

Task-Adaptive AI Agent System for Improving Ultrasonic Imaging Algorithms

Jaejun Lee ^{a,b}, Dayeon Choi ^{a,c}, Hogeon Seo ^{a,b*}

^aKorea Atomic Energy Research Institute, 111, Daedeok-daero 989 beon-gil, Yuseong-gu, Daejeon, 34057, Korea

^bUniversity of Science & Technology, 217, Gajeong-ro, Yuseong-gu, Daejeon, 34113, Korea

^cChungbuk National University, 1, Chungdae-ro, Seowon-gu, Cheongju, 28644, Korea

*Corresponding author: hogeony@hogeony.com

***Keywords :** ultrasonic testing, SAFT, agentic AI, task adaptation

1. Introduction

Imaging algorithms for ultrasonic nondestructive evaluation, such as the synthetic aperture focusing technique (SAFT) [1], are usually refined through repeated cycles in which an expert inspects reconstructed images, hypothesizes a limitation of the current algorithm, modifies the code, and evaluates the result again. This process requires domain expertise and considerable time, and its outcome depends on the experience of the individual practitioner.

Agents built on large language models can generate and revise scientific code, and recent studies have shown that coupling such an agent with an automatic evaluator in an iterative loop can improve algorithms beyond a given starting point [2,3]. This suggests a possible route to partially automating the refinement cycle in nondestructive evaluation. The evaluator plays the role of the inspection step, and the agent plays the role of the expert who revises the algorithm. The internal structure of the agent system may also affect the outcome of this refinement process.

In this study we examine this on a concrete ultrasonic imaging task. Three systems with different processing structures are placed inside an identical evaluation-feedback loop, and each is asked to improve the same baseline SAFT implementation. Their improvement behavior is compared with respect to the attained score, the progression over iterations, and the computational cost.

2. Methods and Results

2.1 Systems

All systems receive the same request, which contains an expert role statement, output constraints on the function signature and the reconstruction grid, and the complete baseline code, and all must return a complete improved implementation, as summarized in Table I. The defect locations and the scoring criteria are never included in the request.

The Direct system passes the request to a single agent call, the reasoning directive system prepends a directive that elicits intermediate reasoning before the answer, in the manner of zero-shot chain-of-thought prompting [4,5], and the proposed system is a task-adaptive multi-agent system, as illustrated in Fig. 1. In the proposed system a decomposer agent converts the request into a subtask graph, a judge agent produces a complexity profile of the request, a deterministic composer derives an agent configuration from the graph and the profile, and the configured agents carry out the execution. The configuration is therefore not fixed in advance but derived from the task itself.

After each attempt the generated code is executed on the evaluation data and scored by a deterministic evaluation module, and no agent takes part in scoring. The feedback is appended to the request of the next attempt. It can include the score trajectory of previous attempts, the per-case scores and code of the best and the

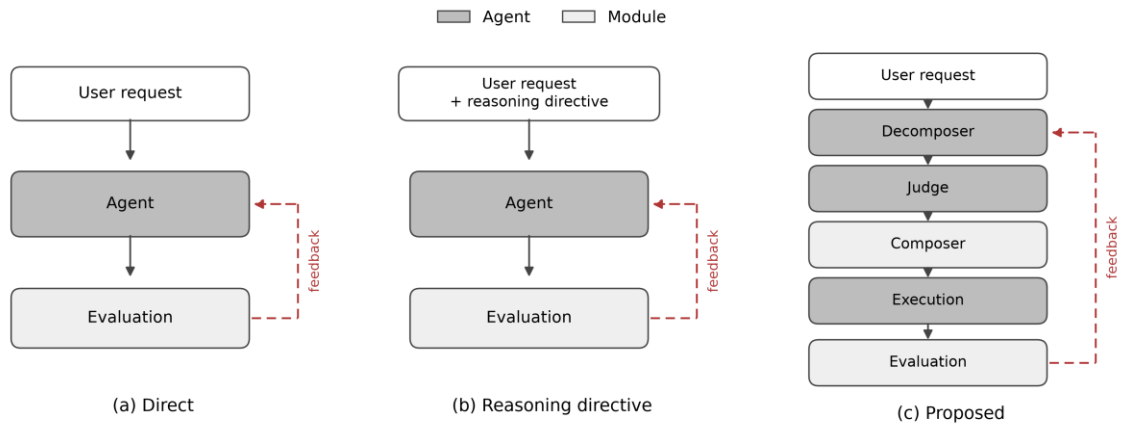


Fig. 1. Structures of the systems.

most recent attempt, and the error message of any attempt that failed to execute. The final output of a run is the highest-scoring attempt within that run.

Table I: Instruction part of the prompt for each method. The output constraints and the baseline code follow identically in all methods.

Method	Instruction
Direct / Proposed	You are an expert in ultrasonic non-destructive evaluation (NDE). Your task is to improve the reconstruction algorithm below so that it produces a more accurate reconstruction.
Reasoning directive	The directive work through the problem step by step, showing your reasoning, and then give your answer in the requested format added in front of the instruction above.

2.2 Task

The task is to improve a delay-and-sum SAFT reconstruction for pulse-echo B-scan data of an aluminum specimen. For a scan position x_i and an image point (x, z) the baseline computes the round-trip delay

$$\tau_i(x, z) = \frac{2\sqrt{(x - x_i)^2 + z^2}}{c},$$

with longitudinal wave speed c , and forms the envelope image

$$I(x, z) = \left| \mathcal{H}_t \left\{ \sum_i w_i(x, z) s_i(\tau_i(x, z)) \right\} \right|,$$

where s_i is the band-pass filtered A-scan at x_i , w_i is a Hann apodization restricted by an F-number aperture, and $\mathcal{H}_t$ denotes the Hilbert transform along depth. Early time samples are gated to suppress the surface echo.

The evaluation data are six B-scans obtained from numerical simulation of pulse-echo inspection and cover cases with different defect positions, depths, widths, and numbers, as shown in Fig. 2. A generated algorithm is applied to all six cases. For each defect the errors of the reconstructed lateral position, depth, and half-maximum width relative to the ground truth are converted to subscores

$$s(e) = \max\{0, 100(1 - |e|/\delta)\}, \quad \delta = 2\text{mm},$$

and the localization score $S_{loc,k}$ of case k is the mean of the three subscores averaged over the defects of that case. An artifact score $S_{art,k} = 100V_{in}/(V_{in} + V_{out})$ measures the fraction of visible pixels above -25 dB relative to the peak that lie inside the defect

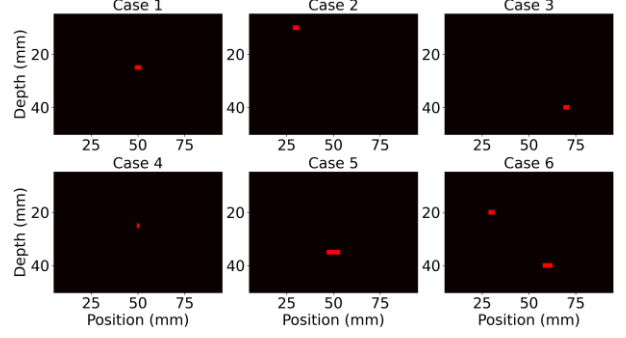


Fig. 2. Defect layouts of the six test cases.

neighborhoods, where V_{in} and V_{out} count visible pixels inside and outside those regions. The run score is

$$S = \frac{1}{6} \sum_{k=1}^6 (0.75S_{loc,k} + 0.25S_{art,k}).$$

The baseline obtains $S = 51.0$ under this metric. Its score is limited mainly by a small bias in the reconstructed depth and by an overestimation of defect width associated with the finite resolution.

2.3 Setup

All systems used Gemini 3.5 Flash-Lite as the underlying model. Each run allowed up to 10 feedback iterations and was stopped early when the best score had not improved by at least 0.1 points for three consecutive iterations. For each system 10 independent runs are reported.

2.4 Results

All three structures improved the baseline on average, as shown in Fig. 3, and we regard this as the primary observation of the study. An agent system operating under a machine-verifiable evaluation loop refined a working imaging algorithm. The improvements are not identical across structures. The reasoning directive and proposed systems reached comparable levels with small spread between runs, Direct remained lower and varied more, and the highest single run was obtained by the proposed system. Since every iteration of the proposed system involves decomposition, judgment, and execution agents, its adaptive configuration can require several times more computation than the single-call systems.

The proposed system improved fastest in the first few iterations, which is consistent with its early attempts already being competitive, and then tended to level off, as shown in Fig. 4. The single-call systems improved more gradually, and the reasoning directive system reached a comparable level in the later iterations. Direct started from the weakest first attempts and improved considerably within a run, yet its final level remained

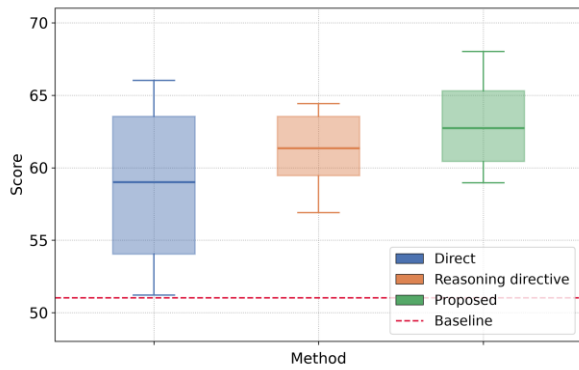


Fig. 3. Distribution of the best score of each run.

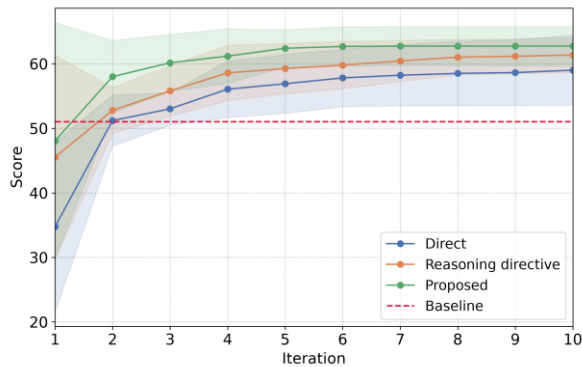


Fig. 4. Mean best score up to each iteration.

below the other two. These trajectories suggest that the processing structure mainly influences how the improvement proceeds rather than the level eventually reached on this task.

The improvement was not uniform over defect configurations. Gains concentrated on cases limited by localization accuracy, where successful modifications corrected the depth bias and suppressed surface reverberation. For the narrow defect the improvement was marginal because the measured width is dominated by the physical resolution limit. For the wide defect the reasoning directive and proposed systems fell slightly below the baseline on average although the overall mean improved. Modifications that sharpen the point response improved most cases while reducing the accuracy of the width estimate for the extended defect. This indicates that an improvement of the mean score does not guarantee an improvement in every case.

3. Conclusions

This study examined whether agent systems with different processing structures can improve a SAFT baseline under an evaluation-feedback loop. All three structures achieved average improvements, and the differences between them appeared mainly in how the improvement proceeded. The proposed task-adaptive system produced competitive early attempts and

improved quickly before leveling off, the single-call systems improved more gradually over the iterations, and all systems benefited from the feedback.

The comparison in Fig. 4 is made per iteration, and the computation per iteration differs between the systems. Each iteration of the proposed system involves several agent calls, whereas the single-call systems issue one, so the early convergence of the proposed system does not indicate higher computational efficiency, and the reasoning directive system reached a comparable level within the same iteration budget. The number of iterations equals the number of evaluation cycles.

Several limitations remain. The evaluation forms a closed loop in which the systems optimize the given metric, and the wide-defect case shows that the mean score can improve while an individual case degrades. The experiment covers six simulated cases, a single baseline, and a single underlying model, so the findings are specific to this setting. Future work includes objectives that account for individual cases, evaluation on additional and measured data, comparisons without score feedback, and a comparison under a matched computational budget.

ACKNOWLEDGMENTS

This work was supported in part by Korea Atomic Energy Research Institute R&D Program under Grant KAERI-526140-26, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25545949).

REFERENCES

- [1] J. Seydel, "Ultrasonic synthetic aperture focusing techniques in NDT", Research techniques in nondestructive testing, Vol. 6, pp. 1-42, 1982.
- [2] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhunoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, "Self-refine: Iterative refinement with self-feedback", Advances in neural information processing systems, Vol. 36, pp. 46534-46594, 2023.
- [3] B. Romera-Paredes, M. Barekatin, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. R. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, A. Fawzi, "Mathematical discoveries from program search with large language models", Nature, Vol. 625, No. 7995, pp. 468-475, 2024.
- [4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models", Advances in neural information processing systems, Vol. 35, pp. 24824-24837, 2022.
- [5] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, "Large language models are zero-shot reasoners", Advances in neural information processing systems, Vol. 35, pp. 22199-22213, 2022.