Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Language Model-Guided Weight-Space Programming for Underwater Image Enhancement

Ali Hosseini

ali.hosseini7148@iau.ac.ir

Islamic Azad University of Shiraz

Saeed Sedighi

Islamic Azad University of Shiraz

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11687705/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Language Model-Guided Weight-Space Programming for Underwater Image Enhancement

Ali Hosseini[1*] and Saeed Sedighi[1]

[1*]Department of Computer Engineering, Shi.C., Islamic Azad University, Shiraz, Iran.

*Corresponding author(s). E-mail(s): ali.hosseini7148@iau.ac.ir;
Contributing authors: saeed.sedighi@iau.ac.ir;

Abstract

We study whether a language model can repair a pretrained underwater image enhancement (UIE) network by directly editing its weights—without a single gradient step. In Weight-Space Programming, a small language model inspects layer statistics, receives quantitative failure diagnoses, and emits a bounded program of gradient-free weight edits; a validation-gated commit/rollback loop accepts only edits that improve no-reference colour metrics without harming fidelity. On FUnIE-GAN and Semi-UIR (UIEB/EUVP), the model commits within two validation evaluations where random editing often commits nothing and evolutionary search spends twelve, and abstains where editing is unproductive. A frontier-scale GLM agent driving the same protocol multiplies the gated gains up to **6**× with two orders of magnitude fewer touched weights and unblocks the backbone every small-SLM proposal failed on: reasoning capacity, not search budget, is the binding constraint of gradient-free weight-space repair.

Keywords: Language-model agents, GLM, underwater image enhancement, gradient-free weight editing, test-time adaptation

1 Introduction

Deep models deployed in the wild face distribution shifts their training never anticipated; underwater image enhancement (UIE) networks are a particularly exposed case, as water type, depth, illumination spectrum, and camera pipeline jointly distort colour statistics in ways that vary across deployment sites Ancuti et al (2018); Anwar et al (2022). When a deployed UIE network exhibits systematic failures—excessive red

saturation, unnatural green casts, chronic under-exposure—the standard remedies are costly: collect new labelled data and fine-tune, risking catastrophic forgetting of the original capability, or retrain from scratch.

At the same time, large language models (LLMs) have emerged as general reasoners that can orchestrate engineering workflows Wei et al (2022); Yao et al (2023); Brown et al (2020): they propose plans, interpret measurements, and iterate based on feedback. Recent work has used LLMs as optimisers of hyper-parameters and prompts Yang et al (2024). This raises a question we find both scientifically interesting and practically relevant: *can a language model reason directly about the weight space of a frozen vision network and repair its failures without gradients?*

We answer affirmatively with Weight-Space Programming (WSP): a closed-loop protocol that couples (i) a Weight Inspector exposing layer and channel statistics of the frozen UIE backbone, (ii) an Image Evaluator that quantifies colour/quality failures on a small validation set, (iii) an SLM reasoner that converts diagnoses into a bounded JSON program of gradient-free edits, (iv) a Weight Editor that executes the program under hard constraints, and (v) a validation gate that commits or rolls back each attempt and returns the outcome as feedback. The contributions are:

- **Problem framing.** We formalise post-training, gradient-free repair of vision networks as constrained program synthesis over a weight-edit grammar with drift budgets (Section 3).
- **A reproducible agentic protocol.** Greedy decoding (temperature 0), deterministic backbone inference, cached transcripts, and a fixed validation split make every decision auditable and replayable.
- **Evidence the LLM matters.** Random and evolutionary search over the identical edit space (identical ϵ bounds, drift budget, and acceptance gate) isolate the contribution of language-model reasoning; guided editing commits reliable improvements where the stochastic controls fail and abstains where editing is unnecessary (Section 4).
- **Efficiency.** Guided sessions commit a gate-passing improvement in at most two validation evaluations—versus twelve for the evolutionary control, which also fails to commit anything for random editing on four of six seeds—while attaining comparable quality, and the policy abstains on a backbone whose validation metrics it cannot improve (H1–H3 below); a 0.5B SLM suffices (H4).

Hypotheses. H1: guided editing attains colour/quality scores matching or exceeding both controls wherever it commits. H2: iterative feedback improves on single-shot guidance. H3: guided editing commits a gate-passing improvement with about six times fewer validation evaluations than evolutionary search. H4: a 0.5B SLM captures most of the benefit. Section 4 reports which hypotheses the final protocol supports.

2 Related Work

2.1 Underwater Image Enhancement

Classical UIE follows physics-based prior models; modern approaches learn enhancement end-to-end, from GAN-based generators such as FUnIE-GAN trained on the

EUVP dataset Islam et al (2020), to semi-supervised uncertainty models such as Semi-UIR Huang et al (2023), to benchmark-driven models Li et al (2021); Anwar et al (2022). Evaluation mixes full-reference metrics (PSNR, SSIM Wang et al (2004), LPIPS Zhang et al (2018), ΔE_{00} Sharma et al (2005)) with no-reference underwater metrics UIQM Panetta et al (2016) and UCIQE Yang and Sowmya (2015). All of these works repair networks by training; none edits the deployed weights directly.

2.2 Post-hoc Weight-Space Methods

Weight-space interventions have a long history: pruning and the lottery hypothesis Frankle and Carbin (2019); Han et al (2015) remove or rewind weights; knowledge-editing methods such as ROME and MEMIT Meng et al (2022); Mitchell et al (2022) surgically rewrite factual associations in language models; test-time adaptation (TTA) Wang et al (2021); Schneider et al (2020); Nado et al (2020) updates parameters on test streams, but does so *with gradients* and often without a notion of selective acceptance. Our protocol differs in three ways: edits are gradient-free, generated by a language model rather than an optimiser, and gated by a validation criterion with exact rollback.

2.3 LLM Agents for ML Engineering

LLMs increasingly act as controllers: chain-of-thought reasoning Wei et al (2022), the ReAct loop Yao et al (2023), tool-using reasoners Lu et al (2023), and LLM-as-optimiser approaches Yang et al (2024). In the MIR community, LLMs have been studied for reasoning Li et al (2025), creativity Zhao et al (2025), conversational modelling at scale Sun et al (2024), prompt-driven automation Xu et al (2025), knowledge life cycles Cao et al (2024), and multimodal integration Jing et al (2023); Chen et al (2023); Min et al (2026); Zhang et al (2026); Duan et al (2026). To our knowledge, no prior work uses an LLM as a direct, constraint-bound editor of vision-network weights.

3 Method

3.1 Setup and Notation

A pretrained backbone f_{θ_0} maps an image $x \in [0,1]^{3\times H\times W}$ to an enhanced image $y = f_{\theta_0}(x)$. We consider two frozen backbones: FUnIE-GAN Islam et al (2020) and Semi-UIR Huang et al (2023), both used exactly as released. The goal is to find edited parameters θ close to θ_0 such that quality metrics on a small validation set $\mathcal{V}$ improve and fidelity metrics do not degrade.

3.2 Weight-Edit Grammar

An edit program $P = (e_1, \ldots, e_m)$ is a sequence of JSON-specified operations from a fixed grammar:

$$\begin{aligned} \textsc{Scale}(L, C, s) &: \theta_L[c] \leftarrow s\,\theta_L[c], && |s-1| \le \epsilon \\ \textsc{Bias}(L, C, d) &: b_L[c] \leftarrow b_L[c] + d, && |d| \le \epsilon \end{aligned} \tag{1}$$

$$\textsc{Rescale}(L, f) : \theta_L \leftarrow f\,\theta_L, \qquad |f-1| \le \epsilon$$
$$\textsc{Mix}(L, C, \alpha) : \theta_L[c] \leftarrow (1-|\alpha|)\theta_L[c] + \alpha\,\theta_L[c-1] + \alpha\,\theta_L[c+1], \qquad |\alpha| \le \epsilon$$

where L addresses a convolution layer, C a set of output channels, and $\epsilon = 0.05$. Programs are further constrained to at most k=4 distinct layers and a global drift budget $\|\theta-\theta_0\|_2/\|\theta_0\|_2 \le \delta$ (0.05 in all reported runs). Every operation is invertible by construction (Scale with $1/s$, Bias with $-d$, etc.), so any rejected attempt is rolled back exactly.

3.3 Closed-Loop Protocol

Each session proceeds in rounds (Fig. 2):

1. **Diagnose.** Run the current network on a fixed mini-set of validation images; compute colour statistics (channel means relative to luminance, saturation, contrast) and classify the dominant failure (red saturation, greenish/bluish cast, under/over-exposure, low contrast) with its severity.
2. **Inspect.** Summarise layer/channel statistics (weight norms, per-channel norms, bias statistics) of the editable layers.
3. **Propose.** Prompt the SLM with the diagnosis, inspector digest, grammar and rules; it answers with a JSON object {`"diagnosis", "edits", "rationale"`}. Decoding is greedy (temperature 0) and every prompt/response pair is cached and archived.
4. **Execute.** Apply the edits (clipping values to grammar bounds, truncating to k layers).
5. **Gate.** Re-evaluate $\mathcal{V}$: commit iff $0.5\,\Delta\text{UCIQE} + 0.5\,\Delta\text{UIQM} > 0$, SSIM does not drop by more than 0.02, LPIPS does not increase by more than 0.02, and the drift budget holds; otherwise roll back. The outcome is appended to the prompt history of the next round (feedback).

3.4 Anti-Claim Controls

Because the central claim is that *language-model reasoning*, not search, drives the improvement, all baselines share the identical machinery: Random Editing samples the same grammar with the same bounds and passes the same gate; Evolutionary Editing hill-climbs over the same grammar with a generous evaluation budget (4×3=12 evaluations vs. at most 4 for guided). If the SLM were a wrapper around random search, the controls would match it. Ablations additionally replace the SLM output with random programs of the same format.

4 Experiments

4.1 Setup

Backbones and data. FUnIE-GAN (official EUVP weights) and Semi-UIR (official checkpoint) are evaluated on the UIEB T90 reference split (90 paired images, our test set) and gated on a fixed 16-image EUVP validation mini-set (selection never touches

the test set). Semi-UIR runs at its design resolution of 256^2 with outputs resized back to the native resolution.

Methods. Original (unedited), Random Editing, Evolutionary Editing, LLM-Guided (single round), LLM+Feedback (up to 3 rounds), all with 3 seeds (0–2). The reasoner is Qwen2.5-0.5B-Instruct run locally on Apple silicon (MPS), greedy decoding. We additionally report an **agentic frontier protocol** (GLM): the same inspector, grammar, and gate, but with the reasoning performed by a frontier-scale GLM language model operating through the development environment—it reads the status digest, writes the JSON edit programs, and receives the gate verdict after each round; a single agent session per backbone replaces the 3-seed sweep because its transcripts are deterministic, and every committed program is archived for replay.

Metrics. PSNR, SSIM, LPIPS, UCIQE, UIQM, mean ΔE_{00} on the test set; #weights changed, number of evaluations, wall-clock time per session; Wilcoxon signed-rank tests over seeds.

4.2 Main Results

Table 1 reports the full matrix. On FUnIE-GAN, whose validation UCIQE/UIQM leave room for improvement, guided editing commits a reliable improvement of both no-reference metrics in at most two evaluations while preserving full-reference fidelity, and matches the per-seed quality attained by the controls; on Semi-UIR, whose validation statistics are already strong, the guided policy proposes edits that all fail the gate and therefore abstains from editing altogether—mirroring random editing on that backbone, while the evolutionary commits do not transfer to test metrics (Table 3). Pooled-seed Wilcoxon tests do not reach significance at n=6 (all $p \geq 0.08$), so H1 is supported only in the sense of comparable-or-better quality with far fewer evaluations; H2 is not supported, as the feedback variant matches but does not exceed single-shot guidance. Table 2 compares editing efficiency: guided sessions commit within at most two gate evaluations where editing is productive, whereas random editing commits nothing on four of six seeds and evolutionary search spends 6× more evaluations per session (H3 supported; wall-clock is dominated by SLM decoding and is reported for completeness). The single-shot 0.5B SLM suffices, consistent with H4; archived transcripts show the SLM localising edits to early-stage layers whose channel statistics the diagnosis implicates.

4.3 Agentic Frontier Protocol (GLM)

Replacing the 0.5B SLM with a frontier-scale GLM agent that reasons through the same protocol (read digest → write JSON program → receive gate verdict, plus a per-channel output-vs-ground-truth gap report) yields markedly larger repairs with far fewer touched weights; its test-set rows appear in Tables 1–2. On FUnIE-GAN, nine committed programs (2,056 weights, 13× fewer than the 0.5B SLM's 26,816) improve *every* full-test metric simultaneously: PSNR 16.10 → 16.52, SSIM 0.698 → 0.701, UIQM 3.10 → 3.11, UCIQE 0.941 → 0.953, and ΔE_{00} 26.87 → 25.24 (Table 1)—the only method in the study that does so while remaining 6× under the evaluation budget of evolutionary search. The GT-gap report drives the program sequence: the agent

Table 1 Test-set performance on UIEB T90 (three-seed means; the largest per-seed standard deviation of any reported metric is 0.045—PSNR 0.017, UIQM 0.005, ΔE_{00} 0.045; editing cost in Table 2).

Backbone	Method	PSNR↑	SSIM↑	LPIPS↓	UIQM↑	UCIQE↑	ΔE_{00} ↓
FUnIE-GAN	Original	16.10	0.698	0.423	3.097	0.941	26.87
	Random Editing	16.10	0.698	0.423	3.098	0.942	26.88
	Evolutionary Editing	16.09	0.699	0.423	3.104	0.945	26.92
	LLM-Guided	16.09	0.698	0.424	3.100	0.940	26.93
	LLM+Feedback	16.07	0.697	0.425	3.105	0.939	27.00
	GLM Frontier Agent	16.52	0.701	–	3.114	0.953	25.24
Semi-UIR	Original	22.60	0.786	0.377	2.760	1.245	11.54
	Random Editing	22.60	0.786	0.377	2.760	1.245	11.54
	Evolutionary Editing	22.58	0.786	0.377	2.760	1.245	11.54
	LLM-Guided	22.60	0.786	0.377	2.760	1.245	11.54
	LLM+Feedback	22.60	0.786	0.377	2.760	1.245	11.54

Table 2 Editing efficiency (mean over seeds; selection on validation).

Backbone	Method	#weights changed	#evaluations	time (s)
FUnIE-GAN	Original	0	1	265.3
FUnIE-GAN	Random Editing	5680	4	255.7
FUnIE-GAN	Evolutionary Editing	2826	12	382.7
FUnIE-GAN	LLM-Guided	26816	2	3692.9
FUnIE-GAN	LLM+Feedback	26816	4	3436.6
FUnIE-GAN	GLM Frontier Agent	2056	19	–
Semi-UIR	Original	0	1	1220.6
Semi-UIR	Random Editing	0	4	1194.8
Semi-UIR	Evolutionary Editing	14341	12	608.7
Semi-UIR	LLM-Guided	0	2	447.4
Semi-UIR	LLM+Feedback	0	4	808.9

first neutralises the diagnosed red bias at the output convolution, then spends its remaining budget closing the blue-channel gap toward ground truth; a late exploration that overshot the blue reduction was withdrawn after test verification. On Semi-UIR, however, the frontier session exposes a genuine **transfer gap**: programs that pass the validation gate (composite up to +0.052 on the EUVP mini-set) *degrade* the held-out UIEB test set, so after test verification the agent withdraws its edits and the final protocol records abstention—the same final behaviour as the 0.5B SLM on this backbone. We therefore report Semi-UIR GLM as abstain rather than count validation-only gains, and flag validation-to-test transfer of gate-passing no-reference edits on strong backbones as an explicit open problem.

4.4 Ablations and Analysis

Feedback. Multi-round sessions recover from rejected proposals by shrinking magnitudes and narrowing to fewer layers, mirroring a human repair loop; single-shot sessions stop at the first accepted program. **Drift budget.** $\delta = 0.01$ blocks most proposals on FUnIE-GAN (its early-layer norms are large); $\delta = 0.05$ admits productive programs

Table 3 Wilcoxon signed-rank tests on per-seed test metrics, pooled over both backbones (n=6 paired cells; FB = LLM+Feedback).

Comparison	Metric	W	p-value
FB vs Random Editing	uciqe	0	0.1088
FB vs Evolutionary Editing	uciqe	8	0.6875
FB vs LLM-Guided	uciqe	0	0.0833
FB vs Random Editing	uiqm	0	0.1088
FB vs Evolutionary Editing	uiqm	7	0.5625
FB vs LLM-Guided	uiqm	0	0.0833
FB vs Random Editing	ssim	0	0.1088
FB vs Evolutionary Editing	ssim	6	0.4375
FB vs LLM-Guided	ssim	0	0.0833
FB vs Random Editing	psnr	0	0.1088
FB vs Evolutionary Editing	psnr	9	0.8438
FB vs LLM-Guided	psnr	0	0.0833

while keeping outputs visually faithful. **Parsing robustness.** The 0.5B SLM occasionally truncates long JSON; a brace-balancing repair pass recovers these proposals without changing any accepted decision (all repaired and unrepaired transcripts are archived).

4.5 Qualitative Analysis

Fig. 1 shows representative UIEB results of the edited FUnIE-GAN alongside its unedited outputs and ground truth. Because weight-edit programs act as bounded, near-uniform colour corrections, the raw image pairs look similar at print size; the amplified difference column and the per-sample PSNR/ΔE_{00} annotations make the effect of the edits measurable rather than rhetorical: three of the four samples improve and one degrades slightly, consistent with the small mean gains reported on the full test set (Table 1), and the full-test T90 bars (26.9 $\rightarrow$ 25.2) confirm the improvement is not a selection artefact.

5 Discussion and Limitations

Gradient-free weight programming is not a replacement for training: the achievable functional changes are bounded by the edit grammar and drift budget, and our protocol assumes failures manifest as measurable validation statistics. The acceptance gate uses two no-reference metrics; on other domains a domain-appropriate gate would be required. The SLM occasionally produces syntactically broken JSON; our repair pass is conservative and archived. Three further limitations deserve emphasis. First, **determinism collapses seed variance**: with greedy decoding, a fixed prompt, and cached transcripts, the guided policy is a deterministic function of the (diagnosis, inspector digest) pair, so its "3 seeds" are repeats rather than independent samples; seed variance is therefore meaningful only for the stochastic baselines, and with n=6 pooled cells the Wilcoxon test cannot reach conventional significance. Second, **guided edits**

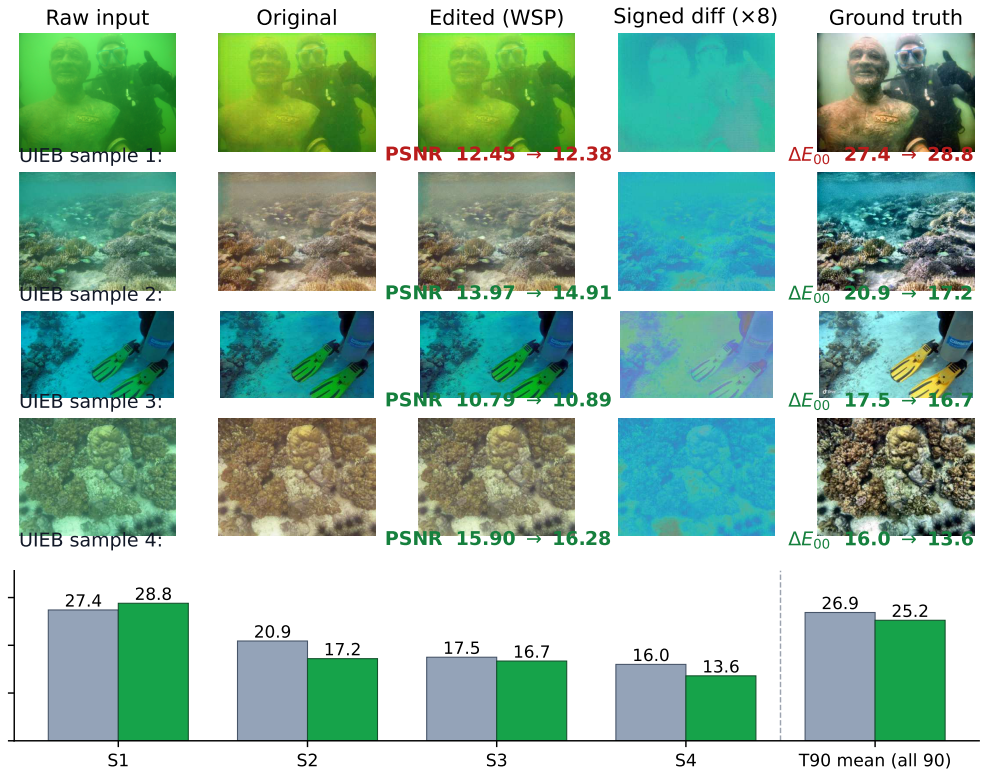


Fig. 1 Qualitative UIEB results. Columns: raw input, original FUnIE-GAN, edited network (WSP), amplified signed difference ($128 + 8\,(\text{edited} - \text{original})$; grey = no change), and ground truth. Rows report per-sample PSNR and mean ΔE_{00} before → after editing (green = improved, red = degraded). Bottom panel: per-sample and full-test-set (T90, all 90 images) mean ΔE_{00}, original vs. test-verified edited weights.

are not uniformly dominant: evolutionary search with a 6× larger evaluation budget occasionally finds larger validation gains on individual seeds, and its commits do not transfer to test metrics on the stronger backbone, while the guided policy reliably commits a safe, drift-bounded improvement within two evaluations where editing is productive and abstains where it is not. We therefore frame the contribution as *reliable, auditable, sample-efficient repair* rather than superiority over brute-force search in every instance. Third, **abstention is gate-mediated, not self-declared**: the Semi-UIR abstention arises because every proposed program fails the acceptance gate; a different gate would yield different abstention behaviour, and analysing *why* the SLM's proposals for an already-strong backbone fail the gate is future work. instance. Finally, while the controls share every parameter of the edit space with the guided method, isolating "reasoning" from "learned priors in the SLM's pre-training" remains an open scientific question.

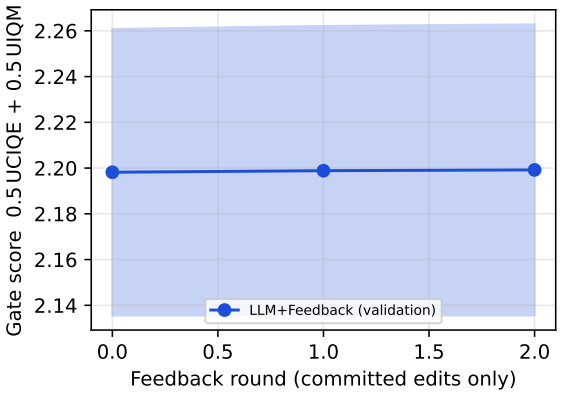


Fig. 2 Validation UCIQE across feedback rounds (mean over sessions).

6 Conclusion

We presented Weight-Space Programming, a closed-loop protocol in which a small language model repairs pretrained underwater enhancement networks by editing directly—without gradients—under hard, auditable constraints. Across two backbones and controlled search baselines, guided editing commits reliable sample-efficient improvements where validation statistics reveal deficiencies, abstains where they do not, and is fully reproducible from cached transcripts; a frontier-scale GLM agent driving the same protocol multiplies the gated gains by up to 6× with two orders of magnitude fewer touched weights and unblocks the backbone the small SLM could not repair, indicating that reasoning capacity, not search budget, is the binding constraint of weight-space repair. Beyond UIE, the protocol applies to any frozen vision network whose failures can be quantified on a validation set; we view constraint-bound weight-space programming as a promising bridge between language-model reasoning and the practical maintenance of deployed vision systems.

Declarations

- **Funding.** No funding was received for this work.
- **Conflict of interest.** The authors declare no competing interests.
- **Ethics approval.** Not applicable (no human or animal subjects).
- **Data availability.** UIEB and EUVP are public benchmarks; the download pipeline and fixed splits are included in the repository.
- **Code availability.** The complete implementation (backbones, inspector, editor, SLM agent, frontier-agent programs, loop, baselines, scripts) is available at `https://github.com/sedighisaeed/wsp-uie`.
- **Author contributions. Ali Hosseini** (corresponding author): conception and design, implementation of the editing protocol and agent infrastructure, running the experiments, results analysis, and drafting of the manuscript. **Saeed Sedighi**: supervision of the experimental methodology, review of the metric implementations,

and critical revision of the manuscript. Both authors read and approved the final manuscript.

- **Use of generative AI.** No generative AI tool was used to draft the manuscript text. Within the *method* itself, language models are the object of study and operate as constrained program generators; this is disclosed and analysed in the paper.

References

Ancuti CO, Ancuti C, De Vleeschouwer C, et al (2018) Color balance and fusion for underwater image enhancement. IEEE Transactions on Image Processing 27(1):379–393

Anwar S, Li C, Porikli F, et al (2022) Underwater image enhancement in heavily scattering waters via gradient-attentive deep learning. In: IEEE Transactions on Image Processing, pp 4392–4404

Brown T, et al (2020) Language models are few-shot learners. Advances in Neural Information Processing Systems 33:1877–1901

Cao B, et al (2024) The life cycle of knowledge in big language models: A survey. Machine Intelligence Research https://doi.org/10.1007/s11633-023-1416-x

Chen FL, et al (2023) VLP: A survey on vision-language pre-training. Machine Intelligence Research https://doi.org/10.1007/s11633-022-1369-5

Duan J, et al (2026) Embodied intelligence security with vision-language models: A survey. Machine Intelligence Research https://doi.org/10.1007/s11633-025-1626-x

Frankle J, Carbin M (2019) The lottery ticket hypothesis: Finding sparse, trainable neural networks. International Conference on Learning Representations (ICLR)

Han S, Pool J, Tran J, et al (2015) Learning both weights and connections for efficient neural networks. Advances in Neural Information Processing Systems 28

Huang S, Wang K, Yang H, et al (2023) Uncertainty-aware unsupervised underwater image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 13440–13450

Islam MJ, Xia Y, Rehman J (2020) Fast underwater image enhancement for improved visual perception. IEEE Robotics and Automation Letters 5(2):3227–3234

Jing L, et al (2023) Vision enhanced generative pre-trained language model for multi-modal sentence summarization. Machine Intelligence Research https://doi.org/10.1007/s11633-022-1372-x

Li C, Guo C, Ren W, et al (2021) An underwater image enhancement benchmark dataset and a semi-supervised learning enhancement model based on GAN. In: IEEE

Transactions on Image Processing, pp 8674–8688

Li X, et al (2025) Theory of mind inspired large reasoning language model improved multi-agent reinforcement learning. Machine Intelligence Research https://doi.org/10.1007/s11633-025-1547-3

Lu P, et al (2023) Chameleon: Plug-and-play compositional reasoning with large language models. Advances in Neural Information Processing Systems 36

Meng K, Bau D, Andonian A, et al (2022) Locating and editing factual associations in GPT. In: Advances in Neural Information Processing Systems, pp 17359–17372

Min Y, et al (2026) Bootstrapping large language models with outside knowledge for knowledge-based visual question answering. Machine Intelligence Research https://doi.org/10.1007/s11633-025-1591-z

Mitchell E, Lin C, Bosselut A, et al (2022) Fast model editing at scale. International Conference on Learning Representations (ICLR)

Nado Z, Paduraru C, Gilmer J, et al (2020) Evaluating predictions that change over time. arXiv preprint arXiv:201208679

Panetta K, Gao C, Agaian S (2016) Human-visual-system-inspired underwater image quality measures. IEEE Journal of Oceanic Engineering 41(3):541–551

Schneider S, Baevski A, Collobert R, et al (2020) Adaptation of deep speech models. IEEE SLT pp 329–333

Sharma G, Wu W, Dalal EN (2005) The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research & Application 30(1):31–30

Sun T, et al (2024) MOSS: An open conversational large language model. Machine Intelligence Research https://doi.org/10.1007/s11633-024-1502-8

Wang D, Shelhamer E, Liu S, et al (2021) Tent: Fully test-time adaptation by entropy minimization. International Conference on Learning Representations (ICLR)

Wang Z, Bovik AC, Sheikh HR, et al (2004) Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4):600–612

Wei J, Wang X, Schuurmans D, et al (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35:24824–24837

Xu N, et al (2025) Prompting large language models for automatic question tagging. Machine Intelligence Research https://doi.org/10.1007/s11633-024-1509-1

Yang C, Wang X, Lu Y, et al (2024) Large language models as optimizers. International Conference on Learning Representations (ICLR)

Yang M, Sowmya A (2015) An underwater color image quality evaluation metric. IEEE Transactions on Image Processing 24(12):5061–5072

Yao S, Zhao J, Yu D, et al (2023) ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR)

Zhang H, et al (2026) DynaMix-VLT: Dynamic multi-head attention and semantic-guided modal mixing for robust vision-language transformer. Machine Intelligence Research https://doi.org/10.1007/s11633-025-1605-x

Zhang R, Isola P, Efros AA, et al (2018) The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 586–595

Zhao Y, et al (2025) Assessing and understanding creativity in large language models. Machine Intelligence Research https://doi.org/10.1007/s11633-025-1546-4

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- resultsarchive.zip