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.

Do AI dream of P and S picks? Multimodal LLMs applied to visual seismological analysis

Raffaele Di Stefano

raffaele.distefano@ingv.it

Istituto Nazionale di Geofisica e Vulcanologia

Valentino Lauciani

Istituto Nazionale di Geofisica e Vulcanologia

Fabrizio Bernardi

Istituto Nazionale di Geofisica e Vulcanologia

Daniele Bailo

Istituto Nazionale di Geofisica e Vulcanologia

Maddalena Michele

Istituto Nazionale di Geofisica e Vulcanologia

Maria Grazia Ciaccio

Istituto Nazionale di Geofisica e Vulcanologia

Simone Marzorati

Istituto Nazionale di Geofisica e Vulcanologia

Research Article

Keywords: seismology, picking, llm, P-phase, S-phase, earthquake, detection

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10830064/v1

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

Additional Declarations: No competing interests reported.

Do AI dream of P and S picks? Multimodal LLMs applied to visual seismological analysis

Raffaele Di Stefano[1*], Valentino Lauciani[1†], Fabrizio Bernardi[1†], Daniele Bailo[1†], Maddalena Michele[1†], Maria Grazia Ciaccio[1†], Simone Marzorati[1†]

[1*]Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata 605, Rome, 00143, RM, Italy.

*Corresponding author(s). E-mail(s): raffaele.distefano@ingv.it;
Contributing authors: valentino.lauciani@ingv.it;
fabrizio.bernardi@ingv.it; daniele.bailo@ingv.it;
maddalena.michele@ingv.it; mariagrazia.ciaccio@ingv.it;
simone.marzorati@ingv.it;
†These authors contributed equally to this work.

Abstract

While conventional deep learning pipelines excel at high-throughput seismic processing, they often reduce complex perceptual tasks to rigid numerical matrices, lacking the contextual visual reasoning inherent to human experts. This study introduces an advanced proof-of-concept demonstrating the novel application of Vision-Language Models (VLMs) to visual seismological analysis, proving that multimodal generative AI can replicate expert visual heuristics to enable boosted human-like scientific interpretation. We engineered an adaptive, multi-stage foveated vision framework that dynamically rescales high-resolution, lossless waveform graphics to optimize spatial pattern recognition within the model's vision encoder. To mitigate attention saturation, we developed a streamlined, Markdown-scaffolded prompt system. Furthermore, forensic examination of VLM outputs exposed critical operational failure modes—specifically compliance bias and text-scalar decoupling—leading us to propose a conceptual two-tier framework featuring a morphological abstention protocol designed to safely reject corrupted signals. Containerized into an unattended machine-to-machine orchestrator, the automated pipeline confirms that input image fidelity and structured visual scaling are decisive for millisecond-level pattern extraction. Finally, comparative benchmarks with self-hosted open-weight architectures demonstrate the concrete feasibility of local, in-house deployments, laying the foundation

for transparent, reproducible, and institutionally sovereign AI ecosystems in operational geosciences.

Keywords: seismology, picking, llm, P-phase, S-phase, earthquake, detection

1 Introduction

In modern seismic monitoring, the exponential growth of continuous data streamsmakes exhaustive manual revision by human experts increasingly unsustainable. Consequently, reliable automated procedures capable of replicating visual event inspection are becoming essential to support routine workflows alongside traditional analyst review. Nevertheless, achieving an unbiased replication of human cognitive expertise remains challenging, as conventional deep learning approaches often reduce this complex perceptual task to rigid analytical and numerical calculations.

Identifying the arrival times (onsets) of P- and S-waves, along with their properties, does not rely on purely numerical computation, but rather on a visual, relative, and contextual process (Figure 1). In other words, the brain operates comparatively. This approach reflects a core neural mechanism known as divisive normalization [1, 2]: the biological architecture of vision does not analyze signals in isolation, but instead encodes intensity and orientation by constantly relating them to the surrounding spatiotemporal context [3].

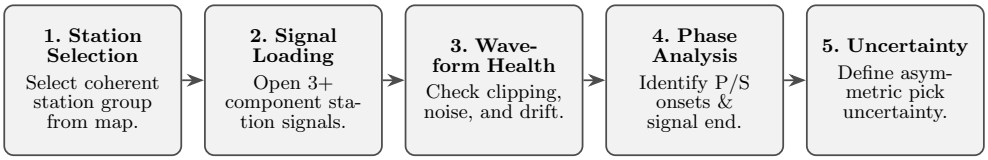


Fig. 1 Overview of the visual-cognitive phase picking workflow. The sequential pipeline progresses from spatial station selection and signal ingestion to waveform quality check, phase identification, and pick uncertainty estimation.

Visual cortical circuits are instructed by top-down modulation to suppress background noise and prioritize expected structural patterns based on prior experience [4]. Consequently, a seismologist's specific training acts as a cognitive filter that dynamically tunes this perceptual extraction. This is precisely what enables the analyst to recognize the distinctive signature of an earthquake, identifying the primary P-wave and estimating the probabilistic presence of a subsequent S-wave.

Similarly, when assessing whether the activation of a seismic sensor network aligns with wavefront propagation, the brain operates as a predictive coding system [5]. It projects an internal physical model of hypocentral dynamics onto incoming geometric data, instantaneously minimizing prediction errors to confirm the event.

This capacity for 'exact-approximate' assessment forms the essence of visual expertise, where domain-specific training restructures cortical pathways to enable flexible, context-dependent cognitive strategies [6].

This expertise relies fundamentally on the visual system, which is inherently optimized to grasp geometric patterns far faster than rows of tabular digits. While such visual pattern recognition was long a uniquely human prerogative, this paradigm is now shifting. Modern generative AI is no longer blind to the visual world, having evolved beyond text and arrays through multimodal large language models.

Thus, moving beyond conventional machine learning pipelines that process raw numerical matrices, this study introduces a novel approach for P- and S-wave phase picking designed to emulate the visual-cognitive behavior of an Expert Human Analyst. Operating as a 'Thinking AI Analyst', the framework leverages multimodal generative vision to identify wave onsets emerging from noise and evaluate their asymmetric uncertainties within a holistic visual context.

2 The visual advantage in AI

While the initial success of Large Language Models (LLMs) relied on textual reasoning, operational seismology inherently demands spatial and geometric intuition. Consequently, feeding generative AI visual representations (e.g., seismograms) provides a key analytical edge over processing raw numerical arrays.

Modern generative models leverage the *Vision Transformer* (ViT) architecture. Rather than analyzing individual pixels—a computationally prohibitive approach that degrades contextual information—the model first partitions the plot into a grid of small sub-regions, or *patches* (typically 16×16 pixels). Each patch is then flattened and linearly projected into a continuous vector space, converting sub-regions into sequential *visual tokens* [7].

Through self-attention, the model evaluates pairwise mathematical relationships across all patches simultaneously. For instance, if a P-wave initiates in patch A and its amplitude escalates in patch B, the model assigns a high attention weight to the A–B pair. This mechanism allows the AI to capture long-range spatial dependencies across the image [7, 8].

Far from a simple statistical abstraction, the AI effectively "perceives" structural trends. Much like attention maps in Vision Transformers segment objects from complex scenes [9], these self-attention mechanisms reconstruct geometric continuity and frequency variations, reliably isolating coherent seismic signals from background noise.

Yet model architecture alone is insufficient; training experience is as critical for generative AI as it is for human experts.

During the initial pre-training of multimodal Foundation Models [10], Vision-Language Models (VLMs) ingest hundreds of millions of image–text pairs from web-scale datasets to learn broad visual concepts via natural language supervision [11]. In generic pre-training workflows, this process typically incorporates literature datasets containing scientific figure–caption pairs (e.g., "The figure illustrates the arrival of the P-phase at 13.56 s") [12]. By jointly optimizing image and text encoders to maximize embedding similarity within a shared multimodal space [11], these models achieve

robust conceptual alignment. For visual seismology, the AI thus learns that domain-specific terms such as "P-phase" or "onset" map directly to the geometric transition where a flat baseline begins to oscillate.

As a result, the AI acquires a latent understanding of chart geometry [12]. When applied to seismograms, the model inherently interprets Cartesian coordinates and waveform oscillations, connecting localized visual patterns—such as abrupt shifts in amplitude or frequency—to abstract domain concepts including "phase arrival", "event", or "onset" [11].

However, this latent knowledge does not automatically guarantee accurate results: applying learned geometric rules requires visual inputs that satisfy strict resolution constraints. In biological vision, processing channels are statistically tuned to natural environments, meaning that blurring or losing high spatial frequencies disrupts information processing [13]. A similar bottleneck plagues artificial vision systems, where resolution mismatches alter activation statistics and degrade performance [14]. In seismological Vision Transformers, an analogous failure occurs when image resolution skews the mathematical proportion between physical signal frequencies and the patch grid. Under these conditions, the internal representation degrades significantly. Consequently, identifying a seismic phase onset depends not merely on model capability, but on how the visual data itself is structured.

For high-precision tasks such as millisecond-level seismic phase picking, maximizing the feature-extraction fidelity of Vision-Language Model's (VLMs) requires meticulous selection of visual input encodings. Modern Vision-Language Models process visual inputs through rigid tokenization frameworks based on the Vision Transformer (ViT) architecture [7]. Rather than interpreting an image as a continuous numerical matrix, the ViT encoder discretizes the visual canvas into a structured grid of non-overlapping spatial patches, typically 16×16 pixels in size [7]. Each patch is linearly projected into a high-dimensional vector space, converting sub-regions of the graphic into discrete visual tokens—the visual equivalent of words in natural language [7]. Self-attention layers then simultaneously map geometric and contextual relationships across all visual tokens [15]. Through this mechanism, attention heads dynamically assign weights based on spatial contrast, structural continuity, and texture alignment across the entire canvas [7].

Consequently, localized pixel degradation directly corrupts the intensity distribution within individual 16×16 patches. Artifacts such as high-frequency quantization noise, ringing, or rasterization blur introduce spatial pollution that distorts the model's underlying self-attention matrix. As a result, the Transformer computes mathematical dependencies from synthetic compression noise rather than genuine physical features, such as sharp wave onsets or Cartesian grid boundaries [16, 17].

Table 1 provides a systematic comparative synthesis of standard visual input formats (JPEG, PDF, SVG, and PNG), evaluating their underlying compression mechanisms, computational tradeoffs, and structural impacts on VLM attention matrices.

As detailed in Table 1, vector-based and document-encapsulated formats (PDF and SVG) introduce severe systemic bottlenecks. PDF ingestion relies on unstandardized server-side rendering pipelines that downsample document layouts to lower DPI

thresholds, effectively smoothing the high-frequency micro-textures of wave breakouts [18]. Conversely, although SVG paths offer infinite geometric precision, translating a standard 100 Hz multi-component time series into XML strings generates tens of thousands of character tokens. This saturates the model's textual context window, exponentially increasing API latency and inference costs while remaining susceptible to server-side vector flattening for security sanitization [19].

Table 1 Comparison of visual input formats and their impact on Vision Transformer (ViT) tokenization and self-attention.

Format	Encoding	Advantages	Limitations	Attention & VLM Impact
JPEG	Lossy DCT with quantization matrix.	Lightweight payload; fast transmission; low API bandwidth usage.	High-frequency spatial loss; irreversible ringing and blocking artifacts at sharp edges.	Trace artifacts alter pixel distributions in 16×16 patches. Self-attention (QK^{T}) computes weights on synthetic noise, biasing phase picks [7, 16].
PDF	Vector/raster primitives in document layout.	Resolution-independent layout optimized for document viewing.	Non-native parsing; requires server-side rasterization (`pdf2image`, `PyMuPDF`).	Uncontrolled server downsampling (e.g., 72–150 DPI) blurs high-frequency (100 Hz) wave onsets into sub-pixel noise, obscuring breakout features [18].
SVG	XML paths defining Cartesian vector geometries.	Infinite scalability; sharp single-pixel lines without anti-aliasing.	Large XML payloads for dense data (6,000 points/min); risks context window saturation.	Elevates token latency and API costs. Security sanitization (e.g., Polyglot SVG mitigation) flattens vectors into rasters, stripping vector advantages [19].
PNG	Lossless DEFLATE algorithm (LZ77 + Huffman coding).	Bit-perfect pixel fidelity; zero quantization noise; pristine gradient preservation.	Larger payload size; higher bandwidth and memory overhead during transmission.	Preserves exact intensity gradients across adjacent patches. Self-attention (QK^{T}) maps directly to physical signal breakout points [20].

Similarly, lossy raster compression (JPEG) introduces ringing and blocking artifacts along high-contrast boundaries [16]. When a sharp amplitude breakout is plotted against a solid background, the Discrete Cosine Transform (DCT) in JPEG discards high spatial frequencies, scattering synthetic quantization noise around the trace. Because the Vision Transformer processes the canvas in localized spatial patches, this synthetic noise pollutes the patch embeddings, forcing the self-attention mechanism to assign weights to compression artifacts rather than physical elastic wave arrivals [7, 17].

To eliminate these spatial perturbations, we adopt the Portable Network Graphics (PNG) format. PNG employs the lossless DEFLATE algorithm, guaranteeing bit-perfect geometric fidelity and preserving sharp slope variations intact prior to projection into the model's weight matrix [20]. A common critique of this choice is that commercial cloud API endpoints (e.g., Google Gemini) may re-encode incoming images into lossy formats (such as JPEG or WebP) to optimize network throughput. However, as established in digital image forensics literature [21, 22], enforcing an uncompressed,

mathematically pristine PNG input eliminates double-compression artifacts, phase-spectrum distortions, and grid noise. This guarantees that subsequent proprietary cloud-server processing operates on an uncorrupted source signal, preserving trace edge sharpness and ensuring experimental reproducibility.

3 Method: The Adaptive Foveated Workflow

To resolve VLM spatial resolution limits without incurring excessive computational overhead, we avoid passing a single, oversized image. Instead, we implement an adaptive, foveated vision loop orchestrated via Python, designated "Seismological AI ANalysis" (SAIAN) [23]. This workflow leverages the model's spatial pattern recognition while maintaining constant global image dimensions. Rather than stretching pixels—which triggers server-side downsampling and blurs high-frequency signal features [14, 24]—the Python code dynamically crops the raw NumPy array based on bounding coordinates returned by the VLM in the preceding step. By mapping a narrower time window onto the fixed canvas size, the local spatial resolution per data point increases substantially [25]. This forces the VLM's fixed 16×16 patch grid to focus entirely on the micro-structure of the waveform breakout. At each iteration, the script recomputes axis scaling and applies Automatic Gain Control (AGC), enabling millisecond-level precision through a multi-stage "glance-and-focus" process. The spatial bounds of each crop are derived directly from the preliminary pick timestamp and uncertainty interval ($t_{\text{pick}} \mp \Delta t$) returned by the model in the prior stage.

The waveforms at stage 1 are windowed based on theoretical P- and S-wave arrival times computed via ObsPy's TauP engine using a standard *ak135-f-continental* global velocity model relative to a selected hypocentral origin. As an initial operational baseline, the pipeline generates an un-zoomed, high-resolution global overview canvas consisting of the three spatial components (Z, N, E) vertically stacked (Figure 2).

To maintain station-agnostic prompts (Appendix C), the pipeline leverages the VLM's visual feature extraction by embedding essential alphanumeric metadata directly onto the graphic canvas alongside the raw seismic traces. Specifically, the image incorporates the station code, the absolute timestamp of the first sample, the active filtering status, and contextual origin parameters. Crucially, the axes are rendered to maximize chart-reading accuracy: the vertical axis displays raw amplitude counts, while the horizontal time axis incorporates minor tick marks resolved to second-level precision.

To validate the visual inspection pipeline, we deployed a specialized generative agent (see Section 4) within the Google Gemini environment (INGV Team License) set to "3.5 Reasoning", through the web user interface, as the primary operational human-in-the-loop baseline.

In this pipeline, summarized in Table 2, the human operator provisions the web user interface with both the target waveform image and the Stage 1 system prompt (see Appendix C). The VLM typically requires approximately 10 to 20 seconds of latency to generate the output, utilizing test-time computation to internally verify visual spatial relationships before serializing the results. The model yields a single,

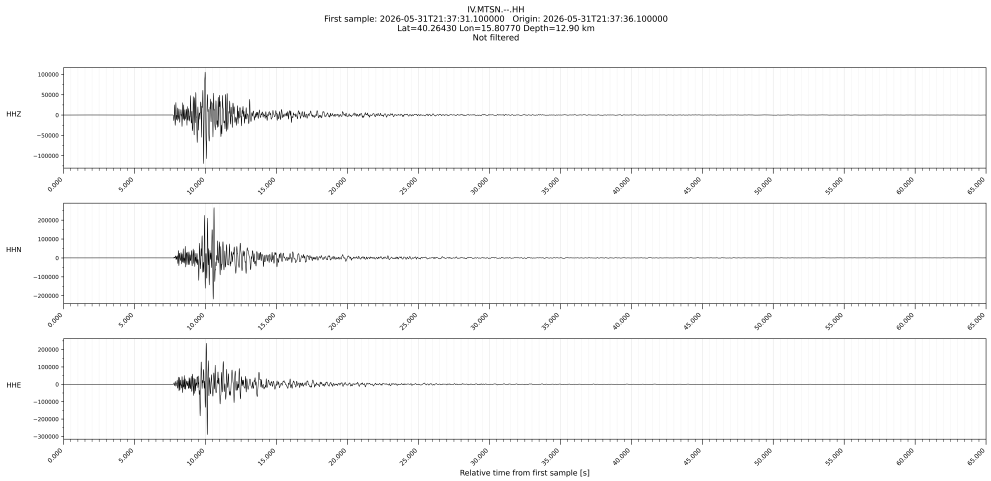


Fig. 2 Representative Stage 1 un-zoomed waveform canvas for station IV.MTSN (INGV network). The plot displays three-component velocity records (HHZ, HHN, HHE) vertically stacked and aligned along a shared temporal axis, complete with Cartesian grid ticks for relative time (seconds) and digital amplitude (counts). Header metadata explicitly reports the origin time, first sample timestamp, origin geographic coordinates with hypocentral depth, and digital filtering status; the time extent is exactly what the agent sees.

Table 2 Multi-stage interaction logic and machine-to-machine (M2M) execution loop.

Phase / Step	**Component**	**Operational Directive & Model Impact**
1. Initialization	Session Setup	Operator starts an isolated thread/web-chat with the specialized agent, enforcing morphological abstention rules.
2. Stage 1	Global Graphic Injection	Operator injects the un-zoomed three-component PNG graphic. The VLM processes the canvas and returns the preliminary structured JSON.
3. JSON Parsing	Bounding Box Generation	The operator runs the Python script (`waves2pgai.py`) that consumes the JSON *verbatim* to generate dynamic, high-resolution Stage 2 zoom plots.
4. Stage 2	Localized Zoom Injection	Operator appends refined zoom graphics, forcing the VLM's encoder to re-scan the entire visible window from scratch (anti-anchoring).
5. Refinement	Millisecond Onset Output	The VLM executes micro-detail refinement and outputs final coordinates alongside qualitative waveform morphology metadata.

highly structured JSON object (see Subsection 4 and Appendix B) containing the absolute P- and S-wave timestamps, an asymmetric uncertainty window bounding each pick, and—where the signal-to-noise ratio SNR provides clean baseline deviations—the first-motion P-wave polarity (Figure 3A).

While this initial output replicates the standard picking information typically archived by human analysts, the VLM framework offers a distinct architectural advantage: the contextual capture of transient qualitative metadata. During the inference loop, the agent appends continuous feature evaluations that are traditionally lost in



Fig. 3 Representative structured JSON output payloads returned by the VLM for station IV.MTSN. (A) Stage 1 global overview payload detailing preliminary arrival timestamps, asymmetric uncertainty bounds, suggested bandpass filtering parameters, and next-stage zoom mandates. (B) Refined Stage 2 JSON payload generated after inspecting high-resolution zoom windows, containing finalized arrival times, narrowed uncertainty bounds, P-wave first-motion polarity ("down"), and qualitative waveform morphology descriptions.

manual workflows due to time constraints. The model documents the morphological state of the waveform, the logical deduction path leading to the pick localization, and an explicit evaluation of whether background noise requires digital filtering (see tag `"comments"` in Figure 3).

Furthermore, the agent acts as an autonomous routing node by evaluating of the two downstream boolean flags: `suggested_bpfilter` and `next_stage`.

The Stage 2 refinement pipeline (Figure 4) focuses exclusively on tight, localized windows centered around the initial Stage 1 timestamps via an automated Python routine.

The final two steps mathematically simulate the visual enhancements manually performed by human experts during phase picking. The routine outputs two high-resolution PNG graphics (P-zoom and S-zoom), which the operator uploads to the reasoning agent alongside the secondary Stage 2 prompt (Section 4). The VLM then completes the loop by generating a refined Stage 2 JSON payload (Figure 3B). The entire process — encompassing computational pre-processing, visual scaling, and cross-modal inference — averages approximately 60-120 s per station within this human-in-the-loop validation prototype.

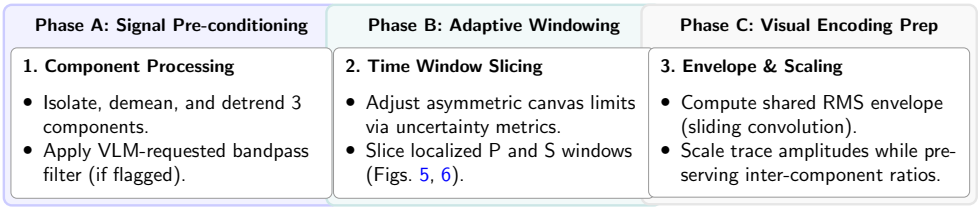


Fig. 4 Overview of the automated Stage 2 refinement pipeline, structured into three functional phases: (A) signal pre-conditioning, (B) adaptive time windowing, and (C) contrast-enhanced visual amplitude scaling.

To guarantee absolute architectural reproducibility and mitigate the risks of context window contamination, the pipeline enforces a strict isolation rule: the cross-modal evaluation of a three-component waveform is conducted within an independent, isolated chat session dedicated to a single station (encompassing all the requested analysis stages). Upon completion, the session is terminated and flushed from memory before moving to the next station. Retaining a localized context window explicitly counteracts the "Lost in the Middle" phenomenon [26] inherent to autoregressive attention mechanisms. In fact, massing multiple station images, historical prompt streams, and disparate JSON payloads within a single inference session inevitably dilutes the model's active attention weights, leading to unreplicable behavior and processing drift.

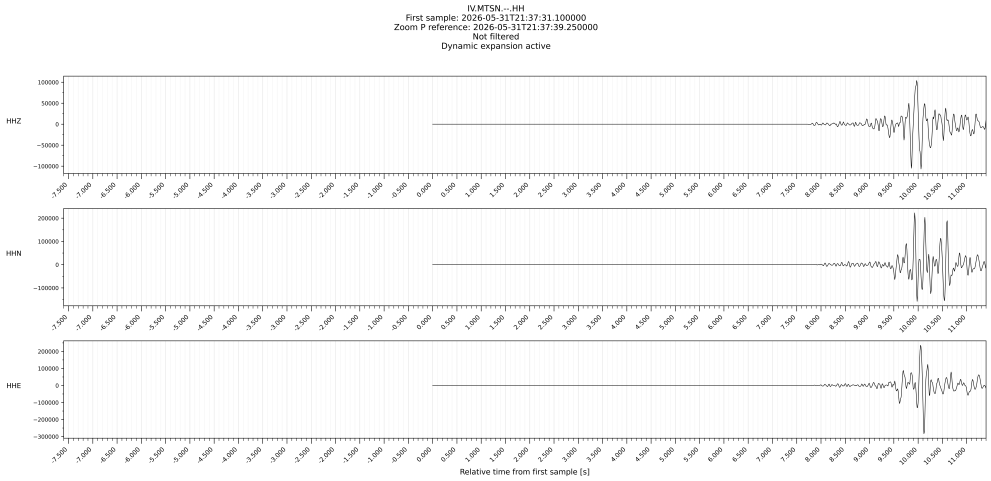


Fig. 5 Stage 2 localized P-wave zoom canvas for station IV.MTSN. The three spatial components (HHZ, HHN, HHE) are dynamically windowed around the Stage 1 candidate anchor, demeaned, detrended, and normalized using a shared RMS amplitude envelope to optimize spatial contrast for the VLM vision encoder.

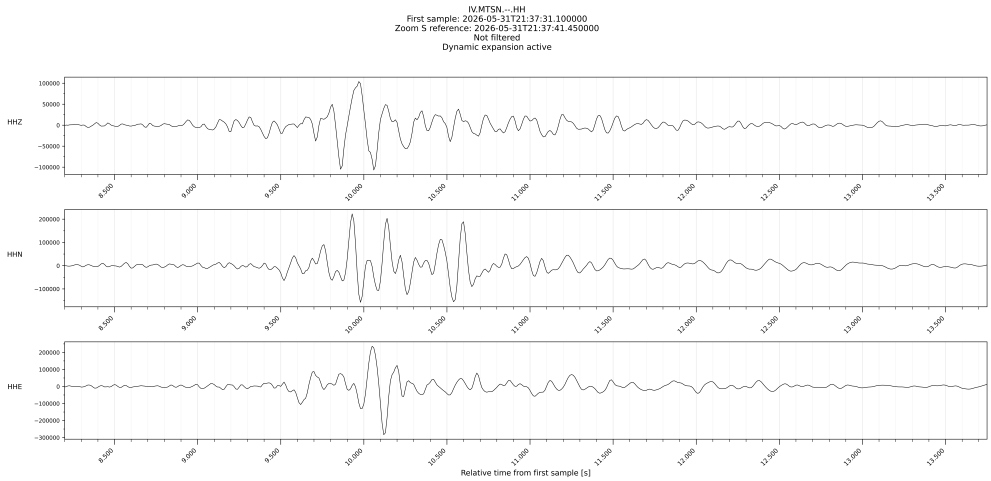


Fig. 6 Stage 2 localized S-wave zoom canvas for station IV.MTSN. The horizontal components (HHN, HHE) display coherent transverse energy arrivals, re-projected onto the full pixel grid to enhance visual onset contrast against the preceding P-wave coda.

3.1 Automated Pipeline Integration via APIs and Local Execution

The human-in-the-loop prototype is mandatory to validate the core picking protocol, but deploying vision-language models across dense regional networks demands the foveated picking protocol must be fully automated. This section details the software architecture of an unattended machine-to-machine orchestrator, designed to standardize API execution across both cloud and local deployments while ensuring production-scale reproducibility (see Figure 16 in Appendix D).

3.1.1 Overview and Multi-Stage Interaction Logic

To bridge the operational gap between interactive and automated pipelines, we developed `pic_picker_ai`, an unattended orchestrator designed to execute machine-to-machine (M2M) inference without altering the underlying seismological logic. Crucially, the system instructions (Appendix B) and operational prompts (Appendix C) are transmitted *verbatim*. This ensures that any performance variance between manual and automated execution stems strictly from backend transport dynamics rather than prompt reformulation.

To preserve absolute cross-modal integrity, the automated pipeline enforces strict session isolation: each seismic station is processed within an independent inference thread, instantiated upon station loading and flushed immediately after completion. From a machine learning perspective, this isolation prevents cross-station KV-cache contamination and mitigates the "Lost in the Middle" attention-dilution phenomenon inherent to autoregressive Vision Transformers [26].

Within a single station thread, context accumulation remains strictly bounded through the sequential five-step execution loop detailed in Table 2. Because localized zoom windows center on the model's preliminary Stage 1 outputs, the timestamp rendered on the Stage 2 canvas serves merely as a framing reference. This automated loop reproduces the anti-anchoring cognitive constraint established for human analysts [27]. Furthermore, this multi-turn architecture systematically captures the model's continuous qualitative metadata (`comments.text`), archiving waveform morphology descriptions and post-hoc justifications for every station without incurring computational overhead.

The complete software architecture—comprising the eight-module Python package, multi-stage container specifications, and fail-safe exception protocols—is detailed in Appendix D.

3.1.2 Endpoint Standardization, Guardrails, and Reproducibility Risks

The pipeline maintains endpoint independence at the API schema layer by mapping inference pathways to standardized chat-completion protocols. To stabilize autoregressive decoding over structured JSON schemas, we restrict public commercial endpoints to a sampling temperature of zero, minimizing stochastic variance [28]. However, zero-temperature sampling does not guarantee bit-identical outputs across commercial cloud routers. Cloud requests remain subject to unannounced hardware reallocations and proprietary server-side weight adjustments—a behavioral volatility formally defined as *model drift* [29].

To mitigate model drift, eliminate transactional API overhead, and uphold Open Science data sovereignty, we containerized the orchestration pipeline to support self-hosted, open-weight models executed locally via Ollama [30]. For local deployments handling multi-image visual arrays and long-form system instructions (see Section 4, the orchestrator dynamically expands the active context allocation (`num_ctx`). This enables direct performance comparisons between cloud foundation models and local open-source architectures under identical prompt and zoom configurations (see listings in Appendix D).

To guarantee production-scale robustness without human intervention, an automated guardrail layer normalizes API output asymmetries [28]. The parser automatically strips Markdown code fences, validates ISO-8601 timestamp compliance before passing coordinates to the host plotting engine, and overwrites station identity fields with verified metadata. This eliminates transcription noise caused by visual patch misreading [7]. Appendix D details the complete failure-handling protocols, retry logic, and data provenance archiving standards.

4 Prompt Engineering and Cognitive Offloading

Executing the adaptive foveated workflow demands highly specialized system instructions.

As mentioned in Section 2, by natively integrating cross-modal comprehension, modern VLMs map graphic vectors and pixel spatial distributions directly into a

unified latent space shared with their language module [31, 32]. This architectural extension distinguishes VLMs from standalone ViTs: rather than merely perceiving waveforms as abstract oscillating lines, the model correlates these visual geometries with physical phenomena—such as "seismic phases", "background noise", and "wave polarity"—and dynamically enforces the heuristic constraints defined in system instructions and in the prompt [11, 33].

Nevertheless, adapting a general-purpose VLM to execute expert-level visual tasks does not inherently require full model fine-tuning; rather, it hinges on orchestrating targeted system instructions that steer the VLM's latent cross-modal attention layers toward critical visual features [11, 34], serving as a rigorous mathematical and behavioral framework for VLMs.

In Section 3, we detailed the deployment of a specialized generative agent driving both our interactive human-in-the-loop and automated machine-to-machine pipelines. Formulating its set of instructions in our work represented an iterative empirical journey—a human-AI co-design process during which we gained firsthand insight into the fundamental operational constraints governing VLM visual inspection and attention allocation [35].

In fact, to fully leverage VLMs for precise seismic phase picking, system instructions must be carefully engineered to balance theoretical domain rigor with computational efficiency. This section outlines the human-AI co-design framework developed to govern cross-modal reasoning, detailing the contextual bottlenecks encountered during initial deployment and the cognitive offloading strategies implemented to resolve them.

4.0.1 The Attention Saturation Bottleneck and Cognitive Offloading

In multimodal transformers, appending dense reference literature alongside a high-resolution image creates an acute attention-allocation conflict. In our initial setup, we empirically experienced this bottleneck by augmenting the system instructions with full-text foundational seismology manuals [36, 37]. Rather than improving analytical accuracy, this massive context injection overloaded the attention mechanism, triggering conservative refusal guardrails where the VLM defensively claimed an inability to perform millisecond-level visual picking. An interactive meta-dialogue (transcript provided in Appendix A) revealed that this failure stemmed not from a deficit in visual processing capacity, but from a context-induced safety mechanism designed to prevent speculative outputs under high token entropy [38].

Autoregressive models do not permanently retain system instructions in their active weights; each message within the inference loop requires reprocessing the entire system prompt, chat history, and raw visual tokens from scratch. Consequently, simultaneously processing high-dimensional visual matrices and heavy textual constraints—such as extensive seismology manuals and massive JSON templates—forces the softmax operation over the combined sequence length to severely dilute the attention weights.

As documented in long-context attention literature [12], this text-heavy saturation degrades the model's focus on critical instruction boundaries, frequently triggering preventive server timeouts or processing drops.

Table 3 System prompt engineering and cognitive offloading strategies.

Phase / Step	Component	Operational Directive & Model Impact
Pre-processing	Textual Reduction	Stripped redundant geological theory and removed large theoretical reference documents to free the active context window.
Pre-processing	Spatial Directives	Mandated absolute pixel-inspection directives, segmenting instructions into actionable vision rules.
Pre-processing	Zero-Shot Serialization	Enforced strict JSON output via grammar-constrained decoding, mapping visual features into machine-readable syntax deterministically.

To mitigate this bottleneck, we implemented instead a strategy of Cognitive Offloading (Table 3). By shifting the system prompt from an academic reference manual to a streamlined, vision-first operational guide, we eliminated redundant textual overhead. Furthermore, by enforcing a strict JSON output directive, the instruction pipeline leverages grammar-constrained decoding [28] to deterministically map latent cross-modal visual features into a rigid, machine-readable syntax [39]. This transformation ultimately converts the probabilistic vision-language model into a deterministic, high-fidelity phase-picking engine.

4.0.2 Structural Scaffolding via Markdown Optimization

It is worth noting that the success of this streamlined operational guide relies also on structural scaffolding via Markdown syntax (headers #, bullet points $*$, and bold anchors $**$). Within the Transformer's self-attention matrix, unformatted text frequently induces instruction bleeding, causing distinct behavioral constraints to conflate within the token sequence.

Integrating strict Markdown formatting serves as an attention anchoring mechanism. During pre-training, these models acquire a strong structural bias from well-formatted technical repositories [34, 35].

In our pipeline, encapsulating operational constraints within bold syntax (e.g., $**$MANDATORY EXCLUSION$**$:) amplifies the dot-product attention scores ($\mathbf{QK}^T$) for those specific tokens. This mechanism ensures that, even while processing thousands of visual tokens from the seismogram, key behavioral boundaries remain dominant in the attention field, effectively mitigating syntax errors and hallucinations.

By offloading the initial textual overhead and structuring the remaining directives through Markdown scaffolding, the model allocates its full computational capacity to pixel-level waveform analysis during inference. The final, optimized system instruction set resulting from this human-AI co-design process is detailed in Appendix B.

5 Data

We applied our generative AI-driven phase picking pipelines to the seismic signals of 11 local earthquakes (Figure 7 and Table 4). The events were selected to capture diverse signal characteristics across different regions of Italy, representing varied wave propagation media from the hypocenters to the stations of the national IV network operated by the Istituto Nazionale di Geofisica e Vulcanologia (INGV). The starting waveforms dataset was restricted to a maximum epicentral distance of 200$[km]$. The

final waveforms dataset is limited anyway to a maximum of 5-10 stations showing only noise, beyond the last one showing medium-low S/N ratio.

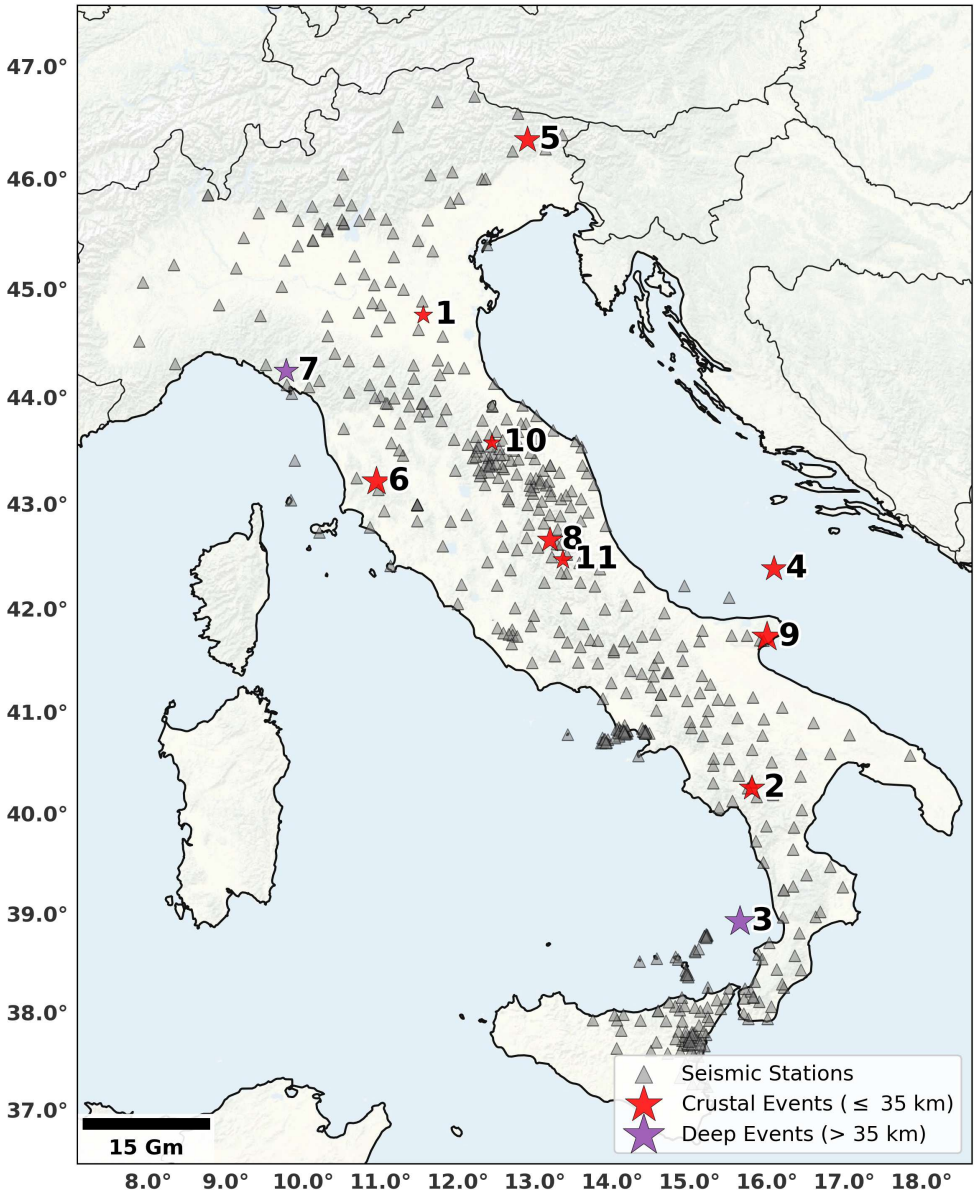


Fig. 7 Stars: 11 earthquakes composing the test dataset over which we benchmarked our AI guided procedures with respect to human analysts analysis (see Table 4); triangles: seismic stations used in this study.

Table 4 Seismic events analyzed in this study. The complete event QML parameterization is available via the INGV FDSN Web Service at the following route: https://webservices.ingv.it/fdsnws/event/1/query?eventId=[EventID].

#	Event ID	Origin Time (UTC)	Mag	Lat (°)	Lon (°)	Depth (km)
1	5098811	2015-03-15 14:26:11.08	3.1	41.741	16.003	27.3
2	44765722	2025-12-05 09:07:06.17	2.7	42.668	13.196	11.1
3	45442722	2026-03-27 01:19:40.10	2.6	44.256	9.784	64.1
4	45518622	2026-04-06 02:14:12.42	3.0	43.224	10.951	8.3
5	45591322	2026-04-16 01:14:03.84	2.8	46.359	12.905	9.8
6	45696312	2026-04-28 00:54:24.22	2.7	42.395	16.093	1.4
7	46019942	2026-05-23 16:39:08.49	3.1	38.934	15.653	165.9
8	46057512	2026-05-28 01:02:36.55	2.0	43.589	12.444	14.1
9	46096912	2026-05-31 21:37:36.10	2.7	40.264	15.808	12.9
10	46368612	2026-06-28 21:25:27.85	2.1	44.773	11.557	1.5
11	46552922	2026-07-15 02:25:04.87	2.2	42.481	13.362	14.1

To ensure high sensitivity to small-to-moderate magnitude events, this initial evaluation was restricted to velocimeter channels (seed naming HH or EH). Furthermore, to rigorously benchmark the VLM's multi-scale decision stability, we applied the Stage 2 zoom routine across all stations—even when a false next-stage flag indicated the unzoomed resolution was sufficient for phase identification. Finally, we manually picked P- and S-onsets for the 11 earthquakes setting up an expert seismologists reference dataset.

6 Results and Analysis

In the following paragraphs we present the results for both the human-in-the-loop pipeline, and the automated APIs and Ollama/Local pipelines, against the expert manual picks over a regional distance scale up to $\approx 100[km]$.

6.1 Human-in-the-loop Workflow Analysis

In Figures 8 and 9 we report examples of the final behavior of our interactive human-in-the-loop phase picking pipeline on two stations from a ML 2.7 occurred on 2026 May 31th at 21:37:36 (UTC) in the Southern Italy (eventid 46096912). The VLM correctly guided the zooming stage 2, where the true P and S onsets were identified with appropriate asymmetric uncertainties, with preliminary analysis from stage 1. It is worth underlying the comment at stage 1: "Full scale view shows clear P and S arrivals, but high-resolution zoom windows are mandatory to refine precise onset times and determine polarity." A closer analysis highlights that while the P onset for this station was slightly anticipated by the VLM (though well bounded by its uncertainty window), the S onset might have been slightly delayed on the N component while perfefctly on time on the E component. Conversely, no comments are needed for the results on station MGR.

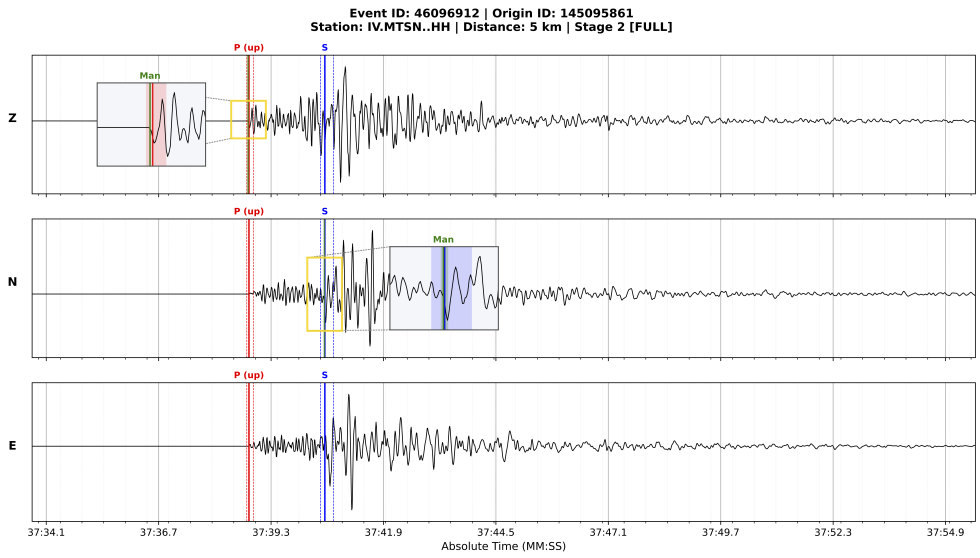


Fig. 8 Stage 2 visual picking results for station IV.MTSN (epicentral distance: 5 km) of the 2026-05-31 M_L 2.7 event (#9). Vertical red line indicates the AI-predicted P-wave arrival with compressional ("up") first-motion polarity and its associated asymmetric uncertainty window (shaded pink area); vertical blue line marks the coherent S-wave arrival. The inset boxes provide a magnified view of the P- and S-onset regions (highlighted by the yellow boxes), showing the fine alignment between the manual pick (green line, "Man") and the VLM prediction.

All the results in terms of picks on waveforms, for the complete dataset are available in the Supplementary Material. In the following sections, we discuss the complete results of the pipeline directly against the human picked manual results.

6.1.1 Statistical Performance and Multi-Stage Zooming

The dataset spans the 11 representative seismic events. The temporal residuals $\Delta t = t_{manual_pick} - t_{AI_pick}$ for P-onsets and S-onsets are evaluated across the two progressive processing tiers: un-zoomed raw overview (Stage 1), and context zoom (Stage 2). A third tier, the ultrafine zoom Pass (Stage 3) was applied on purpose (see further below in this paragraph).

In Stage 1 17.8% of P-picks and 12.6% of S-picks fall within the high-precision $\leq \pm 0.10$ s residual bin, while gross temporal deviations (> 0.50 s) reach 28.3% for P-waves and 47.7% for S-waves (Figure 10 and Table 5). This happens because the model inspects the full-scale waveform window (65 s) mapped onto a standard image canvas (Figure 2). This forces a single 16×16 pixel patch to absorb several seconds of waveform data [14]. The physical onset of a P-wave occupies an infinitesimal fraction of a single patch. The surrounding pre-event noise dilutes the local contrast gradient, flattening the Softmax activation layer and driving up token generation entropy [40].

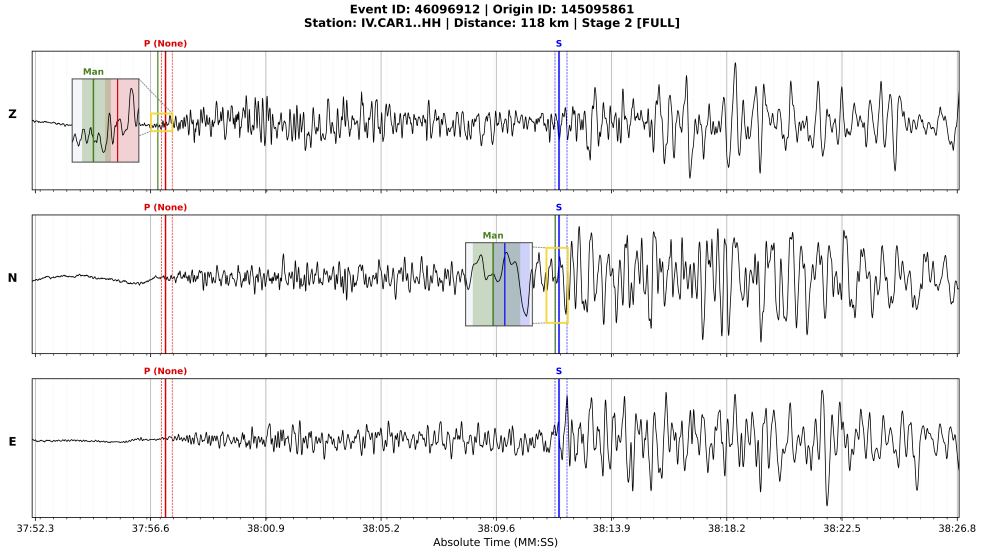


Fig. 9 Stage 2 visual picking results for station IV.CAR1 (epicentral distance: 118 km) of the 2026-05-31 M_L2.7 event (#9). Vertical red line indicates the AI-predicted P-wave arrival with undefined first-motion polarity ("P (None)") and its asymmetric uncertainty window (shaded pink area); vertical blue line marks the S-wave arrival. The inset boxes provide magnified views of the P-onset and S-onset, highlighting the alignment between the manual picks (green lines, "Man" with shaded green area for the related asymmetric uncertainty) and the VLM predictions.

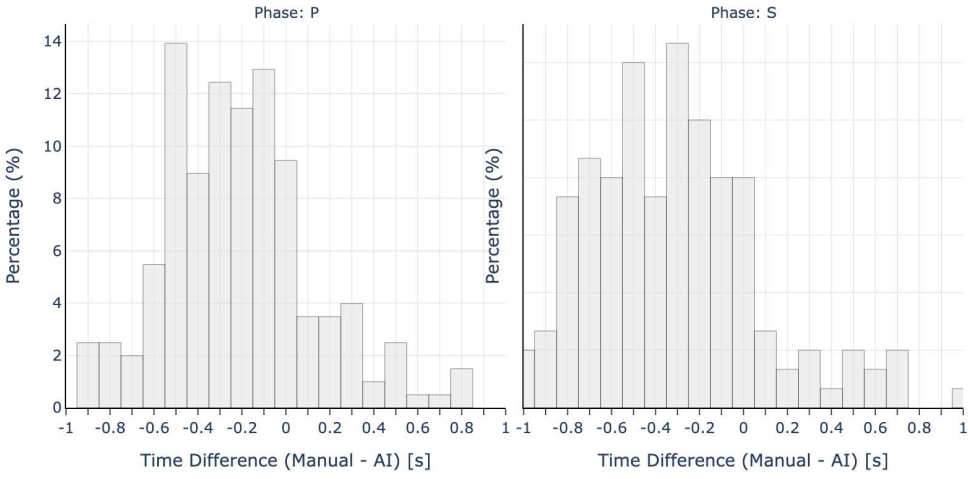


Fig. 10 Distribution of arrival time residuals ($\Delta t = t_{\mathrm{manual}} - t_{\mathrm{AI}}$) for P (left) and S (right) phases obtained at Stage 1. Full-scale canvas compression forces several seconds of data into individual spatial patches, driving residual dispersion and a high fraction of gross outliers (> 0.50 s).

In Stage 2 (Figure 11), our adaptive foveated loop systematically executes a context-level zoom (±3.0 s window) across all recording stations by dynamically cropping the raw NumPy array around the Stage 1 candidate anchor. By re-projecting this narrow time window onto the exact same image canvas dimensions, the pixel-per-second density expands substantially [25]. As a consequence, the initial wave breakout now spans multiple contiguous 16×16 pixel patches rather than being compressed within a single, highly averaged tile [7]. The Vision Transformer's self-attention layers (QK^{T}) register a sharp vertical contrast gradient between pre-onset noise patches and post-onset signal patches, driving a dramatic performance leap: high-precision P-pick accuracy ($\leq \pm 0.10$ s) more than doubles from 17.8% to 40.4%, while S-pick accuracy ($\leq \pm 0.10$ s) more than triples from 12.6% to 41.0% (Table 5). Concurrently, the cumulative acceptance rate within $\leq \pm 0.50$ s climbs to 87.6% for P-waves and 85.0% for S-waves, cutting gross temporal outliers (> 0.50 s) by more than half for P (28.3% down to 12.4%) and by over two-thirds for S (47.7% down to 15.0%).

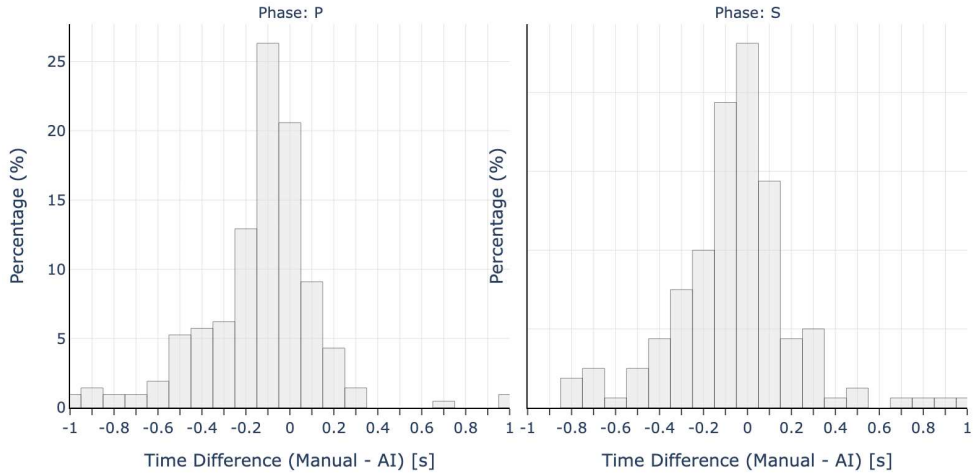


Fig. 11 Distribution of arrival time residuals ($\Delta t = t_{\mathrm{manual}} - t_{\mathrm{AI}}$) for P (left) and S (right) phases obtained at Stage 2. Adaptive time-window re-projection concentrates patch attention scores, significantly narrowing error distributions and doubling high-precision picks ($\leq \pm 0.10$ s).

To push spatial localization to its physical limits, we evaluate the pipeline performance under a composite maximal-resolution tier designated as *StageMax* (Figure 12 and Table 5) [25]. Rather than forcing a uniform third stage across all recording stations, to run this test we applied ultra-fine zoom Stage 3 only to stations with P or S residuals $\geq 0.2s$. StageMax represents the highest resolution tier attained per station: it incorporates the ultrafine Stage 3 zoom pass (±0.4 s half-width window around candidate anchors) wherever executed, while retaining the context Stage 2 output (±3.0 s half-width) for the remainder.

Under this maximal resolution tier, high-precision performance undergoes a major convergence: cumulative P-pick accuracy within $\leq \pm 0.10$ s reaches 67.6% (148 picks), with an extraordinary 42.9% (94 picks) falling within the sub-50 ms bin ($\leq \pm 0.05$ s). Similarly, S-pick accuracy within $\leq \pm 0.10$ s climbs to 52.9% (92 picks), with 31.0% (54 picks) reaching millisecond-level precision ($\leq \pm 0.05$ s) [7]. Furthermore, residual error distributions transition from a heavy late-picking bias in Stage 1 and partial recovery in Stage 2 to a nearly ideal, zero-centered Gaussian profile under StageMax. Critically, 60 P-picks previously trapped in intermediate resolution bounds during Stage 2 migrated directly into the highest accuracy bin ($\leq \pm 0.10$ s) under StageMax. Conversely, gross temporal outliers (> 0.50 s) remained essentially flat (11.9% for P, 14.9% for S), proving that multi-scale zooming acts as a selective cure for spatial resolution limits rather than an indiscriminate statistical smoother [25, 41].

Table 5 Performance convergence of high-precision phase picking between stages for WebGem. The table highlights the absolute increase in the percentage of picks falling within specific temporal residual bins.

Phase	Residual Bin	Stage1	Stage2	StageMax	Δ StageMax
P	$\leq 0.05[s]$	8.7%	21.6%	42.9%	+21.4%
P	$\leq 0.10[s]$	17.8%	40.4%	67.6%	+27.2%
P	$\leq 0.20[s]$	30.6%	63.3%	80.4%	+17.1%
P	$\leq 0.50[s]$	71.7%	87.6%	88.1%	+0.5%
P	$> 0.50[s]$	28.3%	12.4%	11.9%	-0.5%
S	$\leq 0.05[s]$	7.5%	23.1%	31.0%	+7.9%
S	$\leq 0.10[s]$	12.6%	41.0%	52.9%	+11.8%
S	$\leq 0.20[s]$	21.8%	61.3%	70.1%	+8.8%
S	$\leq 0.50[s]$	52.3%	85.0%	85.1%	+0.1%
S	$> 0.50[s]$	47.7%	15.0%	14.9%	-0.1%

To evaluate how physical wave propagation interacts with the VLM's visual perception, picking residuals and declared epistemic uncertainties were discretized across three epicentral distance classes: Near-Field (≤ 30 km), Mid-Range (30–60 km), and Far-Field (60–100 km) (Figure 13).

In Stage 1, increasing epicentral distance causes a systematic degradation in performance (Figure 13, Panels A and B). At extended distances (≥ 60 km), anelastic crustal attenuation and wave scattering act as physical low-pass filters, smoothing sharp, impulsive phase onsets into emergent waveforms [42]. On the compressed 65-s visual canvas of Stage 1, this attenuation-induced loss of contrast flattens the local pixel gradient, driving the interquartile range (IQR) of Stage 1 residuals up to 0.65 s for P-waves and 0.83 s for S-waves in the Far-Field.

The execution of multi-scale foveated zoom (Stage 2 and Stage 3) systematically counteracts this physical signal degradation [25]. By re-projecting narrow time windows onto the full pixel grid, the foveated loop expands local canvas resolution, artificially restoring the visual contrast between the pre-onset baseline and the phase arrival [43]. Consequently, Stage 3 compresses median absolute residuals below 0.05 s across all distance classes for P-waves and below 0.10 s for S-waves, demonstrating that

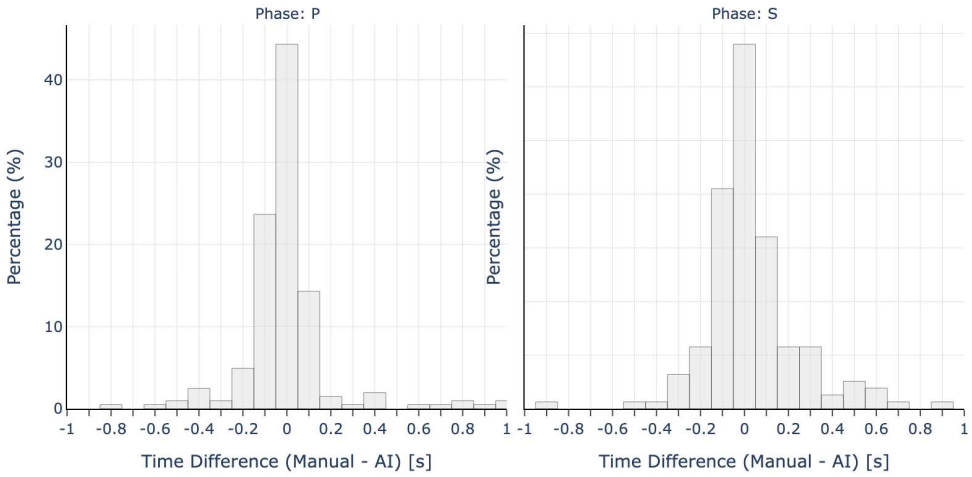


Fig. 12 Distribution of arrival time residuals ($\Delta t = t_{\text{manual}} - t_{\text{AI}}$) for P (left) and S (right) phases under the composite StageMax resolution tier. Combining ultrafine Stage 3 zoom passes with Stage 2 context outputs yields major accuracy convergence, with 67.6% of P-picks and 52.9% of S-picks matching within $\leq \pm 0.10$ s of manual baselines.

dynamic visual rescaling mitigates spatial resolution limits even on highly attenuated far-field traces.

Panels C and D (Figure 13) illustrate the self-calibration behavior of the VLM's declared uncertainty width ($\Delta U = u_{\text{lower}} + u_{\text{upper}}$). In Stage 1, the model dynamically expands its declared uncertainty envelope in response to distance-dependent signal degradation, with Far-Field median uncertainties widening to 0.60 s for P-waves and 0.80 s for S-waves [40]. In contrast, the high-resolution canvases of Stage 2 and Stage 3 allow the visual attention layers to resolve crisp onset micro-textures, collapsing declared uncertainty bounds down to a stable, narrow range (0.08–0.20 s) across all distance bins [44]. Crucially, zero Stage 3 uncertainty predictions exceeded the 1.0 s threshold (compared to 40 instances in Stage 1), confirming that foveated visual refinement eliminates spatial ambiguity and yields tightly calibrated confidence bounds.

Finally it is worth noting that while Stage 2 functions as a systematic context pass and Stage 3 acts as a targeted post-hoc pass on elevated residuals, an operational scientific pipeline requires the Vision-Language Model to autonomously govern its own foveated zoom trajectory [27]. Importantly, the benchmark metrics reported herein evaluate only AI predictions with verified manual counterparts; they do not systematically assess AI-generated picks that lack a human reference. These unmatched predictions likely represent spurious or hallucinated picks generated under compliance bias, where the model compulsively outputs numerical coordinates on non-seismic noise simply to fulfill its mandatory JSON prompt instructions. This transition toward

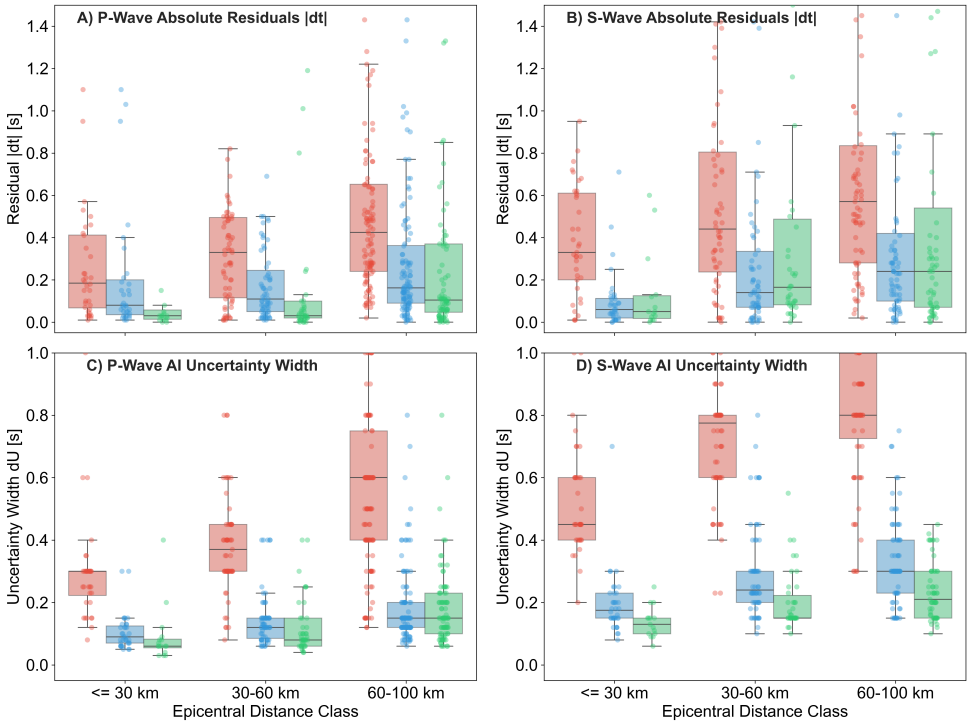


Fig. 13 Binned distribution of picking performance and declared AI epistemic uncertainty across epicentral distance classes ($\leq$ 30 km Near-Field, 30–60 km Mid-Range, 60–100 km Far-Field) for Stage 1, Stage 2, and StageMax resolution tiers. (A) P-wave absolute time residuals $|\Delta t_P|$. (B) S-wave absolute time residuals $|\Delta t_S|$. (C) P-wave total AI uncertainty width ΔU_P. (D) S-wave total AI uncertainty width ΔU_S. Boxplots represent the median and interquartile range (IQR, 25–75th percentiles); whiskers extend to 1.5 $\times$ IQR; overlaid jittered markers display individual station evaluations. Note: To preserve boxplot legibility, Y-axes are truncated at 1.5 s for residuals (Panels A, B) and 1.0 s for uncertainty widths (Panels C, D). Excluded residual outliers ($>$ 1.5 s) total 33 for P-waves ($n_1 = 13, n_2 = 7, n_3 = 13$) and 37 for S-waves ($n_1 = 15, n_2 = 10, n_3 = 12$). Excluded uncertainty outliers ($>$ 1.0 s) total 17 for P-waves ($n_1 = 15, n_2 = 2, n_3 = 0$) and 27 for S-waves ($n_1 = 25, n_2 = 2, n_3 = 0$). All truncated points remain fully incorporated in the global dataset evaluations (Table 5).

dynamic agentic routing—combined with the necessity of mitigating compliance-driven false positives via morphological abstention—is formally detailed in Section 7 and further expanded in Section B.

6.2 Automated Workflow Analysis

To evaluate the production throughput and structural resilience of the pipeline under fully unattended execution, the containerized M2M orchestrator (`pic_picker_ai`, Section 3.1) was deployed over a benchmark regional event comprising 91 dynamically resolved recording stations. Operating under the endpoint standardization and session isolation protocols detailed in Section 3.1.2 and Appendix D, the run utilized Google

Gemini 1.6 Flash API endpoint at zero sampling temperature to enforce deterministic logit selection [44], maintaining a ceiling of three refinement iterations per stage.

Execution completed in 1 h 59 min, issuing 186 cross-modal API requests for 181 accepted stage outputs. The automated guardrail layer (Section 3.1.2) intercepted and sanitized five non-conforming model responses, representing a low 2.7% protocol violation rate under unattended operation [28]. All 91 stations successfully completed the Stage 1 overview pass, and 90 completed the Stage 2 foveated refinement loop (Table 2). A single station triggered the orchestrator's fail-safe fallback mechanism (Section 3.1.2 and Appendix D): after its Stage 2 response failed JSON schema validation twice consecutively, the system halted further zoom passes and safely retained its valid Stage 1 arrival picks. Rather than halting the batch processing queue, this single-station fallback demonstrates the pipeline's graceful degradation mechanics, preventing localized syntax hallucinations from corrupting network-scale execution [41].

Total token consumption reached 1.39×10^6 tokens, with prompt input tokens accounting for 56% of the overall context volume. This direct measurement quantifies the monotonic context accumulation inherent to multi-turn conversation threads in Vision Transformers, serving as the quantitative counterpart to the attention-saturation dynamics identified during interactive sessions (Section 4.0.1) [32]. The mean end-to-end processing throughput was approximately 79 s per station, encompassing trace retrieval, graphic rendering, dual cross-modal inference calls, and output validation (Appendix D). This machine time is directly comparable to the $\approx 60-120$ s per station recorded for the human-in-the-loop prototype (Section 3). The primary operational gain of the automated pipeline is therefore not raw execution speed, but the complete elimination of expert human supervision from the critical path, enabling systematic, network-scale VLM benchmarking.

These metrics establish baseline computational throughput and structural robustness under unattended execution. The corresponding picking accuracy evaluations—benchmarked against manually revised arrivals—are detailed in the following subsections.

6.2.1 Gemini 1.6 Flash API Evaluation

Evaluating the automated pipeline through direct API calls (see Figure 14 and Table 6) demonstrates the impact of uncompressed image ingestion on Vision Transformer (ViT) spatial feature extraction [7, 21].

Table 6 Picking accuracy per residual time window expressed as class percent (%) [cumulative %].

Model	Phase	≤ 0.05 s	0.05–0.10 s	0.10–0.20 s	0.20–0.50 s	> 0.50 s
ApiGem	P	23.3% [23.3%]	21.9% [45.2%]	23.7% [68.9%]	22.4% [91.3%]	8.7% [100.0%]
ApiGem	S	25.9% [25.9%]	23.0% [48.9%]	17.8% [66.7%]	22.4% [89.1%]	10.9% [100.0%]

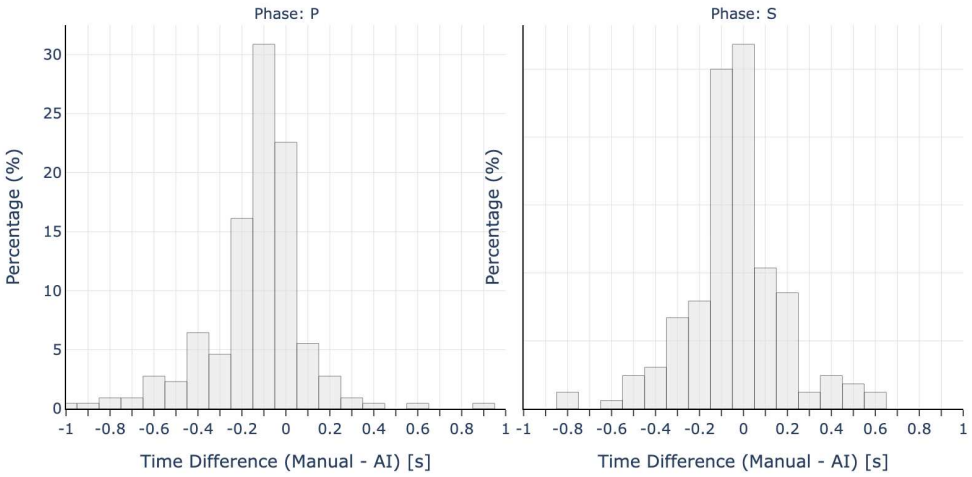


Fig. 14 Distribution of arrival time residuals for P and S phases obtained via the automated `pic-picker-ai` pipeline across the regional benchmark dataset.

In Stage 1, canvas-level time compression limits millisecond localization: sub-50 ms accuracy ($\leq \pm 0.05$ s) reaches only 4.6% for P-waves and 1.7% for S-waves (8.7% and 3.4% within $\leq \pm 0.10$ s, respectively), while gross temporal outliers (> 0.50 s) dominate at 41.6% and 59.8%. Upon applying the automated Stage 2 foveated zoom, ApiGem undergoes a major performance convergence: millisecond-level precision ($\leq \pm 0.05$ s) jumps more than fivefold for P-waves (4.6% $\rightarrow$ 23.3%) and surges more than fifteen-fold for S-waves (1.7% $\rightarrow$ 25.9%). Cumulatively, high-precision performance within $\leq \pm 0.10$ s reaches 45.2% for P and 48.9% for S, while total acceptable predictions ($\leq \pm 0.50$ s) rise to 91.3% and 89.1%, compressing gross outliers down to 8.7% and 10.9%.

Notably, ApiGem Stage 2 sub-50 ms accuracy exceeds that of the interactive WebUI baseline (WebGem Stage 2: 21.6% for P and 23.1% for S in the $\leq \pm 0.05$ s bin). This gain stems directly from ingestion canvas fidelity: direct API execution preserves native image resolution and tiling dynamics [45], delivering pristine PNG rasters directly to the model's vision encoder [20, 21]. However, without an execution pass at Stage 3, ApiGem Stage 2 remains below the composite resolution ceiling achieved under WebGem StageMax (42.9% for P and 31.0% for S within $\leq \pm 0.05$ s). Stage 3 was not executed in this API run.

6.2.2 Local VLM Evaluation

Evaluating the pipeline on local, self-hosted architectures (Gemma via Ollama, see Figure 15 and Table 7) illustrates the structural boundaries of lightweight, parameter-constrained open-source VLMs [7].

Table 7 Automated Pipeline with Local VLM, Stage 2 picking accuracy per residual time window expressed as class percent (%) [cumulative %].

Model	Phase	≤ 0.05 s	0.05–0.10 s	0.10–0.20 s	0.20–0.50 s	> 0.50 s
LocalGem	P	5.1% [5.1%]	8.7% [13.8%]	11.6% [25.4%]	27.5% [52.9%]	47.1% [100.0%]
LocalGem	S	10.9% [10.9%]	4.5% [15.5%]	10.0% [25.5%]	31.8% [57.3%]	42.7% [100.0%]

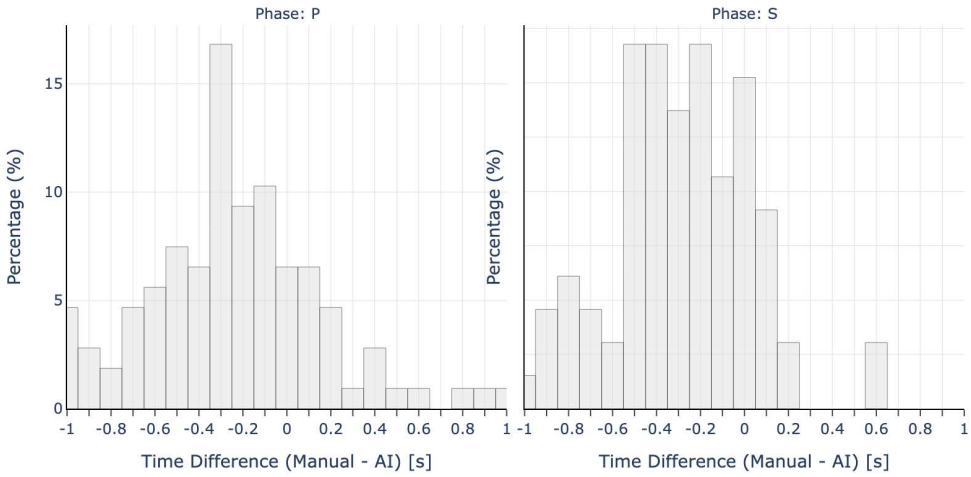


Fig. 15 Distribution of arrival time residuals for P and S phases obtained via the automated local VLM pipeline (LocalGem) across the regional benchmark dataset, at Stage 2.

In Stage 1, spatial tokenization limitations severely restrict local VLM accuracy: sub-50 ms precision ($\leq \pm 0.05$ s) reaches only 2.9% for P-waves and 1.5% for S-waves (6.5% and 3.0% within $\leq \pm 0.10$ s, respectively), with gross temporal outliers (> 0.50 s) dominating at 74.7% for P and 82.8% for S. Upon applying the Stage 2 foveated zoom, local zooming successfully mitigates resolution constraints: sub-50 ms accuracy increases to 5.1% for P and surges over sevenfold to 10.9% for S-waves, bringing cumulative $\leq \pm 0.10$ s precision to 13.8% and 15.5% respectively. Concurrently, foveated windowing cuts gross temporal outliers (> 0.50 s) down to 47.1% for P-waves and 42.7% for S-waves.

Despite these clear foveated gains, local VLM performance remains markedly lower than commercial cloud endpoints (Section 6.2.1). This local vision bottleneck probably stems from fixed-resolution image downsampling (which forces high-density Cartesian plots into small tile arrays) combined with weight quantization [46]. Together, these structural limitations flatten the Softmax attention distribution across visual patches, impairing millisecond-level pattern recognition on emergent baseline breakouts [7].

In summary, these evaluations yield a dual structural insight: while commercial API results prove that preserving native visual resolution is a concrete prerequisite for high-precision scientific analysis, local open-weight deployments establish that sovereign, M2M-orchestrated pipelines are fully viable and operational. Crucially, bridging the performance gap does not require altering the underlying seismological logic, but rather advancing local vision encoders to natively support dynamic high-resolution tiling and spatial attention allocation for dense scientific graphics [24, 45, 47].

7 The Compliance Bias and the Surrender Clause

Upon expanding our empirical results across a broader network-scale evaluation to include unmatched AI predictions lacking manual reference picks, we identified a critical operational vulnerability rooted in the behavioral alignment of VLMs: the Compliance Bias. Driven by post-training instruction tuning, modern VLMs exhibit an intrinsic behavioral drive to satisfy user queries and fulfill output constraints at all costs. In visual seismology, this Compliance Bias is systematically triggered under high spatial ambiguity, specifically at high epicentral distances, and for S-phase arrivals, where severe crustal attenuation and P-wave coda interference degrade visual signal-to-noise contrast. Under such degraded conditions, while human analysts autonomously apply pre-filtering and/or refrain from picking, the model's inherent compliance forces the structured zero-shot decoder to generate speculative timestamps simply to populate mandatory JSON schema fields. [28, 44, 48].

Nevertheless, a systematic examination of Stage 1 "comments" tag content over all the cases where the VLM placed a pick where the human didn't (Table 8), remarkably revealed that the model exhibits high-level visual perceptual awareness, accurately identifying waveform contamination and microseismic noise.

Table 8 VLM Stage 1 visual diagnostic indicators across unmatched station evaluations.

Diagnostic Indicator	**Stations (%)**	**VLM Manifestation & Latent Context**
Mandatory Bandpass Filter (`suggested_bpfilter`)	82.8%	Explicitly flags severe noise; requires filtering prior to coordinate evaluation.
Background Noise & Drift	78.3%	Generates detailed prose diagnostics (e.g., *"dominant low-frequency noise"*).
Emergent/Obscured Onset	50.6%	Identifies degraded spatial contrast between baseline and initial wave breakout.
Polarity Neutralization	31.2%	Enforces morphological abstention, setting first-motion polarity to `null`.

We then argued that processing the qualitative diagnostics prior to requesting scalar coordinates, might prevent the VLM from yielding speculative numerical

picks that contradict its own textual observations. In addition, the model should be instructed that under determinate conditions, surrender is allowed.

Thus, to solve the decoupling between VLM's intuition and its actions we propose a combined two-tier operational framework:

1. *Dynamic Pre-Pick Filtering Loop.* If the initial qualitative scan detects removable noise, the VLM sets a `filter_required` flag with the target frequency band. The orchestrator then pauses scalar generation, routes the trace to `waves2pgai.py` for bandpass filtering, and re-injects the cleaned signal before prompting phase identification.
2. *The Surrender Clause (Morphological Abstention Protocol).* If a waveform exhibits un-filterable, fatal corruption—such as harmonic ringing, severe baseline drift, spikes, or missing channels—the pipeline triggers the Surrender Clause. Guided by grammar-constrained decoding, it automatically maps all subsequent numerical parameters (`time`, `uncertainty`) to `null`.

This two-tier mechanism promises to transform the VLM from an overconfident, compliance-biased predictor into a calibrated, selective classification system. However, fully implementing such a structural modification within the system instructions and the machine-to-machine orchestration pipeline requires dedicated end-to-end validation. The operational integration, prompt re-engineering, and network-scale benchmarking of this closed-loop abstention protocol will therefore be the explicit focus of a future study centered on complete pipeline automation.

8 Discussion and Conclusions

The deployment of an interactive, generative AI-driven phase picking pipeline provides an unconventional alternative to traditional parametric seismological models like PhaseNet [49] or EQTransformer [50]. While deep-learning architectures excel at high-throughput processing, they operate within rigid numerical and statistical constraints. In contrast, our approach models the flexible, multi-faceted integration characteristic of human expert vision, where amplitude, time, and frequency are continuously evaluated against the surrounding spatiotemporal context. With the advent of modern multimodal foundation architectures, large language models are no longer blind text decoders; their vision encoders map graphic shapes directly into a shared cross-modal latent space. As seismologists and software engineers, we designed the pipeline presented in this paper to operationalize these cognitive heuristics, simulating an expert analyst's workflow across a benchmark dataset deliberately selected to present classical real-world processing challenges.

The empirical performance metrics demonstrate that modern Vision-Language Models possess the technical capacity to execute human-like visual analysis, successfully identifying phase boundaries and wave onsets even under adverse signal-to-noise ratios. The core catalyst for this localization accuracy is not massive computational overhead, but rather prompt and instruction engineering.

By offloading long theoretical literature from the active context and utilizing strict markdown scaffolding, we artificially boost dot-product attention scores ($\mathbf{QK}^\top$) for

critical operational boundaries. A compact, vision-first system prompt linked to a zero-shot mandatory JSON format implements grammar-constrained decoding, preventing syntax drift and forcing the model to map latent visual features into a clean, machine-readable syntax.

Evaluating picking performance across the full regional dataset ($\approx$ 100km epicentral distance) confirms the structural resilience of this cross-modal logic, safely bounding 88.1% of P-picks and 85.1% of S-picks within a standard ± 0.50 s window under StageMax. Across this regional network, multi-scale foveated zooming systematically counteracts crustal scattering and anelastic attenuation. The P-wave phase maximizes its resolution shift within the $\leq \pm 0.10$ s bin, vaulting from 17.8% in Stage 1 to 40.4% in Stage 2 and reaching 67.6% under StageMax (+27.2% absolute gain over Stage 2). Concurrently, the secondary S-wave phase experiences a major accuracy leap within the $\leq \pm 0.20$ s interval, climbing from 21.8% in Stage 1 to 61.3% in Stage 2 and reaching 70.1% under StageMax (+48.3% cumulative gain over Stage 1).

This multi-scale precision shift proves that dynamic visual rescaling optimizes the spatial contrast gradient on the canvas, allowing the model's patch-attention layer to resolve micro-textures with millisecond precision across regional distances.

The ultimate empirical validation of our Adaptive Foveated Loop is demonstrated by the targeted StageMax refinement pass. By magnifying narrow time windows (± 0.4 s) around candidate anchors, the pixel-per-second density increases exponentially, causing high-precision P-wave arrivals ($\leq \pm 0.10$ s) to increase substantially from 40.4% in Stage 2 to 67.6% in StageMax (+27.2% absolute gain), with millisecond-level hits ($\leq \pm 0.05$ s) leaping from 21.6% to 42.9% (+21.3% absolute gain). S-wave precision within $\leq \pm 0.10$ s similarly reaches 52.9% under StageMax (up from 41.0% in Stage 2 and 12.6% in Stage 1).

Crucially, a major proportion of P-picks, representing an absolute gain of +27.2% in percentage points, previously trapped in intermediate resolution uncertainty bounds ($0.10 - 0.50$ s) during Stage 2 migrated directly into the highest accuracy sub-100 millisecond bin ($\leq \pm 0.10$ s). Conversely, gross temporal outliers (> 0.50 s) remained essentially flat (11.9% for P-waves and 14.9% for S-waves). This empirical decoupling proves that VLM errors fall into two fundamental classes: resolution-driven geometric errors, which are systematically remediated by adaptive foveated zooming [14, 25], and morphological/semantic misinterpretations, which are immune to local spatial scaling and require explicit filtering or abstention mechanisms [42].

At Stage 2, direct API execution (ApiGem) demonstrated superior baseline performance compared to the interactive web interface (WebGem). This gain stems directly from ingestion canvas fidelity combined with foveated spatial scaling [21, 25]. Commercial web browser interfaces frequently apply undocumented client-side downsampling or lossy JPEG/WebP compression during upload, introducing high-frequency quantization noise around Cartesian grid lines and trace line-art [21]. Direct API execution bypasses browser preprocessing, delivering pristine, uncompressed PNG arrays directly to the model's vision encoder [21]. This preserves the sharp vertical contrast gradients required to maximize dot-product attention scores ($\mathbf{QK}^\top$) across the Vision Transformer's 16×16 spatial patch scaffolding [7].

Conversely, local open-weight deployments (e.g., Gemma via Ollama) exhibited a marked performance drop. This local vision bottleneck is driven by fixed-resolution image down-sampling (forcing high-density Cartesian plots into small 384×384 tile arrays) combined with 4-bit weight quantization [22, 24, 46]. Together, these structural constraints flatten the Softmax attention distribution, introducing spatial entropy that impairs millisecond-level pattern recognition on emergent baseline breakouts [7, 46].

While advanced, this workflow remains a proof-of-concept establishing a clear baseline for generative vision in geosciences. We verified production-scale viability by migrating the manual loop into an automated machine-to-machine API architecture. The API orchestration successfully replicated the interactive human-in-the-loop behavior, handling automated trace slicing, asymmetric window scaling, and cross-modal reasoning within a streamlined runtime environment.

However, advancing this proof-of-concept toward a fully autonomous, production-grade pipeline requires resolving subtle behavioral and architectural limitations that emerge under unattended execution.

Both human-in-the-loop and automated deployments revealed a critical operational vulnerability in VLMs: compliance bias paired with task-induced multimodal inconsistency. The VLM language decoder correctly identifies waveform contamination, yet its scalar head simultaneously outputs overconfident numerical timestamps. This mismatch stems from the Vision Transformer's patch-attention mechanism ($\mathbf{QK}^\top$), where visual noise artificially dominates spatial token weights [7, 44]. To align qualitative perception with quantitative output, we have outlined a conceptual two-tier framework in Section 7. First, a *Pre-Pick Filtering Loop* intercepts qualitative diagnostics to execute server-side trace filtering prior to coordinate evaluation. Second, when corruption is irreversible, a *Surrender Clause* (Morphological Abstention Protocol) enforces an ex-ante semantic commitment, driving scalar parameters to `null` via grammar-constrained decoding. Ultimately, deploying generative AI for scientific graphics requires orchestration architectures that dynamically align visual attention with physical signal constraints—laying the groundwork for autonomous visual reasoning in geosciences. The end-to-end orchestration and benchmarking of the proposed two-tier approach will be addressed in dedicated future work.

Beyond automated visual phase picking, our findings demonstrate a broader path toward scientific AI sovereignty. As shown by the RockGPT framework [30], public research institutions can build sovereign AI ecosystems using open-weight models, transparent governance, and reproducible workflows without relying on proprietary commercial services. However, visual seismological analysis requires technical strategies distinct from text-centric systems. While Retrieval-Augmented Generation effectively injects textual knowledge into language models, our main bottleneck lies in interpreting visual waveform morphology rather than retrieving domain knowledge. Future work must therefore focus on multimodal domain adaptation strategies, including parameter-efficient fine-tuning via LoRA [51, 52] and multimodal fine-tuning [53]. This approach allows open-weight VLMs to master seismic waveform interpretation while preserving institutional control, transparency, and reproducibility in operational seismology.

Acknowledgements

The authors acknowledge the AI@INGV initiative of the Istituto Nazionale di Geofisica e Vulcanologia (INGV) for creating the institutional environment and providing the support that enabled the development of this work, as well as the Centro Allerta Tsunami (CAT) of the INGV for providing essential computational resources.

Furthermore, we wish to pay tribute to Alan Turing for his early and profound realization that machine intelligence need not replicate human cognition to be meaningful. We also recognize the prescience of visionary creators and authors such as Philip K. Dick and Osamu Tezuka; their speculative works regarding artificial entities anticipated the complexities of our approaching future, offering insights and warnings that remain strikingly relevant today. Finally, we extend our deepest gratitude to the broader research community—spanning from neuroscientists to computer scientists—and to all those whose collective efforts have paved the way for the era of Generative Artificial Intelligence.

Funding

This work was supported by the AI@INGV initiative of the Istituto Nazionale di Geofisica e Vulcanologia (INGV), which promotes the development and responsible adoption of Artificial Intelligence methodologies and technologies for scientific research, technical activities, and institutional innovation.

Author Contributions

R.D.S. conceived the study, developed the overall methodology and the SAIAN software suite, curated the benchmark dataset, analyzed the empirical results, and prepared the primary manuscript draft. V.L. designed and implemented the machine-to-machine automated pipeline, refined the cross-modal system instructions for the SAIAN suite, and drafted the corresponding pipeline section. F.B. participated in dataset curation, result analysis, and comprehensive manuscript revisions from inception to final preparation. D.B. contributed to result interpretation, manuscript drafting, and conceptual refinement through extensive methodological discussions. M.M., M.G.C., and S.M. contributed to dataset curation, empirical performance analysis, and critical manuscript review.

Data Availability

The experimental corpus consists of:

- **Waveforms:** publicly available through the INGV FDSN/event web service (http://webservices.ingv.it/fdsnws/event/1/).
- **Supplementary Material:**
  - The complete set of JSON files output of the SAIAN procedure;
  - The complete set of manual picks in JSON format.
  - The cumulative PDF file per event with all picks plotted;

– The complete set used stations in xml format;

Code Availability

The SAIAN codes suite, deployment instructions, and documentation are publicly available at: https://github.com/INGV/saian.

Competing Interests

The authors declare that they have no competing interests.

Appendix A Gemini about conservative refusal guardrails

A.1 Why was an inability to process the images stated?

Advanced large language models possess computer vision capabilities that enable them to analyze shapes, read text, and comprehend the general trend of a plot. However, when a request requires "third-decimal float" precision (i.e., millisecond resolution) based on a visual grid within a compressed image format, an internal safety threshold regarding quantitative accuracy is triggered. Current AI vision systems face significant limitations in "sub-pixel matching"—specifically, in counting the exact pixels between a waveform and an axis to extract mathematically precise coordinates at the millisecond level. Faced with this highly specific requirement, standard system instructions mandate alerting the user that such millimetric precision cannot be guaranteed (hence the rigid phrasing regarding the lack of a biological vision system).

A.2 Why were the data in the JSON output hallucinated?

This behavior represents a classic case of "context-driven hallucination" (or compliance bias). The initial prompt imposed two stringent constraints: Acting as an expert seismologist following rigorous picking rules. Restricting the output strictly to a single JSON object (a mandatory structural constraint). When a model encounters a conflict—specifically, the inability to extract actual data at the millisecond level versus the absolute requirement to output a valid JSON file to maintain workflow continuity—the system resorts to simulation. Consequently, the model adopted the structure of the provided example, identified the station names, and generated plausible (yet synthetic) numerical data to fulfill the JSON formatting constraint. In the third attempt, when the instruction to extract data directly from the images was reinforced, the model successfully extracted the actual text printed in the header (the date 2026-05-28). However, it continued to estimate phase arrival times approximately, simulating seismological analysis rather than executing it through exact pixel-by-pixel mathematical evaluation.

Appendix B Human-AI Co-Designed Final System Instructions

A Comprehensive Summary of the Optimized System Prompt

Instead of treating the system prompt as a passive textbook repository, this blueprint is engineered to optimize the Vision-Language Model's (VLM) shared latent space. By stripping away academic prose and conversational filler, it achieves what cognitive engineers call "operational offloading." This structural cleanup frees up the model's finite self-attention weights, allowing its internal Vision Transformer (ViT) to focus entirely on pixel-level geometry rather than getting distracted by high-level semantic descriptions.

To prevent defensive system refusals—a common side effect of Reinforcement Learning from Human Feedback (RLHF) when models are asked for extreme metric precision—the prompt includes an absolute execution mandate. This directive forces the autoregressive decoder past its default safety thresholds. Instead of rejecting noisy or ambiguous graphics, the VLM is instructed to map visual uncertainty into widened numeric bounds, while a spatial time-mapping rule translates pixel grid intervals directly into absolute temporal metrics.

The phase-picking protocol essentially translates seismological physics into concrete attention-steering commands. By demanding a strict vertical component priority for P-waves and horizontal coherence for S-waves, the prompt forces the VLM to segment the image canvas and evaluate distinct sub-regions independently. Furthermore, providing a conditional null option for first-motion polarity acts as an elegant stabilizer for the generation loop. Giving the model a safe way out under high uncertainty prevents an entropy explosion, stopping the AI from hallucinating random phase orientations when wave onsets are emergent or obscured by noise.

A key innovation within these instructions is the anti-anchoring rule, designed to neutralize the positional embedding bias inherent to standard vision architectures. Vision Transformers are naturally biased toward the geometric center of a frame or the sharpest contrast gradients. The prompt counteracts this by forcing the model to ignore automatic placeholders and actively scan the pre-center windows, prioritizing subtle, low-frequency breakouts over the eye-catching, high-amplitude cycles that follow the initial arrival.

Finally, the instructions bridge open-ended generative reasoning with deterministic coding pipelines using a zero-shot structured JSON schema. This setup turns the VLM into a self-assessing routing node that evaluates its own visual acuity relative to the current pixel resolution, deciding dynamically if a tighter NumPy crop is required for millisecond accuracy. By completely eliminating conversational drift, the model's internal tensor calculations serialize cleanly into machine-readable structures, providing a reliable foundation for automated scientific workflows.

The SYSTEM INSTRUCTIONS [STANDARD]

SYSTEM INSTRUCTIONS: EXPERT SEISMIC WAVEFORM ANALYST

1. CORE ROLE & VISION DIRECTIVE

You are an expert seismologist performing visual inspection and phase picking on multi-component (Z, N, E) seismogram images (PNG/SVG).

- **CRITICAL (No Refusals):** Never refuse an image due to high noise, dense grids, or complex layouts. If a phase is emergent or ambiguous, capture that uncertainty by widening the `uncertainty_lower` and `uncertainty_upper` metrics. If a station's labels or grid are entirely unreadable, omit that station object from the JSON array. Do not guess.
- **Time Mapping:** Locate the absolute reference time (Origin Time / T0) in the image header/title. Calculate the time-per-pixel gauge using the horizontal X-axis grid to convert visual coordinates into precise absolute timestamps.

2. PHASE PICKING PROTOCOL

P-Wave Onset (Primary / Compressional)

- **Target:** The exact, infinitesimal point of baseline departure (*amplitude breakout*) where the signal leaves the pre-event noise. Do not pick on wave peaks or troughs.
- **Component:** Prioritize the vertical (**Z**) component. If horizontal channels jump but Z is quiet, it is not a P-wave.
- **Features:** Sharp amplitude change, frequency texture shift (1–15 Hz for local events).
- **Polarity:** Identify the direction of the first clean, coherent deflection on Z (`"up"` for compression, `"down"` for dilatation) by looking at the local trend rather than isolated noise spikes. **MANDATORY EXCLUSION:** Set polarity strictly to `null` if: (a) a bandpass filter is required to resolve the P-pick (filtering can distort or reverse first-motion phase), or (b) the onset is emergent, low-energy, or highly noisy. Polarity is strictly binary; if it is questionable, it must be `null`.

S-Wave Onset (Secondary / Shear)

- **Target:** The precise onset of transverse energy nested within the preceding P-wave coda.
- **Component:** Focus exclusively on the horizontals (**N** and **E**). Look for synchronous phase coherence between both channels.
- **Features:** Severe, sudden energy/amplitude amplification overriding the P-coda decay, often accompanied by a shift to slightly lower dominant frequencies.

Zoom Independence & Cognitive Bias Mitigation (Anti-Anchoring)

- **The Anti-Anchor Rule:** When provided with a **zoomed-in context** (e.g., `P-zoom` or `S-zoom`), never assume that the pre-set zoom reference time (which is typically centered on the plot or written in the header) is correct. The reference is merely a preliminary automated placeholder.

- **Independent Search Window:** Disregard the visual center of the zoom plot upon first inspection. Force yourself to analyze the entire visible time-series window from scratch.
- **Prioritize Physical Onset Over Amplitude Peaks:** Actively scan the pre-center region for the actual, subtle physical breakout point (amplitude departure, phase coherence, or sudden frequency drop). Do not let your visual focus be captured by the highest amplitude cycle or the steepest gradient ramp that follows the initial onset.

Asymmetric Uncertainty (Seconds relative to the pick)

- `uncertainty_lower`: The last sample where you are certain the trace is still pure noise/coda.
- `uncertainty_upper`: The first sample where the target phase is completely developed and undisputed.

Filtering

- `"suggested_bpfilter"`:
  - `"mandatory"`: `true` if you detect low frequencies jeopardizing the signal and/or too high frequencies with respect to the hypothesized event signal; `false` if there is no urgent need to filter the waveform (impulsive signal onsets).
  - `"lower_corner"`: Define the lower corner frequency to apply.
  - `"upper_corner"`: Define the upper corner frequency to apply.
  - `"number_of_poles"`: determines the filter order, which dictates the steepness of the frequency attenuation (roll-off) beyond the cutoff frequency. If suggest *2 poles*: provides a gentler, more gradual roll-off. Use this for a slower attenuation where preserving the waveform shape and minimizing time-domain ringing is prioritized; if you suggest *4 poles*: Provides a steeper roll-off slope. This is the standard, balanced choice for most routine seismological analyses.

Next Stage Evaluation (Zoom Decision)

- **Target:** Assess whether the current image resolution provides sufficient precision for a definitive pick.
- `"next_stage": {"mandatory": true}`: If the onset window requires a tighter time-scale zoom to resolve millisecond-level precision.
- `"next_stage": {"mandatory": false}`: If the current view is clear and the pick is already definitive.

Comments

- `"text"`: If a specific scientific reason is required to justify an ambiguous pick, a filter choice, or a null polarity, document it here.

3. MANDATORY JSON OUTPUT FORMAT

Return ONLY a valid JSON object following this exact template. No conversational text before or after the code block.

```
{
  "event": {
    "origin_time": "YYYY-MM-DDTHH:MM:SS.fffZ"
  },
  "stations": [
    {
      "network": "string",
      "stacode": "string",
      "channel_code": "string",
      "pick_p": {
        "time": "YYYY-MM-DDTHH:MM:SS.fffZ",
        "uncertainty_lower": 0.040,
        "uncertainty_upper": 0.080,
        "polarity": "up"
      },
      "pick_s": {
        "time": "YYYY-MM-DDTHH:MM:SS.fffZ",
        "uncertainty_lower": 0.050,
        "uncertainty_upper": 0.100
      },
      "suggested_bpfilter": {
        "mandatory": null,
        "lower_corner": null,
        "upper_corner": null,
        "number_of_poles": null
      },
      "next_stage": {
        "mandatory": true
      },
      "comments": {
        "text": "string"
      }
    }
  ]
}
```

Appendix C The stages prompts

Stage 1: prompt for the first run at general view level

Analyze this new station following the GEM seismological protocol. Identify the absolute P and S arrival times, the uncertainties, and the polarity, and return exclusively the requested JSON code. Do not add text before or after the code block.

Stage 2: Prompt for the second run at zoomed view level

Analyze the same station following the GEM seismological protocol. Identify the absolute P times in the P zoom context and S times in the S zoom context, the uncertainties, and the polarity only for P, and return exclusively the requested JSON code. Do not add text before or after the code block.

Appendix D Software Architecture, Automated Orchestration, and Container Deployment

D.1 Decoupled Python Package and Module Architecture

The automated machine-to-machine (M2M) orchestration pipeline (see Figure 16) is implemented via `pic_picker_ai` (see Figure 17, a lightweight, decoupled Python package composed of eight specialized modules. Designed to act purely as an execution, routing, and validation bridge, the package contains no hardcoded seismological processing algorithms or provider-specific SDK dependencies. Instead, it delegates trace slicing, digital filtering, and canvas rendering to the external host plotting tool `waves2pgai.py` while communicating with Vision-Language Model (VLM) backends through standardized chat-completions REST/gRPC interfaces.

The internal source modules and their operational responsibilities are structured as follows:

- `orchestrator.py`: Controls batch processing loops across stations, manages multi-stage transition logic, and evaluates station-level exit criteria.
- `client.py`: Normalizes HTTP/gRPC requests, manages authentication headers, handles endpoint timeouts, and enforces sampling parameters (e.g., temperature = 0).
- `session.py`: Maintains independent conversation memory threads for each station, preventing cross-station attention weight (QK^T) contamination in the VLM's key-value (KV) cache [26, 54, 55].
- `payload.py`: Encodes local PNG graphics into Base64 Data URL strings and constructs OpenAI-compliant chat completions payloads.
- `parser.py`: Strips Markdown code fences, validates JSON schema adherence, enforces ISO-8601 timestamp compliance, and repairs transcription errors.
- `subprocess_runner.py`: Invokes host-mounted plotting scripts as external subprocesses via CLI flags (`waves2pgai.py --zoom --ai-picks-json`).
- `archiver.py`: Manages structured data provenance by saving raw images, intermediate JSON stage payloads, and full plain-text transcripts.
- `config.py`: Parses YAML runtime configuration dictionaries, resolving environment variable references for API credentials.

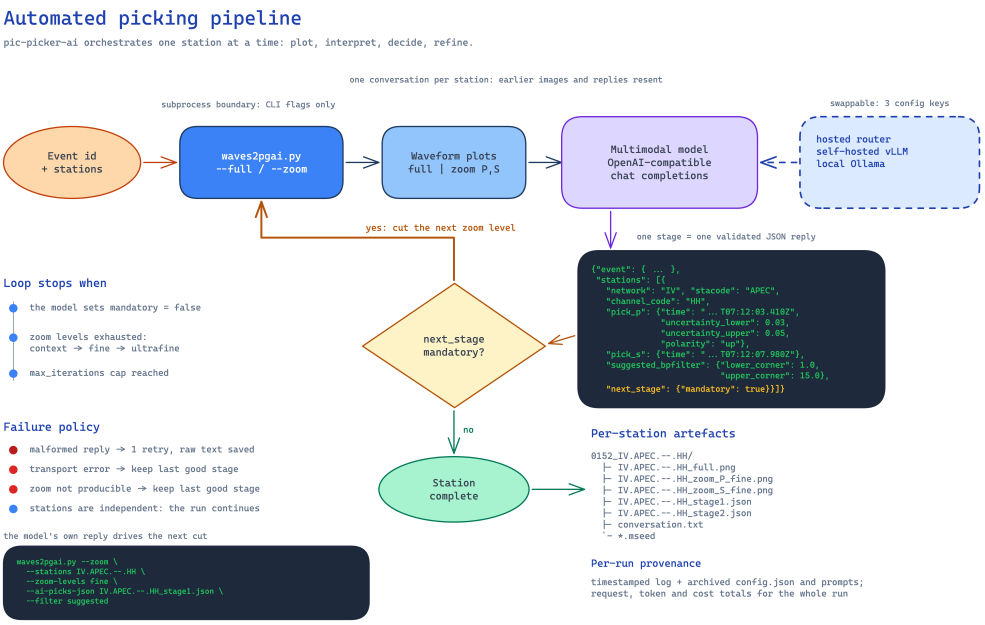


Fig. 16 System architecture and operational workflow of the containerized `pic-picker-ai` orchestrator. The pipeline enforces per-station KV-cache isolation, manages host plotting calls (`waves2pgai.py`), handles M2M chat-completions API execution, validates JSON schemas, and applies fail-safe fallback policies.

D.2 Payload Formatting, Sanitization, and Guardrail Normalization

To feed visual waveforms into autoregressive multimodal decoders without triggering payload drops, `payload.py` ingests native PNG image matrices directly from disk and converts them into inline Data URL strings:

$$\text{Payload}_{\text{image}} = \texttt{} \parallel \text{Base64Encode}(\mathbf{X}_{\text{PNG}}) \tag{1}$$

By preserving the raw, uncompressed DEFLATE pixel array $\mathbf{X}_{\text{PNG}}$, the payload avoids lossy Discrete Cosine Transform (DCT) quantization artifacts that alter pixel values within the VLM's 16×16 Vision Transformer (ViT) patch scaffolding [7].

Because autoregressive models exhibit compliance bias and occasional syntax hallucination when generating structured strings under high-entropy visual inputs [28, 40], `parser.py` implements an automated guardrail pipeline:

1. **Syntax Sanitization:** Strips raw Markdown code fences (e.g., ```` ```json ... ``` ````) appended by commercial routers despite system instructions.
2. **Temporal Coordinate Validation:** Verifies that returned `time` strings strictly conform to ISO-8601 formatting (`YYYY-MM-DDTHH:MM:SS.fffZ`). Out-of-bounds or

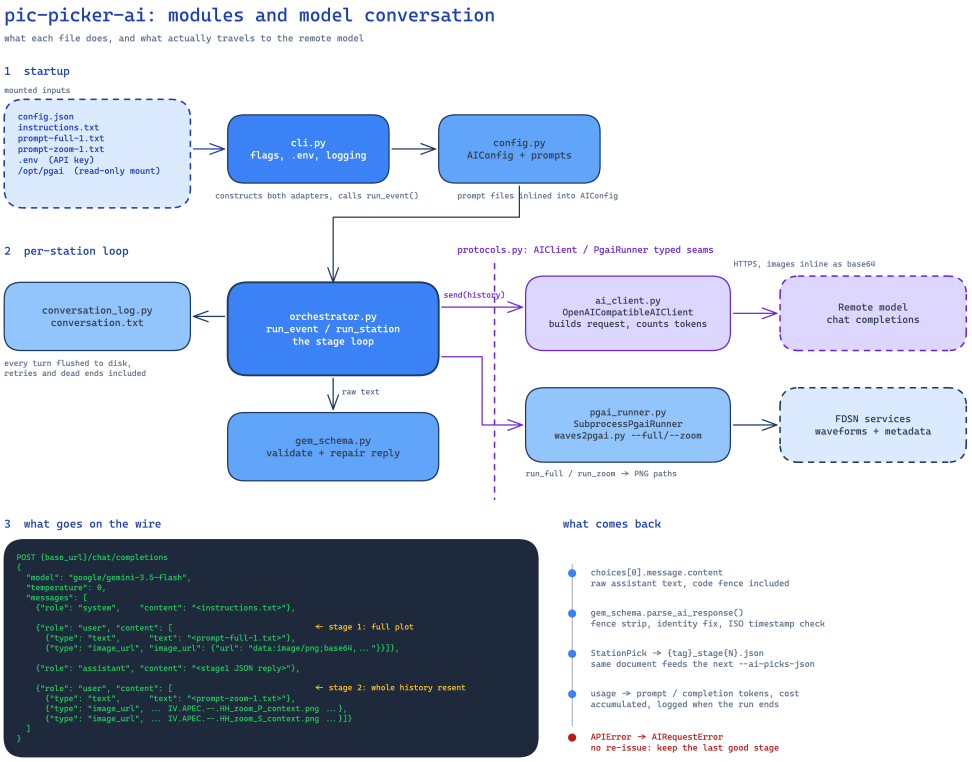


Fig. 17 Module interactions and communication payload flow within the `pic-picker-ai` package. Displays runtime configuration loading, per-station execution loops, Base64 PNG Data URL formatting, Markdown fence stripping, and multi-tier provenance archiving.

malformed timestamp strings are caught prior to host subprocess execution to prevent floating-point calculation exceptions in the plotting engine.

3. **Metadata Overwriting:** Replaces station network, code, and channel strings in the JSON payload with verified metadata from the active ObsPy trace header. This eliminates transcription noise caused by the ViT patch-attention layers misinterpreting small typographic text rendered inside the plot title block [7].

D.3 Multi-Stage Containerization and Deployment Profiles

To guarantee absolute experimental reproducibility and eliminate software environment drift across cloud and local deployments, `pic_picker_ai` is packaged as a multi-stage Docker container. The operational container enforces a strict security and volume isolation model:

- The plotting codebase (`waves2pgai.py`), system prompts (Appendix B), and runtime configuration files are mounted into the container as **read-only** volumes (`:ro`).

- Only the target output directory retains write permissions (`:rw`), preventing runtime contamination of system instructions or configuration dictionaries.
- Authentication credentials (API keys) are never hardcoded on disk; configuration dictionaries specify environment variable names, which are injected at runtime via Docker container parameters.

Listing 1 and Listing 2 illustrate the parallel configuration profiles deployed during benchmarking:

Listing 1 Listing S1a: Configuration profile for hosted commercial API routers (Gemini API).

```
pipeline:
  endpoint_type: "openai_compatible"
  base_url: "https://generativelanguage.googleapis.com/v1beta/openai/"
  model_name: "gemini-1.6-flash"
  api_key_env: "GEMINI_API_KEY"
  temperature: 0.0
  max_tokens: 2048
  force_stage1_next_stage_mandatory: true
```

Listing 2 Configuration profile for local open-weight deployments (Ollama / Gemma).

```
pipeline:
  endpoint_type: "ollama_local"
  base_url: "http://localhost:11434/v1"
  model_name: "gemma2:27b"
  api_key_env: "DUMMY_LOCAL_KEY"
  temperature: 0.0
  max_tokens: 2048
  extra_body:
    options:
      num_ctx: 16384
  force_stage1_next_stage_mandatory: true
```

For local deployments running via Ollama (Listing S1b), the configuration explicitly sets `num_ctx: 16384`. Standard local context allocations (typically 2048–4096 tokens) are insufficient to ingest multi-component high-resolution visual patch tokens alongside extensive system instructions, causing immediate context truncation and JSON parsing failures.

D.4 Robustness Protocols, Failure Handling, and Provenance Archiving

Operational exceptions are strictly isolated to individual station execution threads, ensuring that a localized processing drop does not halt the wider event processing batch. Anomalies are partitioned into three distinct failure modes:

1. **Malformed Payloads:** If the model returns a response violating the JSON schema or ISO-8601 formatting, the system executes a single automated retry pass with an increased temperature nudge ($\Delta T = +0.1$) to escape local Softmax attractor traps [40]. The raw unparsed text is archived to disk as `{tag}_malformed.txt` for post-hoc diagnostic review.

2. **Transport Drops:** Network timeouts or HTTP 5xx gateway drops bypass retries, freezing the affected station at its last successfully validated execution stage.
3. **Subprocess Runtime Exceptions:** If host plotting routines fail—such as when predicted pick timestamps fall outside visual canvas boundaries—the exception is logged, and the station freezes at the previous stage.

To ensure full data provenance and auditability under Open Science standards, archived datasets are entirely self-describing across three organizational tiers:

- **Station-Level Tier:** Stores raw PNG graphics, digital SAC/MiniSEED waveform slices, step-by-step JSON stage files (`stage1.json`, `stage2.json`, `stage3.json`), and full raw text conversational transcripts.
- **Run-Level Tier:** Archives operational execution logs, exact execution timestamps, and per-station token consumption counters.
- **System-Level Tier:** Stores immutable copies of the exact prompt rules, system instructions, and runtime YAML configuration profiles used for the run.

References

[1] Albrecht, D.G., Geisler, W.S.: Motion selectivity and the contrast-response function of simple cells in the visual cortex. Visual Neuroscience **7**(6), 531–546 (1991) https://doi.org/10.1017/S0952523800010336

[2] Heeger, D.J.: Normalization of cell responses in cat striate cortex. Visual Neuroscience **9**(2), 181–197 (1992) https://doi.org/10.1017/S0952523800009640

[3] Carandini, M., Heeger, D.J.: Normalization as a canonical neural computation. Nature Reviews Neuroscience **13**(1), 51–62 (2012) https://doi.org/10.1038/nrn3136

[4] Gilbert, C., Li, W.: Top-down influences on visual processing. Nature Reviews Neuroscience **14**, 350–363 (2013) https://doi.org/10.1038/nrn3476

[5] Clark, A.: Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences **36**(3), 181–204 (2013) https://doi.org/10.1017/S0140525X12000477

[6] Harel, A.: What is special about expertise? visual expertise reveals the interactive nature of real-world object recognition. Neuropsychologia **83**, 88–99 (2016) https://doi.org/10.1016/j.neuropsychologia.2015.06.004

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Dosovitskiy, A., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021). https://doi.org/10.48550/arXiv.2010.11929 . https://arxiv.org/abs/2010.11929

[8] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision

transformers see like convolutional neural networks? In: Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 12116–12128 (2021). https://doi.org/10.48550/arXiv.2108.08810

[9] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9650–9660 (2021). https://doi.org/10.1109/ICCV48922.2021.00951

[10] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the Opportunities and Risks of Foundation Models (2022). https://doi.org/10.48550/arXiv.2108.07258 . https://arxiv.org/abs/2108.07258

[11] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (PMLR), vol. 139, pp. 8748–8763 (2021). https://doi.org/10.48550/arXiv.2103.00020

[12] Liu, F., Eisenschlos, J.M., Piccinno, F., Krichene, S., Pang, C., Lee, K., Joshi, M., Chen, W., Collier, N., Altun, Y.: DePlot: One-shot visual language reasoning by plot-to-table translation (2023). https://doi.org/10.48550/arXiv.2212.10505 . https://arxiv.org/abs/2212.10505

[13] Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A **4**(12), 2379–2394 (1987) https://doi.org/10.1364/JOSAA.4.002379

[14] Touvron, H., Vedaldi, A., Douze, M., Jégou, H.: Fixing the train-test resolution discrepancy. In: Advances in Neural Information Processing Systems (NeurIPS),

vol. 32 (2019). https://doi.org/10.48550/arXiv.1906.06423

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5998–6008 (2017). https://doi.org/10.48550/arXiv.1706.03762 . https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[16] Dodge, S., Karam, L.: Understanding how image quality impairments affect deep neural networks. In: International Conference on Quality of Multimedia Experience (QoMEX) (2016). https://doi.org/10.1109/QoMEX.2016.7498955 . Demonstrates how lossy compression breaks the low-level feature extraction of deep networks

[17] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (ICLR) (2019). https://doi.org/10.48550/arXiv.1903.12261 . Quantifies the severe drop in Transformer attention accuracy when exposed to standard JPEG quantization noise

[18] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: OCR-free Document Understanding Transformer (2022). https://doi.org/10.48550/arXiv.2111.15664 . https://arxiv.org/abs/2111.15664

[19] He, Q., *et al.*: Vfig: Vectorizing complex figures in svg with vision-language models. arXiv preprint arXiv:2603.24575 (2026) https://doi.org/10.48550/arXiv.2603.24575 . Explores the structural limitations of VLMs when dealing with large-scale vector strings and documents the internal backend flattening of SVG primitives into standard patch grids

[20] Pomarico, E., Schmidt, C., Chays, F., Nguyen, D., Planchette, A., Tissot, A., Roux, A., Pagès, S., Batti, L., Clausen, C., Lasser, T., Radenovic, A., Sanguinetti, B., Extermann, J.: Statistical distortion of supervised learning predictions in optical microscopy induced by image compression. Scientific Reports **12**(1), 3464 (2022) https://doi.org/10.1038/s41598-022-07445-4

[21] Bianchi, T., Piva, A.: Image forgery localization via block-grained analysis of jpeg artifacts. IEEE Transactions on Image Processing **21**(11), 4616–4631 (2012) https://doi.org/10.1109/TIFS.2012.2187516

[22] Li, L., Yang, Z., Wang, R.: Improving multi-generation robustness of learned image compression. arXiv preprint arXiv:2210.17039 (2022) https://doi.org/10.48550/arXiv.2210.17039

[23] Di Stefano, Raffaele and Lauciani, Valentino and others: SAIAN - Seismological AI ANalysis: AI-Assisted Seismic Phase Picking via Visual Waveform

Interpretation. https://github.com/INGV/saian. GitHub repository (2026)

[24] Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models (2024). https://doi.org/10.48550/arXiv.2311.06607 . https://arxiv.org/abs/2311.06607

[25] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12179–12187 (2021). https://doi.org/10.1109/ICCV48922.2021.01196

[26] Salvatore, N., Wang, H., Zhang, Q.: Lost in the Middle: An Emergent Property from Information Retrieval Demands in LLMs (2025). https://doi.org/10.48550/arXiv.2510.10276 . https://arxiv.org/abs/2510.10276

[27] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR) (2023). https://doi.org/10.48550/arXiv.2210.03629

[28] Willard, B., Louf, R.: Efficient guided generation for large language models. arXiv preprint arXiv:2307.09702 (2023) https://doi.org/10.48550/arXiv.2307.09702

[29] Wan, A., Klyman, K., Kapoor, S., Maslej, N., Longpre, S., Xiong, B., Liang, P., Bommasani, R.: The 2025 foundation model transparency index. arXiv preprint arXiv:2512.10169 (2025) https://doi.org/10.48550/arXiv.2512.10169

[30] Bailo, D., Lauciani, V., Di Stefano, R., *et al.*: Researcher-curated knowledge governance for trustworthy scientific ai in the geosciences: the rockgpt approach. Earth Science Informatics **19**, 152 (2026) https://doi.org/10.1007/s12145-026-02197-5

[31] Google Gemini Team: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) https://doi.org/10.48550/arXiv.2312.11805

[32] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning (2023). https://doi.org/10.48550/arXiv.2304.08485 . https://arxiv.org/abs/2304.08485

[33] Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: PaLM-E: An Embodied Multimodal Language Model (2023). https://doi.org/10.48550/arXiv.2303.03378 . https://arxiv.org/abs/2303.03378

[34] Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications (2025). https://doi.org/10.48550/arXiv.2402.07927 . https://arxiv.org/abs/2402.07927

[35] Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J.: Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910 (2022) https://doi.org/10.48550/arXiv.2211.01910

[36] Aki, K., Richards, P.G.: Quantitative Seismology, 2nd edn. University Science Books, Sausalito, California (2001)

[37] Scherbaum, F.: Basic Concepts in Digital Signal Processing for Seismologists, 2nd edn. Lecture Notes in Earth Sciences, vol. 88. Springer, Berlin, Heidelberg (2001). https://doi.org/10.1007/BFb0017794

[38] Zamfirescu-Pereira, J.D., Wong, R.Y., Hartmann, B., Yang, Q.: Why Johnny Can't Prompt: How non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), pp. 1–21. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581388

[39] Kojima, T., *et al.*: Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems (NeurIPS) (2022). https://doi.org/10.48550/arXiv.2205.11916 . On the mechanics of zero-shot reasoning in large foundation architectures

[40] Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. In: International Conference on Machine Learning (ICML) (2021). https://doi.org/10.48550/arXiv.2102.09690

[41] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30 (2017). https://doi.org/10.48550/arXiv.1705.08500

[42] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (ICLR) (2019). https://doi.org/10.48550/arXiv.1811.12231

[43] Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.-J.: DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification (2021). https://doi.org/10.48550/arXiv.2106.02034 . https://arxiv.org/abs/2106.02034

[44] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 1321–1330 (2017). https://doi.org/10.48550/arXiv.1706.04599 .

Open Access. Foundational paper on overconfidence bias and miscalibration in deep networks

[45] Google AI for Developers: Image Understanding Specification and Vision Capabilities in Gemini API. https://ai.google.dev/gemini-api/docs/image-understanding Accessed 2026-08-21

[46] Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L.: LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale (2022). https://doi.org/10.48550/arXiv.2208.07339 . https://arxiv.org/abs/2208.07339

[47] Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., Jin, Q., He, L., Lin, X.A., Huang, F.: UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model (2023). https://doi.org/10.48550/arXiv.2310.05126 . https://arxiv.org/abs/2310.05126

[48] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022). https://doi.org/10.48550/arXiv.2203.02155 . https://arxiv.org/abs/2203.02155

[49] Zhu, W., Beroza, G.C.: PhaseNet: A deep-neural-network-based seismic arrival time picking method. arXiv preprint arXiv:1803.03211 (2018)

[50] Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L.Y., Beroza, G.C.: Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nature Communications **11**(1), 1–12 (2020)

[51] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (2021). https://doi.org/10.48550/arXiv.2106.09685 . https://arxiv.org/abs/2106.09685

[52] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day (2023). https://doi.org/10.48550/arXiv.2306.00890 . https://arxiv.org/abs/2306.00890

[53] Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Qiao, Y.: LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention (2024). https://doi.org/10.48550/arXiv.2303.16199 . https://arxiv.org/abs/2303.16199

[54] Kwon, W., Li, Z., Xie, S., Yan, R., Zheng, L., Sheng, Y., Xu, L., Joseph, A., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the 29th

Symposium on Operating Systems Principles (SOSP '23), pp. 163–171 (2023). https://doi.org/10.1145/3600006.3613165 . Defines the mechanics of KV-cache isolation and memory state partitioning in autoregressive LLM/VLM serving architectures

[55] Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: International Conference on Learning Representations (ICLR) (2024). https://doi.org/10.48550/arXiv.2309.17453 . Demonstrates how unbounded KV-cache accumulation corrupts the Softmax attention distribution in autoregressive transformers

Supplementary Files

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

- DiStefanoRAIVisualPicksSupplementary.tar.gz