A Test-time Actor-Critic Approach to News Images Generation

Damianos Galanopoulos[1], Vasileios Mezaris[1]

[1]Information Technologies Institute (ITI), Centre of Research and Technology Hellas (CERTH), Thessaloniki, Greece

Abstract

This paper introduces the CERTH-ITI solution for the NewsImages 2026 challenge, which focuses on generating images related to news headlines. Inspired by the actor-critic paradigm in reinforcement learning, we present a test-time, model-agnostic Actor-Critic Image Generation approach (ACIG). ACIG generates prompts for image creation, produces the images, evaluates the generated results, and if needed refines the image generation prompts accordingly in a feedback loop.

1. Introduction

The MediaEval NewsImages 2026 challenge [1, 2, 3] focuses on generating and/or retrieving images that appropriately illustrate a news item, based only on the news item's headline. Contrary to our previous efforts [4, 5], which were in the direction of image retrieval, this year we tackle the image generation task. For NewsImages 2026, we, the CERTH-ITI team[1] propose Actor-Critic Image Generation (ACIG), a training-free pipeline for generating images from article headlines. Inspired by the actor-critic paradigm in reinforcement learning, ACIG iteratively produces, evaluates, and refines image generation prompts through a closed feedback loop, requiring no gradient updates or fine-tuning of any component model.

2. Related Work

Our ACIG approach leverages test-time optimization to address the challenges of the MediaEval 2026 NewsImages task, particularly the alignment between visual outputs and news headlines. The paradigm of agentic methods, in which models autonomously iterate and complete tasks through self-correction, has gained significant momentum. Recent research [6] [7] demonstrates that Vision-Language Models (VLMs) can serve as effective "critics" to refine text-to-image alignment in real time by analyzing visual-textual correspondence. In contrast to previous approaches that rely on heavy reinforcement learning (RL) training cycles, the ACIG pipeline adopts a test-time optimization strategy. Moreover, effective prompt engineering is increasingly viewed as a creative, iterative process rather than a single instruction [**?**]. This iterative philosophy is further supported by the PACE framework [8], which demonstrates the effectiveness of Actor-Critic editing in refining prompts for large language models and ensuring that generated outputs remain faithful to intended semantic constraints. This modular, agent-led workflow represents a significant advancement in bridging the gap between abstract headline narratives and the concrete visual representations required in modern newsrooms [7].

MediaEval'26: Multimedia Evaluation Workshop, June 15-16, co-located with ACM ICMR 2026, Amsterdam, The Netherlands

dgalanop@iti.gr (D. Galanopoulos); bmezaris@iti.gr (V. Mezaris)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]The code is available at https://github.com/IDT-ITI/actor-critic-image-generation.

3. Approach

3.1. Pipeline Overview

Given a set of news article headlines $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$, ACIG processes each article independently over up to T iterations. At each iteration t, the pipeline goes through three sequential stages: prompt generation (Actor), image synthesis (Image Generator), and quality assessment (Critic). In the latter stage, if an image's critic score meets or exceeds a quality threshold Q, the generation process for the corresponding article is marked as complete, and no further iterations are performed. If T iterations are completed without meeting quality threshold Q, the final output is the generated image that received the highest critic score across all iterations.

3.2. Actor: Prompt Generation

The Actor is a large vision-language model (VLM), i.e., Qwen3-VL-8B-Instruct [9, 10], responsible for transforming a headline into a descriptive prompt suitable for image-generation models. At iteration $t = 0$, the Actor receives the article headline a_i as context and is prompted to produce a single image generation prompt $p_i^{(0)}$, with an explicit instruction to emphasise a non-photorealistic visual style. At subsequent iterations $t > 0$, the Actor is additionally provided with the full scoring history from all prior attempts:

$$\mathcal{H}_i^{(t)} = \left\{ \left(p_i^{(j)},\ \left\{ s_i^{(j,k)} \right\}_{k=1}^{K} \right) \right\}_{j=0}^{t-1} \tag{1}$$

where $s_i^{(j,k)}$ denotes the critic score assigned to the image produced by the k-th generation model at attempt j, and K is the total number of image generators active in the run. This history is formatted as a structured string and passed to the Actor, which is instructed to generate an improved prompt with awareness of previous failures.

3.3. Image Generator: Candidate Images Synthesis

Each Actor-generated prompt is forwarded to one or more diffusion-based image generation models. ACIG is designed to be model-agnostic at this stage; in our experiments, we evaluate three generators: Z-Image-Turbo [11] [12], Qwen-Image (SDNQ uint4 quantised), and Qwen-Image-2512 [13]. For brevity, in Table 1 we refer to these models as ZT, SDNQ, and 2512, respectively. All generators produce images in a 16:9 aspect ratio, and a shared base seed is used across runs. As all of these generators support negative prompting, a fixed negative prompt is also supplied to suppress common artifacts such as low resolution, anatomical distortions, and AI-characteristic over-smoothness.

3.4. Critic: Image Quality Assessment

The Critic is a separate VLM, Qwen3.5-9B [14], tasked with evaluating the relevance of each generated image with respect to its source headline. For each image in the current iteration t, the Critic is queried to rate the generated images on a Likert [1-5] scale on how accurately the image captures the article's headline. The model is constrained to return a single integer. The resulting scores $\{s_i^{(t,k)}\}$ - one per image per article - are stored in the corresponding prompt entry, and the maximum score across all images of a given attempt,

$$s_i^{(t)} = \max_k \, s_i^{(t,k)}, \tag{2}$$

is used to determine if the iterative process should be concluded ($s_i^{(t)} \geq Q$) or continued.

3.5. Duplicates

In order to fully comply with the task rules, after the final image per article headline is generated, we compute the MD5 hashes of all image and identify duplicate images for the same article across runs. For these ducplicates, starting from run #10, we re-run the procedure for the affected articles by altering the seed on the generation models. This process continues sequentially through the remaining runs down to run #2, ensuring that all submitted images are globally unique across the full set of articles.

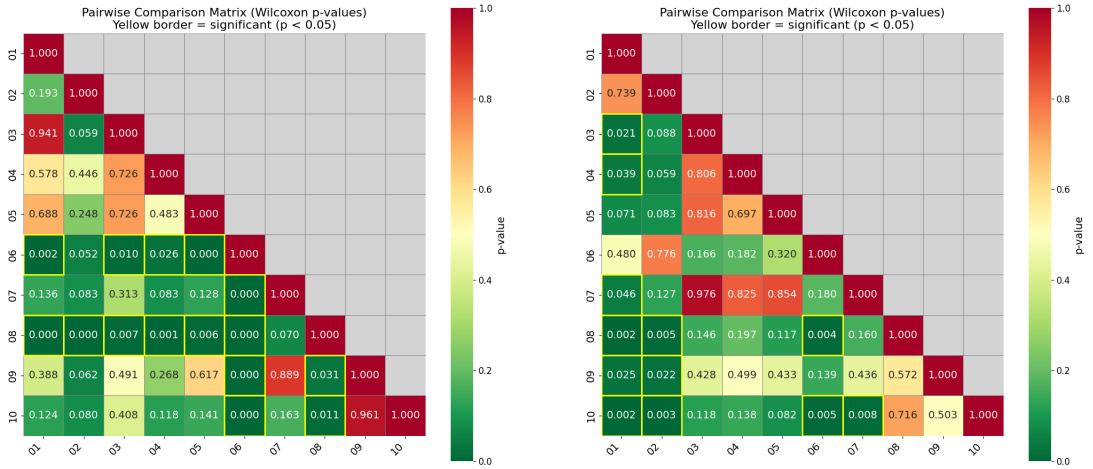
4. Submitted Runs and Results

Table 1
Configuration of submitted runs; Results (Average ratings) from the official collaborative online evaluation event and our Internal experiments. The best / second-best results are in bold / underline, respectively; higher rating values are better; and, per-run computational efficiency statistics are also reported. Column $\mathcal{H}^{(t)}$ indicates whether the critic scoring history is passed back to the Actor for prompt refinement; in the contrastive runs where this is disabled ("—"), the Actor is asked to re-generate a prompt from the headline alone at each iteration. $T = 0$ denotes one-shot generation, without any actor-critic iterations.

Run	Image Generator(s)	Qual. Thres. Q	Iter. T	Hist. $\mathcal{H}^{(t)}$	AVG rating ↑		# of iter. ↓ /article	Time (sec.)↓ /iter.	Total time (sec.) ↓ /article
					Int.	Offic.			
#1	ZT/2512/SDNQ	4	5	✓	**4.86**	3.429	1.2	91.8	111.9
#2	ZT/2512/SDNQ	5	5	✓	<u>4.84</u>	<u>3.517</u>	3.6	91.8	330.4
#3	ZT/2512/SDNQ	5	5	—	4.74	3.362	3.6	91.8	327.6
#4	ZT/2512/SDNQ	—	0	—	4.68	3.431	1	91.8	91.8
#5	ZT	4	5	✓	4.72	3.379	1.6	1.2	2.0
#6	ZT	5	5	✓	4.80	**3.616**	4.6	1.2	5.7
#7	ZT	5	5	—	4.70	3.300	4.5	1.2	5.6
#8	SDNQ	—	0	—	4.53	3.197	1	55.3	55.3
#9	2512	—	0	—	4.60	3.355	1	35.3	35.3
#10	ZT	—	0	—	4.51	3.350	1	1.2	1.2

In Table 1 we present the configuration of the ten submitted runs and whether prompt history and scoring were used during generation. We also report the results from both the official NewsImages 2026 evaluation, and an internal evaluation that we ran. For the latter, we utilized the NewsImages 2025 SMALL test dataset, and we assessed in-house each generated image similarly to the official evaluation protocol, on a Likert [1-5] scale. Moreover, to assess computational efficiency, we report three per-run statistics: the average number of iterations per input article, the average time per iteration (in seconds), and the total generation time per input article (in seconds).

The results show that enabling critic scoring history $\mathcal{H}^{(t)}$ consistently improves performance across both evaluation settings. The two best-performing runs overall (#2 and #6) both use



(a) Official results

(b) Internal results

Figure 1: Pairwise Wilcoxon signed-rank test p-values, for all pairwise comparisons between the 10 evaluated runs. Yellow borders in a cell indicate a pair of runs whose performance difference is statistically significant ($p < 0.05$).

$Q = 5$ with history enabled. Disabling critic scoring history yields a noticeable performance drop in both internal and official ratings. The non-ACIG single-shot baselines (#4, #8, #9, #10) are generally outperformed by the iterative runs with active feedback, though #4 remains relatively competitive on the official metric. Overall, using a single well-performing image generation model (ZT) within the proposed ACIG approach (run #6) yields the best results, while also achieving very competitive computational efficiency.

In Figure 1, we present the pairwise Wilcoxon signed-rank test p-values [15], comparing human evaluation scores across runs. For each pair of runs (x, y), let $d_i^{(x,y)} = r_i^{(x)} - r_i^{(y)}$ denote the difference in ratings for article i. The test statistic is $W^{(x,y)} = \sum_{i=1}^{N} (\text{sgn}(d_i^{(x,y)}) \cdot R_i^{(x,y)})$, where $R_i^{(x,y)}$ is the rank of $|d_i^{(x,y)}|$ among all non-zero differences (in ascending order), and $\text{sgn}(\cdot)$ is the sign function. The p-value is derived from the distribution of $W^{(x,y)}$ under the null hypothesis of no difference; $p^{(x,y)} < 0.05$ indicates a statistically significant difference. Although most p-values in Fig. 1 exceed 0.05, due to the small number of evaluated images, our core finding that ACIG runs #6 and #2 outperform the non-ACIG single-shot baselines (runs #8, #9, #10) is supported by, in most cases, statistically significant differences.

5. Conclusion

In this paper we presented a test-time, model-agnostic Actor-Critic Image Generation approach. We showed the effectiveness of iterative refinement over single-pass generation, demonstrating that the proposed ACIG actor-critic approach yields images that match considerably better the provided textual prompt (article headline). Within ACIG, a single lightweight image generator (ZT) can exceed the performance of much more computationally expensive alternatives.

Acknowledgements

This work was supported by the EU's Horizon Europe programme under grant agreement 101214398 ELLIOT.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to Perform Grammar and spelling checks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Heitz, B. N. Sotic, A. A. Katamjani, Q. Bi, B. Bakker, L. Rossetto, J. Kamps, NewsImages in MediaEval 2026 - Automated Image Recommendations with Retrieval and Generation Techniques for News Articles Thumbnails, in: Working Notes Proceedings of the MediaEval 2026 Workshop, 2026.

[2] L. Heitz, L. Rossetto, B. Kille, A. Lommatzsch, M. Elahi, D.-T. Dang-Nguyen, NewsImages in MediaEval 2025 – Comparing Image Retrieval and Generation for News Articles, in: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.

[3] L. Heitz, L. Rossetto, A. Bernstein, An Empirical Exploration of Perceived Similarity between News Article Texts and Images, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[4] D. Galanopoulos, A. Goulas, V. Mezaris, Cross-modal Image Recommendation for News Articles by Multimodal Foundation Models-based Retrieval-Reranking, in: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.

[5] A. Leventakis, D. Galanopoulos, V. Mezaris, Cross-modal Networks, Fine-Tuning, Data Augmentation and Dual Softmax Operation for MediaEval NewsImages 2023., in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[6] H. Jang, J. Jeon, J.-W. Hwang, K. Lee, Improving Calibration in Test-Time Prompt Tuning for Vision-Language Models via Data-Free Flatness-Aware Prompt Pretraining, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026, pp. 24300–24309.

[7] W. Ye, Z. Liu, Y. Gui, T. Yuan, Y. Su, B. Fang, C. Zhao, Q. Liu, L. Wang, GenPilot: A Multi-Agent System for Test-Time Prompt Optimization in Image Generation, in: EMNLP (Findings), 2025, pp. 929–958. URL: https://aclanthology.org/2025.findings-emnlp.49/.

[8] Y. Dong, K. Luo, X. Jiang, Z. Jin, G. Li, Pace: Improving prompt with actor-critic editing for large language model, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 7304–7323.

[9] Qwen Team, Qwen3 Technical Report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[10] Qwen Team, Qwen2.5-VL Technical Report, arXiv preprint arXiv:2502.13923 (2025).

[11] D. Jiang, D. Liu, Z. Wang, Q. Wu, X. Jin, D. Liu, Z. Li, M. Wang, P. Gao, H. Yang, Distribution Matching Distillation Meets Reinforcement Learning, arXiv preprint arXiv:2511.13649 (2025).

[12] Z-Image Team, Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer, arXiv preprint arXiv:2511.22699 (2025).

[13] Qwen Team, Qwen-Image Technical Report, 2025. URL: https://arxiv.org/abs/2508.02324. arXiv:2508.02324.

[14] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.

[15] F. Wilcoxon, Individual Comparisons by Ranking Methods, Biometrics Bulletin 1 (1945) 80–83. doi:10.2307/3001968.