Adaptive Visual Context

From CARES Research to a Verified Agent Workflow

A rigorous bridge from query-conditioned resolution research to deterministic tooling, executable scientific checks, and narrowly scoped Lean proofs.

Daniel Liezrowice

Presenter · Engineering Software Lab
AI SDLC Consultants

Research attribution

CARES — Kimhi, Shabtay, Giryes, Baskin & Schwartz (2025)
arXiv:2510.19496v3

Independent Apache-2.0 implementation · not affiliated with or endorsed by CARES authors

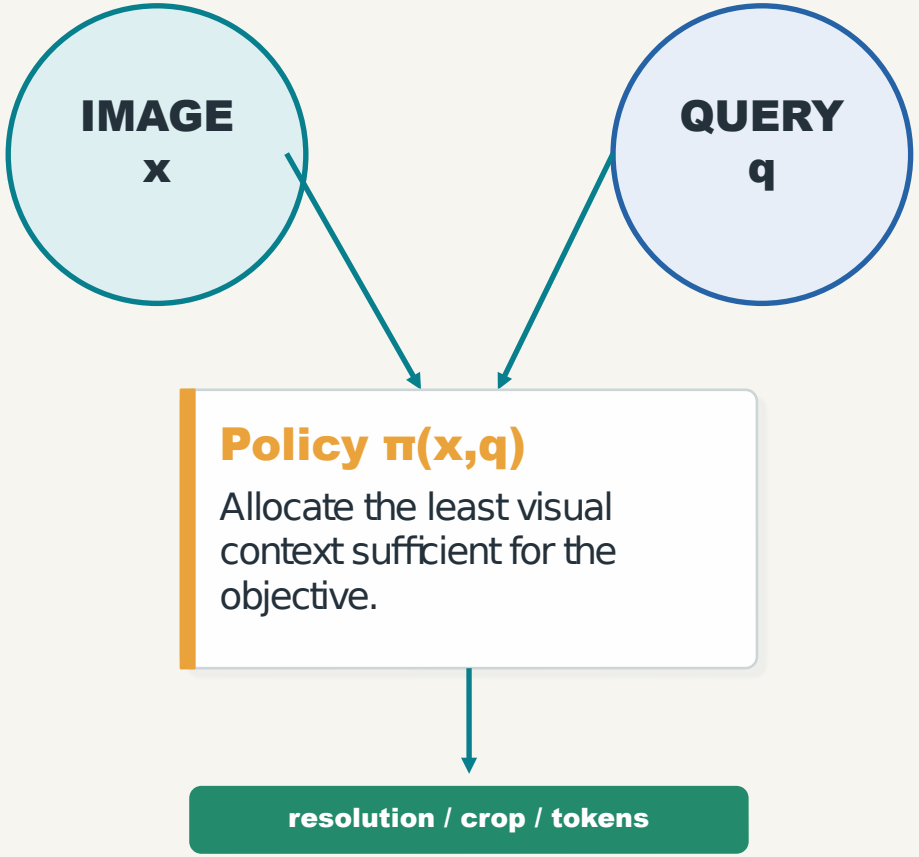
Executive scientific thesis

THESIS

$$r^* = \pi(x, q) \quad \text{— not } \pi(x) \text{ alone}$$

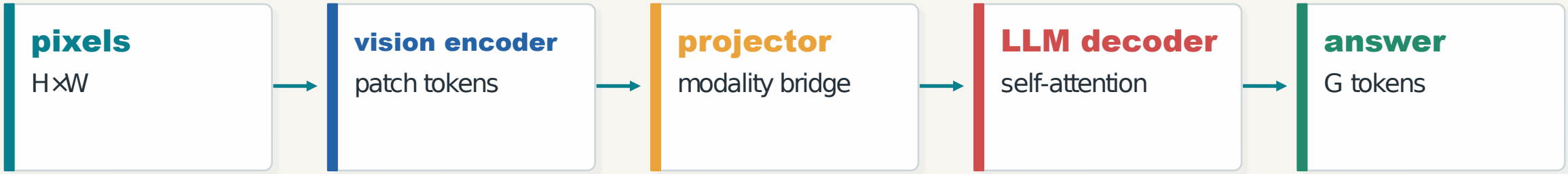
Conditioning changes allocation

The same pixels can require very different evidence: “summarize layout” may tolerate a proxy; “read CVE-2026-…” may require an exact crop. Resolution is therefore a decision conditioned jointly on image x and query q.



https://arxiv.org/abs/2510.19496v3

VLM pipeline — where visual cost enters



TOKENS

$$T = Q + V + G$$

COST

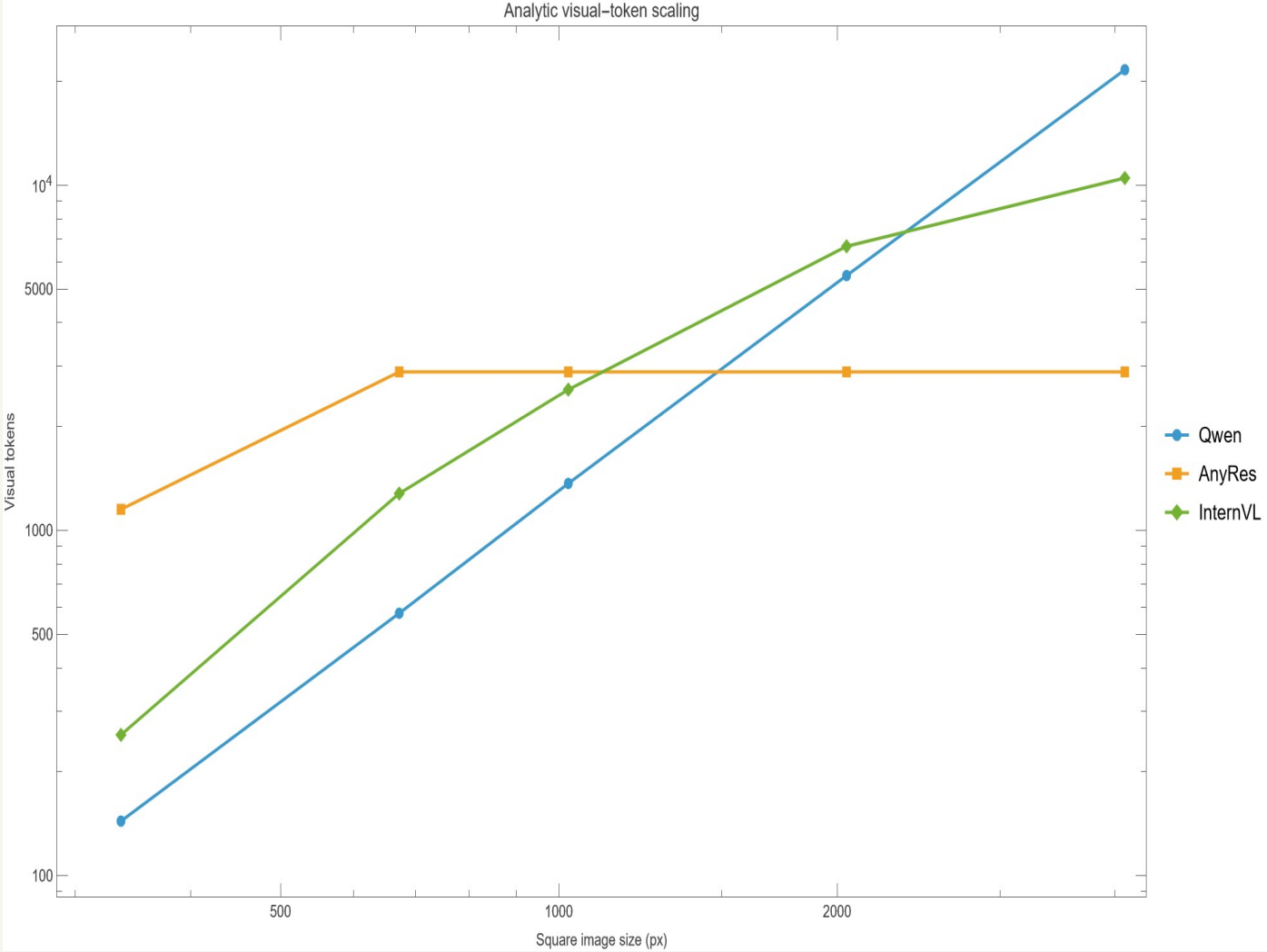
$$\Omega decoder \approx 2s^2h + 12sh^2$$

Consequence

Increasing visual sequence V increases both attention-like s^2h work and projection/MLP-like sh^2 work. Pixel count is only a proxy: tokenization and latency remain target-model specific.

https://arxiv.org/abs/2411.03312

Why visual tokens dominate



At 4096 px

Qwen: 21,609 visual tokens (99.54% with 100 text)
InternVL: 10,496 (99.06%)
AnyRes analytic cap: 2,880 (96.64%)

IDENTITY

$$\text{share\_visual} = V / (V + Q)$$

Scientific boundary

These are analytic token-accounting reproductions of the paper's Table 7 formulas—not measured target latency and not a benchmark-accuracy reproduction.

Project-generated: exports/token_scaling.csv; CARES Table 7

Capability versus allocation

AnyRes / LLaVA-NeXT

Tiling preserves local detail and handles varied aspect ratios. It expands the feasible visual budget; it does not decide whether this query needs that budget.

Native dynamic resolution

Qwen2-VL maps varying image sizes into varying token counts. Dynamic capability is not itself objective-conditioned allocation.

CARES selector

A query-aware selector predicts the smallest sufficient input range before invoking an untouched target VLM.

DISTINCTION

more available tokens ≠ more useful evidence

https://llava-vl.github.io/blog/2024-01-30-llava-next/ · https://arxiv.org/abs/2409.12191

CARES problem definition

OBJECTS

x: image q: query F: target VLM $x^{(r)}$: resized image

UTILITY

$T(r)$: target utility at input range r

GOAL

$r_s = \min \{ r \in R : \text{sufficient}(T, r) \}$

Finite support

$R = \{r_1, \ldots, r_K\}$; deployment chooses among supported target input ranges.

Target untouched

Selection is upstream; F is not retrained or modified by the CARES controller.

Minimal sufficient

The objective is not globally minimal pixels; it is the first supported range satisfying a metric-derived criterion.

https://arxiv.org/abs/2510.19496v3

Label generation — utility and sufficiency

EQ. 1

$$u_i = \mathrm{ANLS}(F(x^{(r_i)}, q), y)$$

EQ. 2

$$u_i \geq \tau \ \wedge \ \max_{j>i}(u_j - u_i) \leq \delta$$

Thresholds

Paper setup: $\tau = 0.85$ utility floor; $\delta = 0.10$ tolerance for any later improvement. Both define labels and encode metric/teacher assumptions.

Non-monotonic utility

Use all higher classes, not only the next one. Example $u = \{.86, .80, .99\}$: r_1 fails because a later gain $.13 > \delta$, despite an intermediate drop.

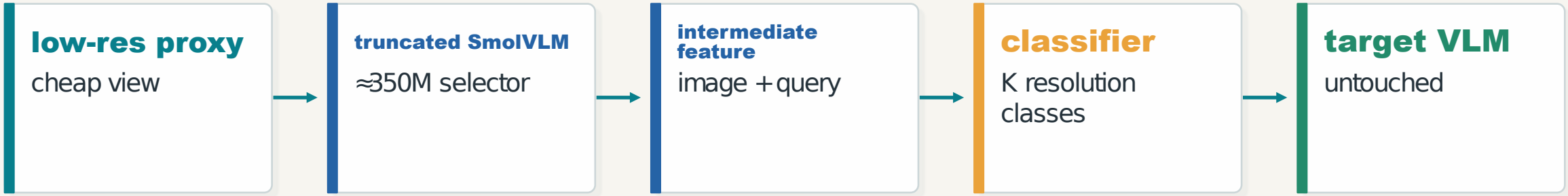
Terminal fallback

If no lower range qualifies, the highest supported class is selected. This is positional selection over finite support.

ANLS = average normalized Levenshtein similarity; labels inherit metric sensitivity.

https://arxiv.org/abs/2510.19496v3

CARES architecture



Key separation

Selector inference is paid once; routed target inference dominates the remaining cost.

Information ceiling

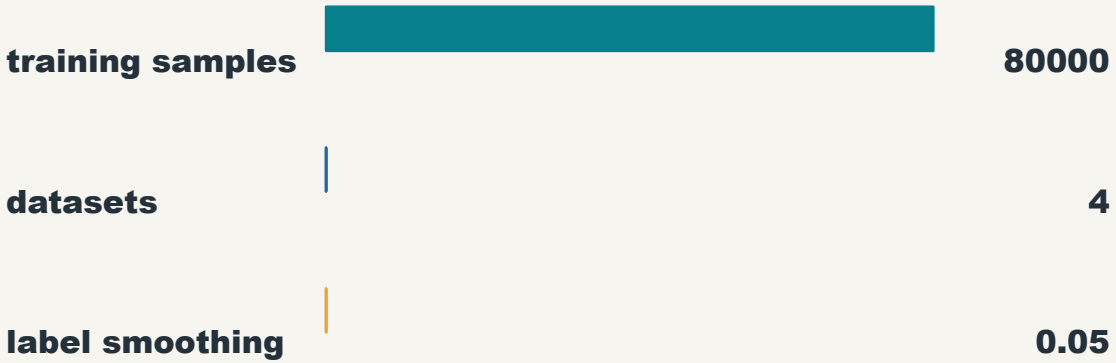
A low-resolution selector cannot reliably infer evidence completely absent from its proxy.

Compatibility

No target fine-tuning is required; predicted resolution is mapped to a supported target input.

https://arxiv.org/abs/2510.19496v3

Training protocol and variant



CE

$$L = -\Sigma_k \tilde{y}_k \log p_k$$

Base selector

Low-resolution proxy + truncated SmolVLM representation + classification head. Labels come from target-VLM utility sweeps.

AR Granite-Docling variant

The paper also reports an autoregressive variant for Granite-Docling-oriented settings; architecture choice remains task/model dependent.

Interpretation

80K / four datasets / 0.05 are reported training facts—not reproduced here.

https://arxiv.org/abs/2510.19496v3

Continuous inference and upward rounding

EQ. 3

$$p = \mathrm{softmax}(z), \quad \tilde{r} = \Sigma_k \, p_k \, r_k$$

336 × 0.10 | 672 × 0.25 | 1024 × 0.45 | 2048 × 0.20

EXAMPLE

$$\tilde{r} = 33.6 + 168 + 460.8 + 409.6 = 1{,}072$$

round upward to 2048

Why upward?

Avoid selecting a supported range below the continuous expectation.

Boundary

Rounding guarantees support membership and order—not empirical sufficiency.

https://arxiv.org/abs/2510.19496v3

Reported evidence across target models

Target	reported score	reported cost
Granite	.59 → .60	–63%
InternVL	.77 → .77	–64%
Qwen2.5-VL-72B	.79 → .80	–70%
GPT-4o	.69 → .68	–55%

Scope

Nine benchmarks · four target models (paper report).

Evidence status

REPORTED, NOT INDEPENDENTLY REPRODUCED. Score aggregation, hardware, routing overhead, and exact cost definition must be checked in an empirical replication.

DO NOT READ AS OUR BENCHMARK

https://arxiv.org/abs/2510.19496v3

Scientific interpretation and limitations

Selector overhead

Break-even requires $C_s + E[C_r] < C_h$. Small target jobs can erase routing gains.

Proxy blindness

The selector cannot see what low resolution hides; OCR and tiny-object evidence are especially exposed.

Metric dependence

ANLS labels inherit teacher, answer, threshold, and benchmark choices.

Interaction scope

Paper focus is single-image / single-turn; video, multi-page and conversation state need new policies.

Safety gap

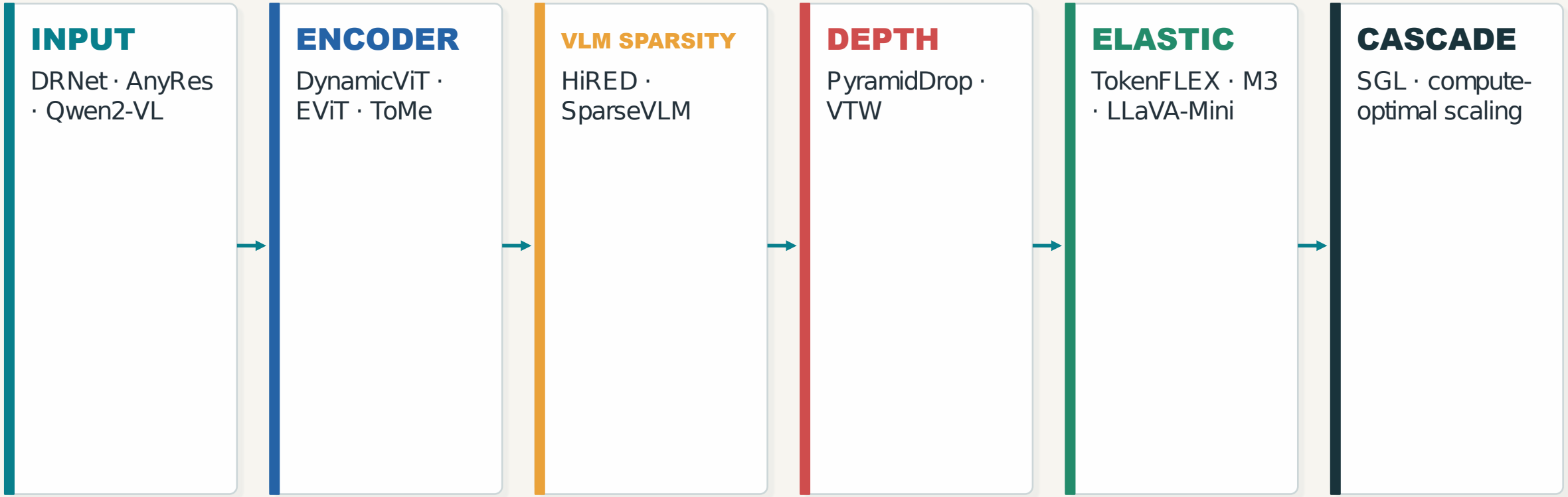
No independent safety, adversarial, calibration, or out-of-distribution study is supplied by this OSS project.

Cost proxy

Pixel or token reduction does not guarantee target-specific latency, memory, energy, or quality.

https://arxiv.org/abs/2510.19496v3

Literature map by intervention stage



MAP

allocation → pruning → merging → withdrawal → elastic training → collaboration

See reference slides 35–39

Input-resolution lineage

DRNet (2021)

Learns per-input dynamic resolution for efficient recognition. Training changes the model; query awareness is not the central VLM objective.

AnyRes / LLaVA-NeXT

Grid/tile strategy increases preserved detail and aspect-ratio support. Mechanism exposes capacity, but does not choose based on question.

Qwen2-VL

Native dynamic resolution maps arbitrary image size to variable visual-token counts. Again: capability, not necessarily allocation.

AXIS

query-aware? CARES: yes | DRNet / AnyRes / Qwen2-VL: not as the same controller objective

ViT token reduction mechanisms

Dynamic ViT

$$\text{keep ratio } \rho = |S_{keep}| / N$$

EViT

$$\text{fused token} = \Sigma_i a_i x_i / \Sigma_i a_i$$

ToMe

$$m(a,b) = (s_a a + s_\beta b)/(s_a + s_\beta)$$

DynamicViT

Predicts which tokens to retain; attention masking supports differentiable training.

EViT

Keeps attentive tokens and fuses less attentive tokens into a summary token.

ToMe

Merges similar tokens, preserving aggregate “size” weights; no token classifier is required.

Composition warning

Reducers can remove evidence selected by upstream resolution. Joint calibration—not independent optimization—is required.

VLM inference sparsification

HiRED

Allocates a visual-token budget using high-resolution encoder signals, retaining spatially informative tokens before the language model.

BUDGET

$$b_i = B \cdot w_i \,/\, \Sigma_j w_j$$

SparseVLM

Uses text-to-vision relevance and visual-token ranking to sparsify inference while seeking task-aware retention.

RELEVANCE

$$rel(v_i, q) = \max_j sim(v_i, t_j)$$

SCHEMATIC

$$rank(v_i) = rel(v_i, q) + \lambda \cdot centrality(v_i)$$

Equations here summarize controller structure; consult papers for exact definitions and implementation.

Depth-wise visual-token reduction

SCHEDULE

$$s_0 \to s_1 \to \ldots \to sL, \quad s\ell_{+1} \leq s\ell$$

DECODER

$$\Omega \approx 2s^2h + 12sh^2$$

PyramidDrop

Drops visual tokens progressively across LLM depth, exploiting the hypothesis that later layers need less raw visual detail. Savings compound because s enters both quadratic and linear terms.

Visual Tokens Withdrawal

Withdraws visual tokens at selected layers after their information has influenced textual states. Timing is a fidelity/control variable.

resolution controls input evidence → depth controls retention

Elastic trained models and controller gap

TokenFLEX

Trains operation over flexible visual-token budgets; supports runtime budget choice but requires elastic training.

Matryoshka Multimodal Models

Nested representations aim to remain useful at multiple token granularities; deployment still needs a budget controller.

LLaVA-Mini

Compression-oriented architecture reduces visual-token burden through learned modality interaction; architecture is not a query policy.

GAP

$$\text{elastic capability} + \text{controller objective} + \text{calibrated gate} = \text{adaptive system}$$

Small–large collaboration and compute optimality

COST

$$F\_inf = O(NT), \quad T = Q + V + G$$

SCALING

$$Y(N,T) = (A/N^{\alpha}) \cdot (B/T^{\beta}) + D$$

SGL: small–large collaboration

A smaller model can route, draft, or handle easier work while escalating hard cases to a larger model. Selector overhead and error correlation determine value.

Compute-optimal VLMs

Allocation across model size N and token budget T exposes a joint design space. More image tokens are not universally optimal.

OCR caveat

Tiny text is discontinuous: below a readability threshold, extra model capacity cannot recover missing characters.

Composition thesis — a generalized controller

ACTION

$$a = [r, \text{crop}, B_{enc}, \text{merge}, \text{withdrawal layer}, \text{model}, \text{escalate}]$$

CARES upstream
select evidence scale

→

encoder/projector
prune or merge

→

decoder depth
withdraw progressively

→

cascade
choose model / escalate

Control objective
Minimize expected cost subject to fidelity, safety, support, and evidence-retention constraints.

OSS project goals and evidence philosophy

Practical runtime

Portable Agent Skill + deterministic Pillow CLI. It applies an explicit keyword policy—not the trained CARES ≈350M selector.

Executable science

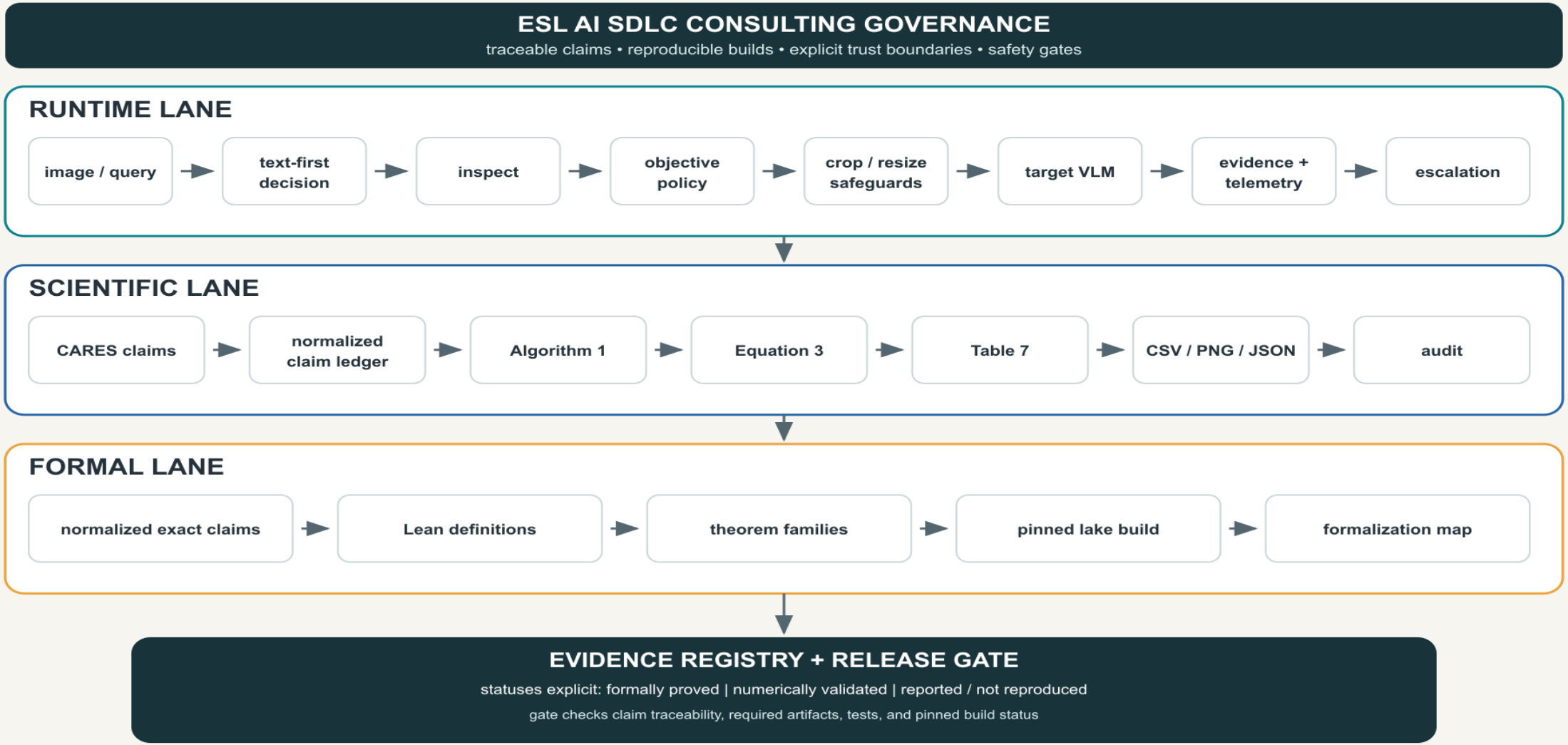
Wolfram reproduces Algorithm 1, Equation 3, Table 7 accounting, sensitivity, and break-even economics.

Narrow formalization

Lean proves selected exact statements under explicit assumptions. It does not prove the paper or empirical accuracy.

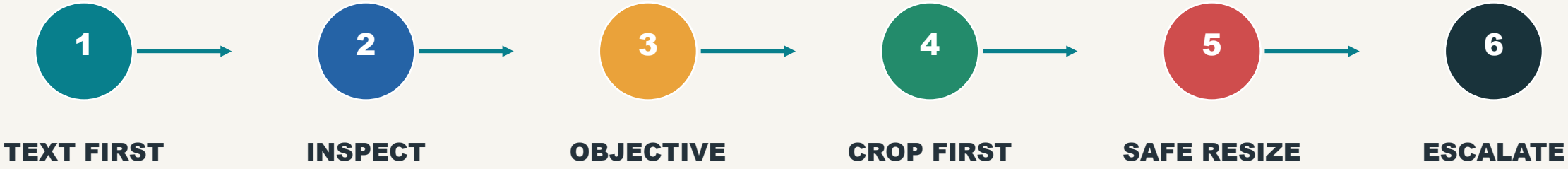
trace every claim to an evidence class

End-to-end project workflow



Original diagram: presentation/project-workflow.svg

Agent Skill runtime behavior



Text-first gate

If source text, DOM, OCR, or structured data already answers the question, avoid image inference.

Fidelity policy

overview 1280/J PEG88; detail 2048/PNG; exact and critical preserve dimensions by default.

Escalation

Warnings and consequential evidence trigger source verification or higher-fidelity processing.

Python implementation — deterministic by design

CLI / API

inspect PATH
prepare PATH --objective …
--mode auto|overview|detail|exact|critical
--crop X,Y,W,H
--max-edge N

Safeguards

EXIF transpose before coordinates; crop before resize; no upscale; no overwrite; frame count and color-mode checks.

Deterministic JSON

policy + reason; original/result dimensions; crop; format; pixel reduction; max-edge provenance; warning.

POLICY

classifier = ordered keyword boundary matching

Verification

19 focused tests reported in the repository; current QA reruns pytest.

Pixel reduction is only a proxy for target-specific visual-token and latency reduction.

Worked path — exact evidence in a report

report.png
2200×1400

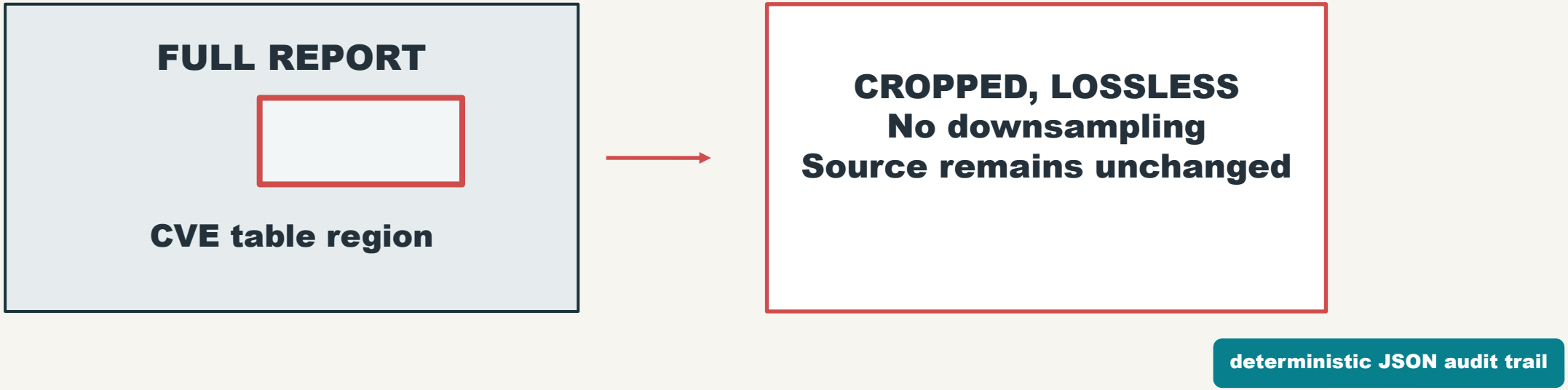
objective
"Read exact CVE + version"

policy
exact

crop
840,120,620,400

output
evidence.png

target VLM
verify source



deterministic JSON audit trail

Wolfram architecture and generated artifacts

CARESModel.wl
validated functions

→

CARESReproduction.nb
executable narrative

→

build_project.wls
deterministic build

→

proof_audit.wls
claim checks

→

exports/
CSV · PNG · JSON · reports

Package boundary

Input validation, exact selector semantics, analytic token formulas, break-even predicate.

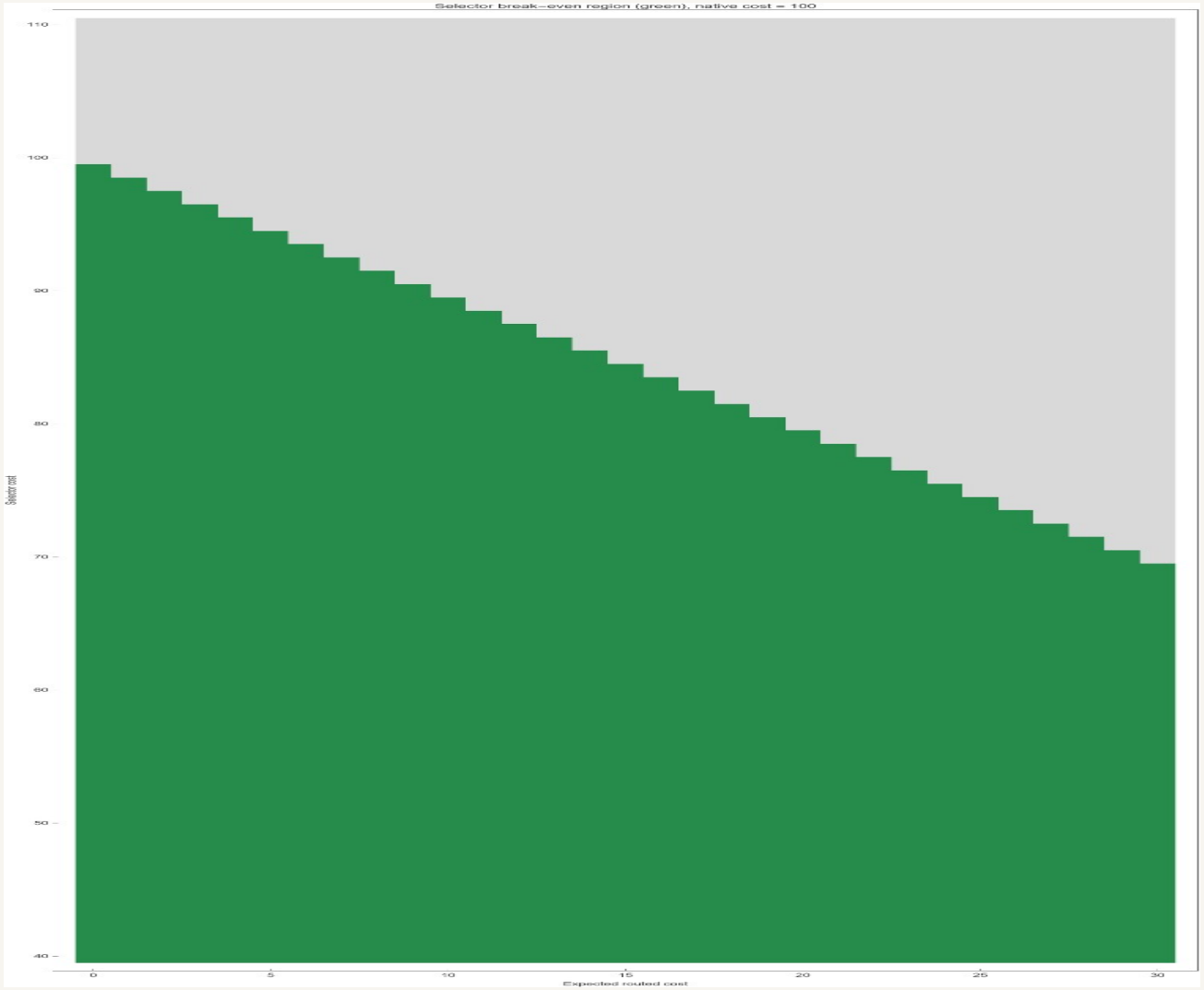
Build boundary

Seeded / deterministic data generation; required-artifact check; validation and audit reports.

Evidence boundary

10/10 numerical/analytic checks passed. Empirical benchmark performance remains out of scope.

Wolfram scientific checks



BREAK-EVEN

$$C_s + E[C_r] < C_h$$

Algorithm 1

First supported class passing τ and all-higher-gain δ; non-monotonic case included.

Equation 3

Weighted expectation, one-hot recovery, deterministic bounds.

Table 7

Hard-coded published token-count targets and percentages checked analytically.

Lean formalization — trust boundary

Definitions

Finite three-class expected resolution; upward rounding; positional sufficiency selector.

Theorem families

Bounds, one-hot identities, upward mass transfer, supported rounding, first-pass/fallback behavior.

Pinned build

Lean 4.19.0; Mathlib v4.19 revision c44e0c8…; lake build exit 0; all theorem proofs complete.

Critical boundary

Lean proves these mathematical statements under assumptions. It does NOT prove CARES empirical accuracy, datasets, implementation fidelity, compute savings, or the paper as a whole.

Lean proof sketches and assumptions

BOUNDS

$$r_1 \le \Sigma_k p_k r_k \le r_3$$

IDENTITY

$$\text{oneHot}(k) \Rightarrow E[r] = r_k$$

TRANSFER

$$\text{mass} \uparrow \Rightarrow E[r] \text{ nondecreasing}$$

Assumptions

$p_k \ge 0$; $\Sigma p_k = 1$; ordered support $r_1 \le r_2 \le r_3$. Upward transfer preserves total mass and nonnegativity.

Round-up

Returned value is supported, no lower than the continuous estimate, and no higher than the largest support value—within defined cases.

Selector semantics

Proof is positional first-pass / second-pass / highest fallback over booleans. It is not a theorem of numeric minimality over $\mathbb{R}$.

no “Lean proves the paper” claim

Evidence matrix — do not collapse categories

Claim	Formally proved	Numerically validated	Reported / not reproduced
Expectation bounds / one-hot	✓	✓	
Algorithm 1 examples		✓	paper mechanism
Table 7 token accounting		✓	published values
Benchmark accuracy / savings			✓
Python policy safeguards		19 tests	project behavior
CARES trained selector accuracy			✓

FORMAL ≠ NUMERICAL ≠ EMPIRICAL

CI, release, licensing, and provenance

CI / release gate

pytest → artifact checks → pinned lake build. Wolfram is a documented licensed-local release check. Evidence statuses remain explicit.

Licensing

This independent implementation: Apache-2.0. CARES paper: CC BY-SA 4.0. Paper attribution does not transfer software provenance.

Upstream boundary

CARES repository lacked a software license at inspected commit 1bedb45d32ac1e6ed641fdcbb32952d9d89aa96a; no upstream source was copied/adapted.

PROVENANCE

claim → source → executable check → evidence status → release decision

Repository is independent and not affiliated with or endorsed by CARES authors or institutions.

ESL — AI SDLC Consultants

ESL Engineering Software Lab

Discovery
claims · risks · constraints

Architecture
trust boundaries · controls

Governance
evidence · release gates

Reproducibility
Pinned tools, generated artifacts, auditable scientific checks.

Formal verification
Narrow theorem statements, explicit assumptions, mapped claims.

Safety + adoption
Escalation policy, fidelity gates, latency/quality/user metrics.

https://eswlab.com

Roadmap — from policy prototype to measured controller

v0.2

OCR / text extraction

frame selection

v0.3

plugin hooks

empirical harness

v0.4

learned router

calibration / OOD

v1.0

joint resolution + token controller

safety & adoption gates

SUCCESS

measure quality, latency, memory, energy, escalation, and calibration together

Conclusion — adaptive context needs evidence discipline

Daniel Liezrowice

linkedin.com/in/liezrowice

1 · Scientific thesis

Resolution is conditioned on image and query; capability alone is not allocation.

2 · Engineering thesis

Deterministic safeguards and telemetry make practical behavior auditable.

3 · Verification thesis

Formal, numerical, and reported evidence must remain separate.

VERIFY THE CLAIM · PRESERVE THE EVIDENCE

References I — CARES and model foundations

01 **CARES · Kimhi et al. · arXiv:2510.19496v3**

https://arxiv.org/abs/2510.19496v3

02 **SmolVLM · arXiv:2504.05299**

https://arxiv.org/abs/2504.05299

03 **Granite Vision · arXiv:2502.09927**

https://arxiv.org/abs/2502.09927

04 **lmms-eval · arXiv:2407.12772**

https://arxiv.org/abs/2407.12772

References II — resolution architectures

05 **Dynamic Resolution Network · arXiv:2106.02898**

https://arxiv.org/abs/2106.02898

06 **LLaVA-NeXT / AnyRes**

https://llava-vl.github.io/blog/2024-01-30-llava-next/

07 **Qwen2-VL · arXiv:2409.12191**

https://arxiv.org/abs/2409.12191

08 **DynamicViT · arXiv:2106.02034**

https://arxiv.org/abs/2106.02034

References III — token reduction and sparsity

09 **EViT · arXiv:2202.07800**
https://arxiv.org/abs/2202.07800

10 **Token Merging (ToMe) · arXiv:2210.09461**
https://arxiv.org/abs/2210.09461

11 **HiRED · arXiv:2408.10945**
https://arxiv.org/abs/2408.10945

12 **SparseVLM · arXiv:2410.04417**
https://arxiv.org/abs/2410.04417

References IV — depth and elastic models

13 **PyramidDrop · arXiv:2410.17247**

https://arxiv.org/abs/2410.17247

14 **Visual Tokens Withdrawal · arXiv:2405.05803**

https://arxiv.org/abs/2405.05803

15 **TokenFLEX · arXiv:2504.03154**

https://arxiv.org/abs/2504.03154

16 **Matryoshka Multimodal Models · arXiv:2405.17430**

https://arxiv.org/abs/2405.17430

References V — cascade, scaling, and project

17 **LLaVA-Mini · arXiv:2501.03895**
https://arxiv.org/abs/2501.03895

18 **SGL · arXiv:2412.03324**
https://arxiv.org/abs/2412.03324

19 **Compute-optimal VLMs · arXiv:2411.03312**
https://arxiv.org/abs/2411.03312

20 **Adaptive Visual Context source**
https://github.com/zuwasi/adaptive-visual-context

21 **Engineering Software Lab**
https://eswlab.com