The Reviewer-Proof Paper Playbook

A checklist-driven guide to writing research papers that close every door a reviewer might use to reject you.

Strategy across Narrative · Write-up · Literature · Diagrams · Methodology · Experiments · Visualization · Failure Analysis · Rebuttal

How to use this playbook. Reviewers rarely reject a paper for a single fatal flaw; they reject for an *accumulation* of small openings an unmotivated method, a missing baseline, a claim without evidence, an ugly figure. This playbook is organized as a defensive checklist: each box you tick removes one opening. Walk it top to bottom during writing, then again before submission. The goal is not perfection it is leaving the reviewer with **no honest reason to say no**.

1 Narrative & Framing

The paper is an argument, not a report. Every reader should be able to recover your story arc from the abstract, the introduction, and the figure captions alone.

1.1 The story arc: Problem → Gap → Insight → Method → Evidence

- ☐ State the **problem** in concrete, high-stakes terms before any method appears.
- ☐ Describe **existing approaches** fairly, then name their **specific drawback / limitation / failure mode**.
- ☐ Articulate the **gap**: what is missing or broken that nobody has fixed.
- ☐ Deliver the **insight** the one idea that makes your method work in a single sentence.
- ☐ Show the **method** as the natural consequence of that insight (not a bag of tricks).
- ☐ Close with the **evidence** preview: the headline result that proves the insight was right.
  - ☐ Run this arc at three scales: abstract (3–4 sentences), introduction (3–4 paragraphs), and the paper as a whole.
  - ☐ Make sure each transition is *causal* ("because X fails, we do Y"), not merely sequential.

REVIEWER'S LENS A reviewer who cannot answer "what problem is this and why is the prior work insufficient?" after the first page will read the rest looking for reasons to reject.

1.2 Contributions: stated, scoped, and defensible

- ☐ List contributions explicitly either as a bulleted list or as "*First, we propose the first . . . ; second, we show . . .*".
- ☐ For each contribution, give one sentence of context so the reader knows *why it is hard / novel*.
- ☐ Ensure every contribution is later **delivered and demonstrated** in the paper (no orphan claims).

- [ ] Calibrate "first ever" / "novel" claims over-claiming invites a single citation that sinks the paper.
  - [ ] Scope each claim with the conditions under which it holds (dataset, regime, assumption).
  - [ ] Map each contribution to the exact section, table, or figure that backs it.

REJECTION TRAP "We are the first to . . . " with no qualifier is the easiest claim in the world to falsify. One reviewer who knows a 2019 paper you missed can reject on novelty alone. Qualify, or downgrade to "among the first."

2 Write-up & Prose

Good ideas die in bad prose. Reviewers are tired, fast, and unforgiving of friction.

- [ ] Lead every section and paragraph with its **point first**, then support it.
- [ ] Use the **problem** → **gap** → **insight** → **method** → **evidence** pattern locally, not just globally.
- [ ] Define every acronym on first use; keep a consistent name for every concept throughout.
- [ ] Prefer active voice and concrete verbs; cut hedging ("somewhat", "arguably", "we believe").
- [ ] Keep sentences short enough to parse on one breath; one idea per sentence.
- [ ] Make the abstract self-contained: problem, approach, headline numbers, and significance.
  - [ ] Put at least one concrete quantitative result in the abstract ("+4.2 AUC, 2.1× faster").
  - [ ] Re-read the introduction last it should promise exactly what the paper delivers.
- [ ] Caption every figure and table so it is **readable standalone** (what, how, takeaway).
- [ ] Run a spell-check and a grammar pass; fix every broken `??` reference and undefined citation.
- [ ] Match notation in prose to notation in equations, figures, and algorithms.

PRO TIP Write the figure captions *before* the body text. If you can tell the whole story by reading only the captions in order, the paper's skeleton is sound.

3 Naming Sections: Headings That Land

Headings are the skeleton a reviewer navigates by. At top venues, keep them *standard enough to navigate* but *sharpened to inform*. Reserve real creativity for the paper title, figure takeaways, and analysis subsections.

3.1 Rules for headings

- [ ] Keep headings **parallel in form** within a section (all noun phrases, or all questions not a mix).
- [ ] Prefer **informative over cute**: an Area Chair must find "Experiments" and "Ablations" instantly.
- [ ] For analysis subsections, a **question or mini-claim** header works well ("What Drives the Gains?").
- [ ] Match **venue conventions**: CVPR/NeurIPS readers expect names like *Experiments*, *Ablation Study*, *Implementation Details*.
- [ ] Be **consistent**: if one sibling subsection is a question, make them all questions (or none).
- [ ] Never sacrifice findability for wordplay if a reviewer cannot locate your metrics or baselines, that is a real cost.

3.2 Heading alternatives by paper region

Standard heading	Stronger / catchier alternatives
Introduction & framing	
Introduction	Introduction; Motivation and Overview
Motivation	Motivation; Why X Matters; The Case for X
Problem Statement	Problem Setup; Problem Formulation; The X Problem
Contributions	Our Contributions; Summary of Contributions; Key Contributions
Background, notation & method	
Background	Background; Preliminaries; Foundations
Preliminaries	Preliminaries; Notation and Background; Preliminaries and Notation
Notation	Notation; Notation and Definitions
Problem Formulation	Problem Setup; Task Formulation; Problem Definition
Method / Methodology	Method; Approach; Our Approach; The Proposed Method; *Method*
Overview	Method Overview; Framework Overview; Approach at a Glance
Architecture	Model Architecture; Network Architecture
Objective / Loss	Training Objective; Learning Objective; Loss Function
Theoretical Analysis	Theoretical Analysis; Why It Works; Theoretical Justification
Literature review	
Related Work	Background and Related Work; Prior Art; Situating Our Work; The X Landscape
Current State of X	State of the Art in X; Where the Field Stands; X: Progress and Open Problems
Positioning	How We Differ; Distinction from Prior Work; Connections to Prior Work
Experiments & results	
Experiments	Experiments; Empirical Evaluation; Empirical Study
Experimental Settings	Experimental Setup; Experimental Protocol; Study Design
Dataset and Settings	Datasets and Protocols; Data and Benchmarks; Datasets and Splits
Training Details	Training Protocol; Optimization Details; Training Recipe
Details of Implementation	Implementation Details; Reproducibility Details
Evaluation Metrics	Metrics; Evaluation Protocol; What We Measure
Comparison with State of the Art	Main Results; Comparison to Prior Work; Benchmarking Against SOTA; How We Compare
Quantitative Results	Quantitative Evaluation; Quantitative Comparison
Qualitative Results	Qualitative Analysis; Visual Results; Qualitative Comparison
Ablation Studies	Ablations; Component Analysis; Dissecting *Method*; What Drives the Gains?
Hyperparameter Sensitivity	Sensitivity Analysis; Hyperparameter Analysis; Robustness to Hyperparameters
Robustness	Robustness Analysis; Robustness and Generalization; Out-of-Distribution Evaluation; Stress Tests
Efficiency / Complexity	Efficiency Analysis; Computational Cost; Runtime Analysis; Complexity Analysis
Statistical Analysis	Statistical Significance; Significance Testing
Analysis, discussion & failure	
Analysis	Analysis; In-Depth Analysis; A Closer Look
Discussion	Discussion; Analysis and Discussion; What We Learned
Case Study	Case Study; A Closer Look; Deep Dive

Standard heading	Stronger / catchier alternatives
Failure Cases	Failure Analysis; Error Analysis; When It Fails
Limitations	Limitations and Future Work; Scope and Limitations; When Does It Fail?; Failure Modes
Conclusion & back matter	
Conclusion / Concluding Remarks	Conclusion; Discussion and Conclusion; Summary and Outlook; Takeaways and Future Directions
Future Work	Future Work; Future Directions; Outlook; Open Problems
Broader Impact	Broader Impacts; Societal Impact; Ethical Considerations
Reproducibility	Reproducibility Statement; Reproducibility
Supplementary / appendix	
Appendix	Appendix; Supplementary Material; Additional Material
Additional Experiments	Further Results; Extended Results; Additional Evaluations
Proofs	Proofs; Derivations; Theoretical Proofs
Baselines	Baseline Methods; Competing Approaches; Methods Compared
Benchmarks	Benchmark Suite; Evaluation Benchmarks

X = your topic/field; Method = your method's name. List the preferred option first.

PRO TIP A subheading can carry a finding: instead of "Ablation Study," a sub-subsection titled "Attention is the primary driver" tells the reviewer your conclusion before they read the numbers. Use sparingly, and only where the claim is backed below.

REJECTION TRAP Over-clever headings ("The Plot Thickens," "Numbers Don't Lie") frustrate reviewers hunting for standard content and read as unserious at a top venue. Keep the wit for the title.

4 Word Choice & Sentence Patterns

Precise, varied, confident language signals a mature author. Upgrade weak verbs, vary your connectives, scope your claims, and lean on reusable scaffolds for the moves every paper makes.

4.1 Technical terminology upgrades

Instead of	Prefer (varied, precise)
Representations & features	
embedding(s)	latent representation, feature representation, latent code, encoded features
vision embeddings	visual representations, visual features
text embeddings	textual representations, language features
vector	representation, feature vector, latent vector
hidden state	latent state, hidden representation
feature map	activation map, feature representation
patch features	patch-level representations, instance embeddings, instance-level features
attention scores	attention weights, attention distribution
representation space	latent space, embedding space, feature space
dimension	dimensionality, feature dimension

Instead of	Prefer (varied, precise)
Models, methods & components	
model	architecture, network, framework, approach
method / our thing	approach, framework, technique, pipeline, the proposed method
algorithm	procedure, scheme, formulation
backbone	backbone, feature extractor, encoder
part	component, module, block, element
layer	layer, block, stage
head	prediction head, classification head, output head
branch	pathway, stream, branch
step	stage, phase
way	approach, strategy, mechanism, scheme
trick	technique, mechanism, strategy
setup / setting	configuration, regime, scenario
idea	insight, intuition, hypothesis, principle
Data & labels	
big dataset	large-scale dataset, extensive corpus
training data	training set, training corpus
amount of data	data volume, sample size, dataset scale
data point / example	sample, instance, observation
pictures / images	images, samples, instances; (pathology) tiles, patches, whole-slide images (WSIs)
slide	whole-slide image (WSI)
class	category, class
labels	annotations, ground-truth labels, supervision signal
ground truth	ground-truth annotations, reference labels
noisy data	noisy / corrupted samples
unseen data	held-out data, out-of-distribution (OOD) data
split	partition, fold, data split
Learning & training	
train	train, optimize, fit
learn	learn, acquire, capture
fine-tune	fine-tune, adapt
pretrain	pre-train
training process	training procedure, optimization
from scratch	from random initialization
update weights	update parameters, optimize parameters
converge	converge, stabilize
Performance & results	
good / great results	strong results, competitive performance, favorable results, state-of-the-art
works well	performs well, is effective, achieves strong performance
better / a lot better	superior, improved; substantially better, better by a large margin
worse	inferior, degraded, lower
bad performance	poor / suboptimal performance
beats	outperforms, surpasses, exceeds
results are similar	comparable, on par with, competitive with
consistent	consistent, robust across settings
gains	improvements, gains, performance boosts

Instead of	Prefer (varied, precise)
big improvement	substantial / considerable gain (reserve "significant" for statistical significance)
Efficiency & compute	
fast / runs fast	efficient, computationally efficient, low-latency, high-throughput
speed	throughput, latency, inference time, wall-clock time
memory	memory footprint, GPU memory, peak VRAM usage
model size	parameter count, number of parameters
compute	computational cost, FLOPs, GPU-hours
small model	lightweight, compact, parameter-efficient
Analysis & evidence	
shows that	demonstrates that, reveals that, indicates that, establishes that
this means	this implies, this suggests, this indicates
the reason is	we attribute this to, this is driven by
we think	we hypothesize, we posit, we conjecture
because of	attributable to, owing to, driven by
look at the effect	examine the effect, analyze the impact of
as expected	consistent with our hypothesis
Problems, gaps & contributions	
problem	challenge, limitation, bottleneck, shortcoming, open problem
drawback	limitation, shortcoming, weakness
gap	open challenge, limitation in prior work
fix / solve	address, resolve, mitigate, alleviate, tackle, overcome
hard	challenging, non-trivial, difficult
new	novel
our contribution	we propose, we introduce, we present, we develop
Quantity, degree & frequency	
a lot of	numerous, substantial, considerable, a wide range of
many	numerous, a multitude of, a wide variety of
a few / some	several, a number of, multiple
small (amount)	modest, limited, marginal
around / about	approximately, roughly, on the order of
more and more	increasingly
always	consistently, invariably (use only if literally always)
usually	typically, generally, commonly
sometimes	occasionally, in certain cases
Comparison & similarity	
compared to	relative to, in comparison with, versus
like / similar to	analogous to, akin to, comparable to
different from	distinct from, in contrast to, unlike
same as	identical to, equivalent to, on par with

4.2 Stronger verbs

Weak / plain	Stronger academic verbs
use	employ, leverage, exploit, harness, adopt, apply
utilize	use, employ, leverage (*note:* "utilize" is often just inflated "use" prefer "use" unless stressing effective use of a resource)
show	demonstrate, reveal, illustrate, establish, indicate
get / obtain	achieve, attain, yield, derive
make / build	construct, formulate, devise, develop
improve	enhance, boost, advance, refine, strengthen
combine	integrate, fuse, unify, aggregate, couple
look at / study	investigate, examine, analyze, explore
find	observe, identify, discover
check / test	verify, validate, assess, evaluate, benchmark
fix (a problem)	address, resolve, mitigate, alleviate, overcome
help	facilitate, enable, support, aid
need	require, necessitate, demand
based on	built upon, grounded in, derived from, inspired by

4.3 Transitions & connectives (by function)

Vary these so paragraphs do not all start with the same word; pick the connective whose *logical* meaning is exact.

Function	Connectives
Addition	moreover, furthermore, in addition, additionally, beyond this, likewise
Contrast	however, nevertheless, nonetheless, yet, whereas, in contrast, conversely, by comparison
Cause	because, since, as, owing to, due to, given that
Result	therefore, thus, hence, consequently, as a result, accordingly, it follows that
Sequence	first / firstly, subsequently, next, then, finally, ultimately
Emphasis	notably, importantly, crucially, in particular, indeed, remarkably
Example	for instance, for example (e.g.), specifically, namely, to illustrate
Concession	although, while, despite, even though, granted that
Restatement	in other words, that is (i.e.), put differently
Conclusion	in summary, overall, taken together, to summarize, in conclusion

4.4 Scope your claims (hedging & precision)

Avoid / risky	Prefer (defensible)
"proves"	suggests, indicates, provides evidence that (reserve "prove" for mathematics)
"always" / "never"	typically, generally, in most cases (unless literally always)
"significantly" (casual)	substantially, considerably ("significant" implies a statistical test)
"the best method"	achieves the best results on *{benchmarks}* (scope it)
"very" / "really"	delete; state the precise magnitude instead
"clearly" / "obviously"	delete; let the evidence carry the claim

4.5 Reusable sentence scaffolds

Templates for the moves every paper makes fill the slots and keep the rhythm.

- ☐ **Motivation:** *"While {prior work} has achieved {X}, it remains limited by {Y}."*
- ☐ **Gap:** *"However, existing approaches fail to / overlook / cannot {Z}."*
- ☐ **Insight:** *"Our key insight is that {...}."*
- ☐ **Contribution:** *"To this end, we propose {Method}, a {one-line description}."*
- ☐ **Method specificity:** *"Concretely / Formally / Specifically, {...}."*
- ☐ **Design rationale:** *"We adopt {choice} because {reason}; we validate this in Sec. {ref}."*
- ☐ **Result highlight:** *"Notably, {Method} achieves {number}, improving over {baseline} by {Δ}."*
- ☐ **Analysis:** *"We attribute this to {...}, which suggests that {...}."*
- ☐ **Address a limitation:** *"To address this limitation, we {...}."*
- ☐ **Ablation framing:** *"To isolate the contribution of {component}, we {...}."*

REJECTION TRAP Upgraded vocabulary is a tool, not a costume. Do not have every sentence "leverage a novel framework to harness" something over-stuffed prose reads as AI-generated or insecure. Vary word choice, and when in doubt choose the **precise** word over the fancy one.

5 Literature Review

A literature review is a map of the field that ends with an X marking where you stand.

- ☐ Organize related work into **thematic subsections by category**, not a flat citation dump.
- ☐ Make the subsections **cohesive**: each should advance a single sub-narrative of the field.
- ☐ Have the subsections **add up** to a picture of the current state of the art and its frontier.
- ☐ End the section by positioning your work against this frontier the gap you fill.
- ☐ Include a **comparison / summary table** of related methods.
  - ☐ Columns = the dimensions that matter (assumptions, inputs, supervision, scalability, what they lack).
  - ☐ Add a final row for *your* method so the contrast is immediate and visual.
  - ☐ Use checkmarks / crosses or short tags so a reviewer reads it in five seconds.
- ☐ Where possible, add a **visual literature review**: a figure laying out the architectural components and strategies competing methods use.
  - ☐ Group methods by family; show shared vs. distinguishing components side by side.
  - ☐ Make your method's novel block visually pop against the shared backbone.
- ☐ Cite the obvious recent and seminal works a missing key citation reads as carelessness.
- ☐ Represent rival work **accurately and charitably**; authors are often your reviewers.

REVIEWER'S LENS Reviewers feel ownership of the literature. A fair, well-structured related-work section signals you understand the field and earns goodwill for the rest of the paper.

6 Diagrams & Visual Communication

Most reviewers form an opinion from the figures before reading a word.

6.1 The hook diagram (Figure 1)

- ☐ Build an eye-catching **hook diagram** that shows off your key result on first glance.

- Make the performance gap **visually striking** your bar/curve clearly dominates.
- Plot the right axes: accuracy vs. cost, quality vs. speed, performance vs. parameters.
- Label the takeaway directly on the figure ("Ours", arrows, Δ callouts).

- Optionally add a second hook showing **how your approach is fundamentally different** in design.
  - Contrast the prior paradigm and your paradigm side by side (before / after).
  - Highlight the one structural change that drives the gains.

6.2 Method & architecture figures

- Provide a clean architecture / pipeline figure that mirrors the methodology's flow exactly.
- Use a consistent visual language: one colour per data type, one shape per operation type.
- Show tensor shapes / dimensions on the data flow where it aids understanding.
- Ensure every block in the figure maps to a named component in the text.

6.3 Figure craft (applies to all figures)

- Use a colour-blind-safe palette and keep colours consistent across the whole paper.
- Make fonts in figures match (or exceed) the body text size when printed.
- Export as vector (PDF/SVG); never ship blurry raster screenshots.
- Keep one clear message per figure; split overloaded figures.
- Verify legibility in grayscale (some reviewers print).

REJECTION TRAP A cramped, low-resolution, or unlabeled Figure 1 primes the reviewer to expect sloppy work everywhere. Spend disproportionate effort on the first figure.

7 Methodology

Rigor here pre-empts the "not technically sound" rejection.

- Break the method into **many subsections** so the flow is explicit and easy to follow.
- **Establish notation and definitions early**, before they are used.
- Keep **symbols and notation consistent** everywhere (text, equations, figures, algorithms).
- Be **mathematically rigorous**: define objects, state assumptions, write the objective precisely.
- Present core procedures as numbered **algorithm blocks** with clear inputs/outputs.
- Include **PyTorch-style pseudocode** for the key module so it is implementation-ready.
  - Make the pseudocode runnable-looking: real tensor ops, shapes in comments.
  - Keep it short enough to read but complete enough to reproduce.
- **Rationalize every design decision** say *why* this choice, not just *what*.
  - Tie each choice back to the insight from the introduction.
  - Foreshadow the ablation that will later validate the choice.
- State complexity (time / memory) of the method where relevant.
- Make reproducibility explicit: hyperparameters, initialization, hardware, and seeds.

REVIEWER'S LENS "Why this and not the obvious alternative?" is the question a methods reviewer asks of every design choice. Answer it pre-emptively or it becomes a weakness in the review.

8 Starting the Methodology: Problem Formulation & Notation

The methodology section's opening pair — the *problem formulation* and the *notation* — determines whether the rest of the section reads as rigorous or hand-wavy. Get them right and every subsequent equation refers to objects the reviewer already accepts. Get them wrong and the methods reviewer flags an "imprecise problem statement" on page two.

Treat these as two separate jobs: the problem formulation states *what task you are solving and under what assumptions*; the notation states *the vocabulary you will use to describe it.* Confusing the two is the most common opening failure.

8.1 The problem formulation subsection

A problem formulation answers four questions, in this order: **(1)** what are the inputs and outputs? **(2)** what supervision is available? **(3)** what is the learning objective? **(4)** what assumptions distinguish this setting from related ones? Anything else — backbones, architectures, training tricks — belongs elsewhere.

- ☐ Open with one sentence naming the **task** and the **setting** ("*We consider few-shot weakly-supervised whole-slide image classification.*"). Reviewers should know within ten words what they're reading.
- ☐ State the **inputs** as a typed mathematical object: *a whole-slide image $\mathcal{S}$, represented as a bag of N patch features* $\mathbf{X} = \{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^D$.
- ☐ State the **outputs** the same way: *a class label* $y \in \mathcal{Y} = \{1, \ldots, C\}$.
- ☐ Specify the **supervision regime** explicitly: weakly-supervised (slide-level labels only), few-shot (k labeled slides per class), training-free, or otherwise.
- ☐ Write the **learning objective** as a single equation: $\arg\min_\theta \mathbb{E}_{(\mathcal{S},y)} \mathcal{L}(f_\theta(\mathcal{S}), y)$. Name the loss $\mathcal{L}$, the model f_θ, the parameter set θ, and the data distribution.
- ☐ List the **assumptions** that distinguish your setting from the obvious neighbors — standard MIL, fully-supervised, multi-task, continual — in one short paragraph. This is where you preempt the "not actually few-shot" or "not actually weakly-supervised" objection.
- ☐ Reference the formulation throughout the paper as "the FSWC setting (Sec. 3.1)" — give it a name so reviewers can talk about it.

REVIEWER'S LENS A methods reviewer reads the problem formulation looking for two things: *is this task well-defined?* and *is this task different from {related setting they know}?* If either is unclear by the end of the subsection, the rest of the method is read with suspicion.

8.2 Worked example: a clean problem formulation (FSWC)

REVIEWER'S LENS **Anti-example: vague and undertyped.**
"We consider few-shot WSI classification. Each slide has many patches and a label. The goal is to learn from few examples. We use a model with parameters θ that maps slides to labels."

No types, no objective, no distinguishing assumptions, "few" undefined. A reviewer cannot tell this apart from standard supervised classification.

PRO TIP **Effective version.**

We consider ***few-shot weakly-supervised whole-slide image classification (FSWC).*** *Each whole-slide image $\mathcal{S}$ is represented as a multi-scale bag of pre-extracted patch features*

$$\mathbf{X} = \{ (x_i^{(s)}, c_i^{(s)}) \}_{i=1,\, s\in\mathcal{R}}, \quad x_i^{(s)} \in \mathbb{R}^D,\ c_i^{(s)} \in \mathbb{R}^2,$$

where $\mathcal{R}$ is the set of magnifications (e.g. $\{5\times, 20\times\}$), $x_i^{(s)}$ is the i-th patch's frozen encoder feature at scale s, and $c_i^{(s)}$ is its spatial coordinate. Each slide carries a single label $y \in \mathcal{Y} = \{1, \ldots, C\}$ (weak supervision: no patch-level annotation). We are given a k-shot training set $\mathcal{D}_{\text{tr}} = \{(\mathbf{X}_n, y_n)\}_{n=1}^{kC}$ containing exactly k slides per class (in our experiments $k \in \{1, 4, 16\}$), and seek a model f_θ minimizing

$$\theta^* = \arg\min_\theta \mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}_{\text{tr}}} \mathcal{L}\big(f_\theta(\mathbf{X}),\, y\big), \tag{1}$$

evaluated on a disjoint held-out set $\mathcal{D}_{\text{te}}$ drawn from the same distribution. Two assumptions distinguish FSWC from related settings: (i) slide labels but no patch labels (vs. fully-supervised MIL), and (ii) a fixed small k throughout training (vs. continual or low-shot fine-tuning regimes).

This version (a) types every object, (b) makes the supervision explicit, (c) writes the objective as a labeled equation that later sections can reference, and (d) lists exactly the two assumptions that distinguish FSWC from the nearest settings. The methods reviewer now has no opening.

8.3 The notation subsection

Notation is the vocabulary the rest of the methodology — and every equation, figure, and pseudocode listing — will speak. A reviewer who can't map the notation in your equations to the notation in your figures stops trusting both.

- ☐ Place notation **immediately after the problem formulation,** before the architecture or training procedure. The reader needs the vocabulary before the verbs.
- ☐ Introduce each symbol **once,** with its type ($\mathbf{W} \in \mathbb{R}^{C\times D}$) and one phrase of intuition (*"the alignment projection"*).
- ☐ **One symbol, one meaning.** If θ is the parameter set, never reuse it for a temperature later.
- ☐ Pick a **convention** and apply it throughout: matrices bold uppercase, vectors bold lowercase, scalars italic, sets calligraphic, indices roman. Stating the convention once saves the reader from re-deducing it per equation.
- ☐ Distinguish **free vs. indexed** symbols: x alone vs. x_i (one patch), $x_i^{(s)}$ (one patch at one scale).
- ☐ Match the notation in equations to the notation in figures, pseudocode, and tables. The fastest reviewer-credibility loss in the methodology is a figure that says "Features" where the equation says $\mathbf{X}$.
- ☐ Defer notation you only need in the appendix — a methodology section dense with symbols you use twice reads as overengineered.

8.4 Worked example: a notation block as a compact table

For a paper with more than ∼8 symbols, a notation table is easier to scan than a paragraph. Reviewers refer back to it without losing their place.

Symbol	Type	Meaning
$\mathcal{S}$	—	a whole-slide image
N	$\mathbb{N}$	number of patches in a bag (variable per slide)
D	$\mathbb{N}$	patch feature dimension (e.g. 512 for CONCH)
$\mathcal{R}$	set	magnifications used, e.g. $\{5\times, 20\times\}$
$x_i^{(s)}$	$\mathbb{R}^D$	feature of patch i at scale s
$\mathbf{X}$	$\bigcup_s \mathbb{R}^{N_s \times D}$	multi-scale bag of features
y	$\{1, \ldots, C\}$	slide-level class label
k	$\mathbb{N}$	shots per class in the few-shot regime
f_θ	$\mathbf{X} \mapsto \mathbb{R}^C$	model with parameter set θ
$\mathcal{L}$	loss	training objective (cross-entropy by default)
$\mathbf{P} \in \mathbb{R}^{C \times D}$	matrix	text-prototype matrix, one row per class
τ	$\mathbb{R}_{>0}$	alignment temperature

Bold uppercase: matrices. Bold lowercase: vectors. Italic: scalars. Calligraphic: sets / spaces.

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{8pt}
\begin{tabular}{l l l}
\toprule
\textbf{Symbol} & \textbf{Type} & \textbf{Meaning}\\
\midrule
$\mathcal{S}$            & ---                              & a whole-slide image\\
$N$                      & $\mathbb{N}$                     & number of patches in a bag (variable per slide)\\
$D$                      & $\mathbb{N}$                     & patch feature dimension\\
$\mathcal{R}$            & set                              & magnifications used, e.g.\ $\{5\times,20\times\}$
    \\
$x_i^{(s)}$              & $\mathbb{R}^D$                   & feature of patch $i$ at scale $s$\\
$\mathbf{X}$             & $\bigcup_s \mathbb{R}^{N_s\times D}$ & multi-scale bag of features\\
$y$                      & $\{1,\ldots,C\}$                 & slide-level class label\\
$k$                      & $\mathbb{N}$                     & shots per class\\
$f_\theta$               & $\mathbf{X}\!\mapsto\!\mathbb{R}^C$ & model with parameters $\theta$\\
$\mathcal{L}$            & loss                             & training objective\\
$\mathbf{P}\in\mathbb{R}^{C\times D}$ & matrix                & text-prototype matrix\\
$\tau$                   & $\mathbb{R}_{>0}$                & alignment temperature\\
\bottomrule
\end{tabular}
\end{center}
```

8.5 Universal rules for the methodology opening

- ☐ **Problem before method.** The first time a reviewer reads about your architecture should be *after* they accept the problem.
- ☐ **Types before behavior.** Every symbol in your method has a type; introduce the type before the verb.
- ☐ **Constraints up front.** "Frozen encoder," "no patch labels," "single-GPU memory limit" — state these in the formulation, not as fine print in §5.
- ☐ **Name the setting.** A named formulation ("the FSWC setting," "the IL-WSI regime") becomes a citable handle for your paper.
- ☐ **Number the objective.** Eq. (1) should be referenceable from later sections (loss decompositions, ablations, theory).
- ☐ **Refer back, often.** "Recall from Eq. (1) that…" costs nothing and binds the paper to its own framing.

REJECTION TRAP Two failure modes are nearly universal in junior writing. The first is *notational churn*: introducing $\mathbf{X}$ in §3.1, then silently switching to F in the architecture figure, then $\mathbf{H}$ in the loss. The second is the *implicit problem*: launching into the architecture without ever writing the objective, so the reader has to reverse-engineer what task is being solved. Both are

caught by reading the methodology aloud with a pen, asking "what is this symbol?" at every encounter.

PRO TIP A useful sanity check: after writing the problem formulation and notation, give *only those two subsections* to a colleague and ask them to state, in their own words, (i) what task you are solving, (ii) what they are given as input, and (iii) what they need to output. If their answers don't match yours within two sentences, the subsection is not yet doing its job — no amount of method description downstream will fix that.

9 Writing PyTorch-Style Pseudocode

A short, readable listing of your core module does more for reproducibility and trust than a page of prose. Top-tier papers (SimCLR, MoCo, ViT) made their key idea legible in ∼20 lines. Aim for that.

9.1 What good pseudocode looks like

- ☐ Show the **key contribution**, not the whole training loop one forward pass of the novel module.
- ☐ Use **real, recognizable ops** (`einsum`, `softmax`, `matmul`) so it reads as runnable.
- ☐ Annotate **tensor shapes in comments** at every important line (`# [B, N, D]`).
- ☐ Keep variable names consistent with the **notation in your equations** ($Q, K, V \rightarrow$ `q,k,v`).
- ☐ Strip boilerplate (imports, device handling, logging) unless it is the point.
- ☐ Caption it like a figure and **reference it from the methodology text**.
- ☐ Keep it to roughly 15–30 lines if longer, factor out helpers or move detail to the appendix.

9.2 Template

Below is the canonical NeurIPS/CVPR "numpy-/torch-like" style: a self-contained function, shapes in comments, names matching the math.

Listing 1: Core module as PyTorch-style pseudocode. Shapes track the notation in Eq. (1)–(3). B batch, N tokens/instances, D dim.

```
def forward(x, mask=None):
    # x: bag of instance features            [B, N, D]
    q = Wq(x)                                 # query   [B, N, D]
    k = Wk(x)                                 # key     [B, N, D]
    v = Wv(x)                                 # value   [B, N, D]

    # scaled dot-product attention over instances
    attn = einsum('bnd,bmd->bnm', q, k) / D**0.5   # [B, N, N]
    if mask is not None:
        attn = attn.masked_fill(mask, -1e9)
    attn = softmax(attn, dim=-1)              # row-normalized [B, N, N]

    ctx = einsum('bnm,bmd->bnd', attn, v)     # context [B, N, D]
    z = LayerNorm(x + ctx)                    # residual + norm [B, N, D]

    # task-aware gated aggregation -> bag embedding
    gate = sigmoid(Wg(z))                     # per-instance weight [B, N, 1]
    bag = (gate * z).sum(dim=1)               # bag vector [B, D]
    return bag, attn                          # return attn for visualization
```

PRO TIP Return your attention / weight maps from the forward function (as above). You will reuse them for the qualitative figures and the failure analysis one tensor, three sections of evidence.

- ☐ Decide format: `lstlisting` (Python-styled, as above) or `algorithm2e`/`algpseudocodex` for math-heavy procedures.
- ☐ Use a code listing when the contribution is a *module*; use an algorithm block when it is a *procedure* with loops and conditionals.
- ☐ Make the listing color-light and grayscale-safe (subtle background, dark keywords).
- ☐ Verify the pseudocode actually matches your released code reviewers do check.

10 Reading Maps & Cross-References

A mature paper is *navigable*: a reviewer flipping through it can always answer "where am I, and where is the thing I want?" Two small disciplines do almost all of this work — a one-paragraph *reading map* at the start of each major section, and dense *cross-references* throughout. Both are pure editorial craft. They cost almost no words, can be added in a final pass, and are among the clearest signals of authorial care.

10.1 The reading map

A reading map is a short paragraph — typically one to three sentences — placed immediately after a section's opening claim, that tells the reader the structure of what follows. It is the single most efficient way to convert a section's table of contents into an *argument* the reviewer can hold in their head.

- ☐ Place the reading map **at the start of every multi-subsection section**: Methodology, Experiments, Related Work, and any analysis section with $\geq$3 subsections.
- ☐ Name each subsection **by what it establishes,** not by its number. "Sec. 3.1 formalizes the problem" beats "First, we present Sec. 3.1."
- ☐ Make the map **causal** where possible ("*we first define X; this lets us state Y in Sec. 3.2*"), not merely sequential.
- ☐ Keep the map **short**. If it runs longer than three sentences, the section probably has too many subsections.
- ☐ Use a non-breaking space (`~`) between "Sec."/"Fig."/"Tab." and the number, so the reference never wraps onto two lines.

10.2 Worked example: a methodology reading map

REVIEWER'S LENS **Anti-example: no map.**
"In this section we present our method. We first describe how patch features are encoded. Then we describe the aggregator. Then we describe the prompt engine. Finally we describe training."
This is a table of contents pretending to be a paragraph. It tells the reviewer what is coming but not *why* those pieces are in that order, and gives no hook for the central claim.

PRO TIP **Effective version.**
This section formalizes few-shot weakly-supervised WSI classification and develops ***FSWC-Bench****'s two-slot architecture. We first state the problem and notation (Sec. 3.1). Sec. 3.2*

then introduces the instance-to-bag aggregator slot*, the design axis that prior work most often varies; Sec. 3.3 introduces the* prompt engine slot*, which we find to be the larger driver of few-shot performance (Sec. 5.1). The training objective and inference procedure follow in Sec. 3.4.*

The effective version (a) tells the reviewer where the contribution lives in the structure, (b) plants the headline finding so the rest of the section reads in its service, and (c) keeps the navigation typographically consistent. The whole thing is four sentences.

10.3 Reading maps for other sections

- ☐ **Experiments.** "*Sec. 5.1 reports main results on three datasets; Sec. 5.2 ablates each slot in isolation; Sec. 5.3 reports compute cost; Sec. 5.4 analyzes failure modes.*"
- ☐ **Related Work.** "*We organize related work along the two contested slots: Sec. 2.1 surveys instance-to-bag aggregation, and Sec. 2.2 surveys prompt engineering for vision-language MIL. Sec. 2.3 positions our framework as the first to vary both axes under one pipeline.*"
- ☐ **Ablations.** "*We isolate each design axis in turn: encoder (Sec. 5.2.1), aggregator family (Sec. 5.2.2), prompt source (Sec. 5.2.3), and token-handling strategy (Sec. 5.2.4). Sec. 5.2.5 summarizes which axes dominate.*"

10.4 Cross-references: rules

A cross-reference is a pointer from one part of the paper to another. Mature papers point at themselves constantly — the paper feels *woven* rather than *written*. The discipline is to make every non-trivial claim *locally verifiable*: alongside the claim sits a pointer to the table, figure, equation, or section that backs it.

- ☐ **Every claim should carry a pointer.** "Our method improves over the prior best (Tab. 1)." Not "Our method improves over the prior best."
- ☐ **Forward and backward both work.** Forward ("*we will show in Sec. 5.1 that…*") primes the reader; backward ("*recall from Eq. (1) that…*") binds new content to established framing.
- ☐ **Be consistent in abbreviation.** Pick one of "Sec. 3.2," "Section 3.2," or "§3.2" and use it everywhere. Mixing forms within a paper is a junior tell.
- ☐ **Non-breaking space** between "Sec."/"Fig."/"Tab." and the number (`Sec.~3.2`) so a reference never wraps onto two lines.
- ☐ **Be specific.** "See Tab. 1, row 4" beats "see Table 1."
- ☐ **Don't reference forward into the supplementary** from the introduction; reviewers read the main paper first.

10.5 Worked example: dense cross-referencing

REVIEWER'S LENS **Anti-example: orphan claims.**

"Our method outperforms prior work on three pathology benchmarks. We use a text-guided prototype decoder, which is more effective than attention alone. The improvement is largest at low shot counts. Even when the encoder is fixed, our method wins."

Four claims, zero pointers. The reviewer must hunt for the evidence behind each one — and a reviewer who has to hunt loses charitable reading.

PRO TIP **Effective version.**
Our method outperforms prior work on three pathology benchmarks (Tab. 1). We attribute this to the text-guided prototype decoder introduced in Sec. 3.2, which Tab. 3 shows to be more effective than attention alone (+2.8 macro-F1). The improvement is largest at low shot counts (Fig. 4, $k = 1, 4$). The gain holds under the fairness control of Sec. 5.3, in which all methods use the same frozen CONCH encoder.

Same four claims, four pointers, fifteen extra words. Every claim is locally auditable; the paragraph reads as a guided tour of the evidence rather than a list of assertions.

10.6 Cross-reference patterns by location

Different parts of the paper take different reference loads. Use this as a rough budget:

- ☐ **Abstract:** no cross-references (the abstract must stand alone).
- ☐ **Introduction:** one forward reference per contribution, pointing at the section that delivers it ("*We introduce a unified framework (Sec. 3) and benchmark 13 methods under it (Sec. 5.1).*"). This is what tells the reviewer the paper actually delivers what the intro promises.
- ☐ **Methodology:** backward references to the problem formulation and notation; forward references to the experiments that validate each design choice.
- ☐ **Experiments:** reference back to the methodology for every result ("*validating the design of Sec. 3.2*"). One pointer per result paragraph.
- ☐ **Ablations:** forward to the supplementary for the exhaustive grid, backward to the design decision the ablation tests.
- ☐ **Conclusion:** backward references to the main result table and the limitations section. No new claims.

10.7 Universal rules

- ☐ **Open every long section with a reading map.** It costs three sentences and reframes the section as an argument.
- ☐ **Cross-reference defensively.** If a reviewer could ask "where is the evidence?" point at it preemptively.
- ☐ **Use one reference style consistently** throughout the paper — "Sec. 3" or "Section 3," not both.
- ☐ **Compile twice** before checking references. LaTeX needs two passes to resolve them; a stray "Sec. ??" in your final PDF is a credibility leak that costs you nothing to fix.

REVIEWER'S LENS On a final read, scan a printed copy of your paper holding a pencil. Every time you read a claim, draw a small arrow to the figure, table, or section that backs it. Sentences with no arrow are orphan claims — either they need a pointer, or they shouldn't be there. A mature paper has an arrow on nearly every sentence of the introduction and the experiments.

REJECTION TRAP The opposite failure is cross-reference *overload*: a sentence that points to three figures and two tables reads as defensive and breaks the flow. Cross-reference *specific* claims, not generic ones. "*Our method works* (Tab. 1, Tab. 3, Fig. 4, Fig. 5)" is worse than the unreferenced version — pick the single strongest pointer.

11 Results & Experiments

This is where reviewers look hardest for a reason to reject. Be comprehensive.

11.1 Baselines & coverage

- ☐ Choose **strong, fair baselines**, including the current state of the art.
- ☐ **Justify why each baseline was chosen** (representative of a family, prior SOTA, ablated variant).
- ☐ Tune baselines as carefully as your own method a weak baseline is a credibility hole.
- ☐ Evaluate on **at least 3–4 datasets** spanning relevant settings.
- ☐ Report **at least 2–3 metrics** per dataset (capture different facets of performance).

11.2 Profiling & efficiency

- ☐ Report **wall-clock time** for training and/or inference.
- ☐ Report **GPU-hours** and the hardware used.
- ☐ Report **FLOPs** and **model size / memory footprint**.
  - ☐ Present efficiency vs. accuracy as a trade-off (table or Pareto plot) so gains are honest.

11.3 Ablations & sensitivity

- ☐ Include **ablation studies** that remove each component to show its contribution.
- ☐ Include **hyperparameter sensitivity analysis** over the key knobs.
  - ☐ Show the method is not brittle to reasonable hyperparameter choices.
  - ☐ Connect each ablation back to the design decision it validates.

11.4 Robustness & generalization

- ☐ Run **out-of-domain / distribution-shift** evaluation.
- ☐ Add other relevant robustness tests (noise, perturbations, low-data, cross-dataset transfer).
- ☐ Ensure every **robustness claim is backed by a specific analysis** no unsupported adjectives.

11.5 Statistical rigor

- ☐ Report results over **multiple seeds / runs** with mean and variance / confidence intervals.
- ☐ Perform **statistical significance testing** for headline comparisons (and name the test).
- ☐ Do not claim an improvement that lies within the noise band.

REJECTION TRAP A single-run number beating a baseline by a hair, with no error bars and no significance test, is the most common "insufficient evidence" rejection. Variance and a p-value cost little and close the gap.

12 Main Results Tables: Multi-Dataset, Encoder, Shot & Prompt

A benchmarking paper lives or dies by its main table. The reader should grasp "who wins, by how much, and is it fair?" in one scan. Choose the layout that matches the axis you vary, and make fairness visually obvious.

12.1 Universal layout rules

- ☐ Choose orientation by what you have more of: **many methods → methods as rows**; many settings → settings as rows.
- ☐ Group related columns with `\multicolumn` headers and `\cmidrule` underlines (one metric per sub-column).
- ☐ Always add an **Average** column, and an **Average-Rank** column when spanning many datasets.
- ☐ **Bold the best** and underline the second best *per column*; keep decimals consistent.
- ☐ Report **std as a subscript** (87.1±0.2) to stay compact.
- ☐ **Hold the backbone/encoder fixed** across methods for fairness and state it in the caption.
- ☐ Separate baseline families with `\midrule` and a small italic group label; shade the **Ours** row.
- ☐ Never leave a cell blank use " " or "OOM" so the reader knows it is intentional.

12.2 Style A Multi-dataset, grouped metrics

Methods as rows; each dataset is a `multicolumn` block of its metrics. The workhorse main-results table.

Method	CAMELYON16		TCGA-NSCLC		TCGA-RCC		Avg.
	AUC	F1	AUC	F1	AUC	F1	
Classic MIL							
ABMIL	90.1±0.6	84.2	94.0±0.4	88.1	97.1±0.3	91.0	90.8
DSMIL	91.4±0.5	85.0	94.6±0.3	88.9	97.4±0.2	91.6	91.5
TransMIL	92.8±0.5	86.7	95.1±0.3	89.4	97.6±0.2	92.1	92.3
Ours	**94.0**±0.4	**88.1**	**95.9**±0.3	**90.6**	**98.2**±0.2	**93.0**	**93.3**

All methods use the same frozen UNI encoder and identical 5-fold splits. Bold: best; underline: second.

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{5pt}
\begin{tabular}{l cc cc cc c}
\toprule
\multirow{2}{*}{\textbf{Method}}
  & \multicolumn{2}{c}{\textbf{CAMELYON16}}
  & \multicolumn{2}{c}{\textbf{TCGA-NSCLC}}
  & \multicolumn{2}{c}{\textbf{TCGA-RCC}}
  & \multirow{2}{*}{\textbf{Avg.}}\\
\cmidrule(lr){2-3}\cmidrule(lr){4-5}\cmidrule(lr){6-7}
  & AUC & F1 & AUC & F1 & AUC & F1 & \\
\midrule
\multicolumn{8}{l}{\textit{\scriptsize Classic MIL}}\\
ABMIL    & $90.1{\scriptstyle\pm0.6}$ & $84.2$ & $94.0{\scriptstyle\pm0.4}$ & $88.1$ & $97.1{\scriptstyle\pm0
     .3}$ & $91.0$ & $90.8$ \\
DSMIL    & $91.4{\scriptstyle\pm0.5}$ & $85.0$ & $94.6{\scriptstyle\pm0.3}$ & $88.9$ & $97.4{\scriptstyle\pm0
     .2}$ & $91.6$ & $91.5$ \\
TransMIL & $\underline{92.8}{\scriptstyle\pm0.5}$ & $\underline{86.7}$ & $\underline{95.1}{\scriptstyle\pm0
     .3}$ & $\underline{89.4}$ & $97.6{\scriptstyle\pm0.2}$ & $\underline{92.1}$ & $\underline{92.3}$ \\
\midrule
\rowcolor{soft}
\textbf{Ours} & $\mathbf{94.0}{\scriptstyle\pm0.4}$ & $\mathbf{88.1}$ & $\mathbf{95.9}{\scriptstyle\pm0.3}$ &
      $\mathbf{90.6}$ & $\mathbf{98.2}{\scriptstyle\pm0.2}$ & $\mathbf{93.0}$ & $\mathbf{93.3}$ \\
\bottomrule
\end{tabular}
\end{center}
% needs: booktabs, multirow, colortbl (\rowcolor); define color "soft"
```

12.3 Style B Multi-encoder benchmark

Your method (or each method) held fixed while the **frozen foundation-model encoder** is swapped. Add Average and Average-Rank to summarize across encoders.

MIL method	CTransPath	UNI	CONCH	Virchow	GigaPath	Avg.	Rank
ABMIL	86.0	90.1	89.4	90.8	91.2	89.5	3.8
DSMIL	86.9	91.4	90.0	91.3	91.9	90.3	2.6
TransMIL	87.6	92.8	91.1	92.0	92.6	91.2	1.8
Ours	**88.9**	**94.0**	**92.3**	**93.1**	**93.7**	**92.4**	**1.0**

Balanced accuracy (%) on TCGA-NSCLC; encoders frozen, only the MIL head trained. Rank = mean rank across the five encoders.

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{6pt}
\begin{tabular}{l ccccc cc}
\toprule
\textbf{MIL method} & \textbf{CTransPath} & \textbf{UNI} & \textbf{CONCH}
  & \textbf{Virchow} & \textbf{GigaPath} & \textbf{Avg.} & \textbf{Rank}\\
\midrule
ABMIL    & $86.0$ & $90.1$ & $89.4$ & $90.8$ & $91.2$ & $89.5$ & $3.8$\\
DSMIL    & $86.9$ & $91.4$ & $90.0$ & $91.3$ & $91.9$ & $90.3$ & $2.6$\\
TransMIL & $87.6$ & $\underline{92.8}$ & $\underline{91.1}$ & $\underline{92.0}$ & $\underline{92.6}$ & $\
    underline{91.2}$ & $\underline{1.8}$\\
\midrule
\rowcolor{soft}
\textbf{Ours} & $\mathbf{88.9}$ & $\mathbf{94.0}$ & $\mathbf{92.3}$ & $\mathbf{93.1}$ & $\mathbf{93.7}$ & $\
    mathbf{92.4}$ & $\mathbf{1.0}$\\
\bottomrule
\end{tabular}
\end{center}
```

PRO TIP **Average rank** is more robust than average score when datasets or encoders differ in scale or difficulty a method that is consistently 2nd can lose on mean score to one that is occasionally 1st and often last. Report both.

12.4 Style C Few-shot / label-efficiency

Settings (shots or label fractions) as columns, grouped by dataset with `multirow`. Shows the trend, not just one operating point.

Dataset	Method	1-shot	4-shot	8-shot	16-shot	Full
CAMELYON16	TransMIL	61.2	72.4	79.0	84.1	92.8
	Ours	**66.8**	**77.1**	**82.6**	**86.9**	**94.0**
TCGA-NSCLC	TransMIL	70.3	80.1	85.5	89.0	95.1
	Ours	**74.9**	**83.4**	**87.8**	**90.7**	**95.9**

AUC (%) vs. shots per class. Gains are largest in the low-shot regime state this in the text.

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{6pt}
\begin{tabular}{l l ccccc}
\toprule
\textbf{Dataset} & \textbf{Method} & \textbf{1-shot} & \textbf{4-shot}
  & \textbf{8-shot} & \textbf{16-shot} & \textbf{Full}\\
\midrule
\multirow{2}{*}{CAMELYON16}
 & TransMIL & $61.2$ & $72.4$ & $79.0$ & $84.1$ & $92.8$\\
```

```
 & \cellcolor{soft}\textbf{Ours} & \cellcolor{soft}$\mathbf{66.8}$ & \cellcolor{soft}$\mathbf{77.1}$ & \
    cellcolor{soft}$\mathbf{82.6}$ & \cellcolor{soft}$\mathbf{86.9}$ & \cellcolor{soft}$\mathbf{94.0}$\\
\midrule
\multirow{2}{*}{TCGA-NSCLC}
 & TransMIL & $70.3$ & $80.1$ & $85.5$ & $89.0$ & $95.1$\\
 & \cellcolor{soft}\textbf{Ours} & \cellcolor{soft}$\mathbf{74.9}$ & \cellcolor{soft}$\mathbf{83.4}$ & \
    cellcolor{soft}$\mathbf{87.8}$ & \cellcolor{soft}$\mathbf{90.7}$ & \cellcolor{soft}$\mathbf{95.9}$\\
\bottomrule
\end{tabular}
\end{center}
% \cellcolor shades individual cells; here it tints the whole "Ours" row
```

- [ ] For label-efficiency, also **plot** metric vs. shots a curve where your line sits above the baseline is the strongest visual.
- [ ] Report the **full-data** column too, so reviewers see you do not trade asymptotic performance for low-shot gains.

12.5 Style D Prompting / inference strategy (VLMs)

Strategies as rows, datasets/metrics as columns. Build it as a cumulative ladder so the contribution of each strategy is visible.

Prompting strategy	PathVQA	PathMMU	Quilt-VQA	Avg.
Zero-shot	42.1	38.5	44.0	41.5
+ Few-shot (k=4)	46.3	41.9	47.8	45.3
+ Chain-of-thought	48.0	43.6	49.1	46.9
+ Retrieval-augmented (**Ours**)	**52.7**	**47.2**	**53.0**	**51.0**

Accuracy (%); same backbone VLM throughout, only the prompting/inference strategy varies.

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{6pt}
\begin{tabular}{l ccc c}
\toprule
\textbf{Prompting strategy} & \textbf{PathVQA} & \textbf{PathMMU}
  & \textbf{Quilt-VQA} & \textbf{Avg.}\\
\midrule
Zero-shot                       & $42.1$ & $38.5$ & $44.0$ & $41.5$\\
\;+ Few-shot ($k{=}4$)          & $46.3$ & $41.9$ & $47.8$ & $45.3$\\
\;+ Chain-of-thought            & $48.0$ & $43.6$ & $49.1$ & $46.9$\\
\rowcolor{soft}
\;+ Retrieval-augmented (\textbf{Ours}) & $\mathbf{52.7}$ & $\mathbf{47.2}$ & $\mathbf{53.0}$ & $\mathbf
    {51.0}$\\
\bottomrule
\end{tabular}
\end{center}
```

12.6 Style E Heatmap grid & win/tie/loss summary

- [ ] For large grids (encoder × method, or dataset × setting), **shade each cell by value** (via `cellcolor`) so trends pop without reading every number.
- [ ] When you span many datasets, add a one-line **win/tie/loss** or "#best" summary against the strongest baseline (e.g. "Ours wins on 9/11 datasets").
- [ ] Keep the heatmap colormap **monotonic and grayscale-safe**; put the exact numbers in the cells anyway.

REVIEWER'S LENS The first thing a benchmarking reviewer checks is *fairness*: same encoder, same splits, same metric, comparable parameter budget. A table whose caption and column

headers make fairness obvious disarms the "unfair comparison" critique before it is written.

REJECTION TRAP Bolding your own row when a baseline is actually higher, or comparing your method on a strong encoder against baselines on a weaker one, is caught instantly and poisons trust in every other table. Bold strictly by the numbers; equalize the backbone.

13 Ablation Studies: Formats That Convince

"We did an ablation" is not enough. The *format* determines whether a reviewer believes each component earns its place. Pick the layout that answers the question you are asking, report variance, and bold the winner.

13.1 Universal rules for every ablation table

- ☐ State the **question** the table answers in its caption ("Does component X help?").
- ☐ Report **mean ± std** over multiple seeds, not a single run.
- ☐ **Bold the best** value per column; optionally underline the second best.
- ☐ Show the **full model in one clearly marked row** (shaded "Ours") for instant comparison.
- ☐ Add a **Δ column** (gain over baseline or over the previous row) so the contribution is explicit.
- ☐ Change **one factor at a time**; hold everything else fixed and say so.
- ☐ Use ✓/✗ or filled/empty cells for present/absent components readable in seconds.

13.2 Format 1 Cumulative (additive) ablation

Best when components are meant to *stack*. Start from a plain baseline and add one component per row, showing monotone gains that justify the full design.

Configuration	Attn.	Gate	Aux loss	Acc. (%)	Δ
Baseline (mean-pool)	✗	✗	✗	81.2±0.4	
+ Attention	✓	✗	✗	84.0±0.3	+2.8
+ Gated aggregation	✓	✓	✗	85.6±0.3	+1.6
+ Auxiliary loss (**Ours**)	✓	✓	✓	**87.1**±0.2	**+1.5**

LaTeX SOURCE

```
\begin{center}\small
\begin{tabular}{l ccc c c}
\toprule
\textbf{Configuration} & \textbf{Attn.} & \textbf{Gate} & \textbf{Aux loss}
  & \textbf{Acc.\ (\%)} & \textbf{$\Delta$} \\
\midrule
Baseline (mean-pool) & \xmark & \xmark & \xmark & $81.2{\scriptstyle\pm0.4}$ &   \\
\;+ Attention        & \cmark & \xmark & \xmark & $84.0{\scriptstyle\pm0.3}$ & +2.8 \\
\;+ Gated aggregation& \cmark & \cmark & \xmark & $85.6{\scriptstyle\pm0.3}$ & +1.6 \\
\rowcolor{soft}
\;+ Auxiliary loss (\textbf{Ours}) & \cmark & \cmark & \cmark & $\mathbf{87.1}{\scriptstyle\pm0.2}$ & \textbf
    {+1.5} \\
\bottomrule
\end{tabular}
\end{center}
% \cmark / \xmark are user macros: \newcommand{\cmark}{\textcolor{teal}{\ding{51}}}
%                                  \newcommand{\xmark}{\textcolor{red}{\ding{55}}}  (needs pifont)
```

13.3 Format 2 Leave-one-out ablation

Best when you want to prove *each* component is necessary in the full system. Start from the full model and remove one component per row; every drop is evidence.

Variant	Dataset A	Dataset B	Dataset C
Full model (**Ours**)	**87.1**±0.2	**79.4**±0.5	**83.0**±0.3
– Attention	84.3±0.3	76.1±0.6	80.2±0.4
– Gated aggregation	85.8±0.3	77.9±0.5	81.6±0.3
– Auxiliary loss	85.9±0.2	78.2±0.4	81.9±0.3

LaTeX SOURCE

```
\begin{center}\small
\begin{tabular}{l c c c}
\toprule
\textbf{Variant} & \textbf{Dataset A} & \textbf{Dataset B} & \textbf{Dataset C} \\
\midrule
\rowcolor{soft}
Full model (\textbf{Ours}) & $\mathbf{87.1}{\scriptstyle\pm0.2}$ & $\mathbf{79.4}{\scriptstyle\pm0.5}$ & $\
    mathbf{83.0}{\scriptstyle\pm0.3}$ \\
\;$-$ Attention         & $84.3{\scriptstyle\pm0.3}$ & $76.1{\scriptstyle\pm0.6}$ & $80.2{\scriptstyle\pm0.4}
    $ \\
\;$-$ Gated aggregation & $85.8{\scriptstyle\pm0.3}$ & $77.9{\scriptstyle\pm0.5}$ & $81.6{\scriptstyle\pm0.3}
    $ \\
\;$-$ Auxiliary loss    & $85.9{\scriptstyle\pm0.2}$ & $78.2{\scriptstyle\pm0.4}$ & $81.9{\scriptstyle\pm0.3}
    $ \\
\bottomrule
\end{tabular}
\end{center}
```

13.4 Format 3 Design-choice / hyperparameter grid

Best for "which variant of a component is right?" (e.g. aggregation type, depth, fusion strategy). One block per design axis; bold the choice you ship.

Design axis	Choice	Acc. (%)	Params (M)
Aggregation	Mean	84.0±0.3	12.1
	Max	83.1±0.4	12.1
	Gated (ours)	**87.1**±0.2	12.4
# Layers	1	85.2±0.3	8.0
	2 (ours)	**87.1**±0.2	12.4
	4	86.8±0.4	20.9

LaTeX SOURCE

```
\begin{center}\small
\begin{tabular}{l l c c}
\toprule
\textbf{Design axis} & \textbf{Choice} & \textbf{Acc.\ (\%)} & \textbf{Params (M)} \\
\midrule
\multirow{3}{*}{Aggregation}
  & Mean                  & $84.0{\scriptstyle\pm0.3}$ & 12.1 \\
  & Max                   & $83.1{\scriptstyle\pm0.4}$ & 12.1 \\
  & \textbf{Gated (ours)} & $\mathbf{87.1}{\scriptstyle\pm0.2}$ & 12.4 \\
\midrule
\multirow{3}{*}{\# Layers}
  & 1                & $85.2{\scriptstyle\pm0.3}$ & \phantom{0}8.0 \\
  & \textbf{2 (ours)} & $\mathbf{87.1}{\scriptstyle\pm0.2}$ & 12.4 \\
  & 4                & $86.8{\scriptstyle\pm0.4}$ & 20.9 \\
\bottomrule
\end{tabular}
\end{center}
```

REVIEWER'S LENS A reviewer reads an ablation to answer one question: *"would the paper survive without this part?"* If a removed component barely changes the number, expect the reviewer to ask why it is a contribution. Either show a real drop, or reframe it as a minor design detail.

- ☐ Pick the format (cumulative / leave-one-out / grid) that matches your claim.
- ☐ Run the key ablation across **multiple datasets** so the conclusion is not dataset-specific.
- ☐ For sensitivity, plot the metric vs. the hyperparameter (a curve reads faster than a table) and mark your chosen value.
- ☐ Tie every ablation row back to a **design decision you claimed in the methodology**.
- ☐ Put exhaustive ablations in the supplementary; keep the headline ones in the main paper.

14 More Results Table Styles

Five additional layouts for specific reviewer questions: per-class fairness, statistical significance, efficiency, cross-dataset transfer, and a compact vs.-SOTA table. All reuse the macros already defined (✓, ✗, `soft`, booktabs, multirow, colortbl). Style I additionally needs `\usepackage{diagbox}`.

14.1 Style F Per-class breakdown

Use when class imbalance matters and a reviewer wants to confirm you don't only win on the majority class. One row per class, plus macro/weighted averages.

Class	Precision	Recall	F1	Support
LUAD	91.2	89.4	90.3	412
LUSC	88.7	90.9	89.8	388
Macro avg	90.0	90.2	90.1	800
Weighted avg	**90.0**	**90.1**	**90.1**	800

*Per-class results for **Ours** on TCGA-NSCLC (CONCH, 16-shot). Support = test slides per class.*

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{6pt}
\begin{tabular}{l cccc}
\toprule
\textbf{Class} & \textbf{Precision} & \textbf{Recall} & \textbf{F1} & \textbf{Support} \\
\midrule
LUAD          & $91.2$ & $89.4$ & $90.3$ & 412 \\
LUSC          & $88.7$ & $90.9$ & $89.8$ & 388 \\
\midrule
Macro avg     & $90.0$ & $90.2$ & $90.1$ & 800 \\
\rowcolor{soft}
Weighted avg  & $\mathbf{90.0}$ & $\mathbf{90.1}$ & $\mathbf{90.1}$ & 800 \\
\bottomrule
\end{tabular}
\end{center}
```

14.2 Style G Significance-annotated comparison

Mean±std with markers showing where your gain over the best baseline is statistically significant. Directly answers the "is this real?" reviewer.

Method	CAMELYON16	TCGA-NSCLC	TCGA-RCC
ABMIL	90.1±0.6	94.0±0.4	97.1±0.3
TransMIL	92.8±0.5	95.1±0.3	97.6±0.2
Ours	**94.0**±0.4*	**95.9**±0.3*	**98.2**±0.2†

*AUC (%), 5 seeds. */†: improvement over the best baseline is significant at $p < 0.01$/$p < 0.05$ (paired Wilcoxon signed-rank).*

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{6pt}
\begin{tabular}{l ccc}
\toprule
\textbf{Method} & \textbf{CAMELYON16} & \textbf{TCGA-NSCLC} & \textbf{TCGA-RCC} \\
\midrule
ABMIL    & $90.1{\scriptstyle\pm0.6}$ & $94.0{\scriptstyle\pm0.4}$ & $97.1{\scriptstyle\pm0.3}$ \\
TransMIL & $92.8{\scriptstyle\pm0.5}$ & $95.1{\scriptstyle\pm0.3}$ & $97.6{\scriptstyle\pm0.2}$ \\
\rowcolor{soft}
\textbf{Ours} & $\mathbf{94.0}{\scriptstyle\pm0.4}^{\ast}$ & $\mathbf{95.9}{\scriptstyle\pm0.3}^{\ast}$ & $\
    mathbf{98.2}{\scriptstyle\pm0.2}^{\dagger}$ \\
\bottomrule
\end{tabular}
\end{center}
% ^\ast / ^\dagger mark significance vs. best baseline; name the test in the caption.
```

14.3 Style H Efficiency / cost table

Params, FLOPs, memory, throughput, and a headline metric side by side the profiling numbers the Results section asks for.

Method	Params (M)	FLOPs (G)	Mem (GB)	Slides/s	AUC
TransMIL	2.7	5.1	7.8	12.4	92.8
DSMIL	0.5	1.2	3.1	**41.0**	91.4
Ours	**0.4**	**1.0**	**2.9**	38.6	**94.0**

Inference cost at 20×, one H100, CONCH features. Best per column in bold; Ours matches the lightest baseline while topping AUC.

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{6pt}
\begin{tabular}{l rrrr c}
\toprule
\textbf{Method} & \textbf{Params (M)} & \textbf{FLOPs (G)} & \textbf{Mem (GB)} & \textbf{Slides/s} & \textbf{
    AUC} \\
\midrule
TransMIL      & $2.7$ & $5.1$ & $7.8$ & $12.4$ & $92.8$ \\
DSMIL         & $0.5$ & $1.2$ & $3.1$ & $\mathbf{41.0}$ & $91.4$ \\
\rowcolor{soft}
\textbf{Ours} & $\mathbf{0.4}$ & $\mathbf{1.0}$ & $\mathbf{2.9}$ & $38.6$ & $\mathbf{94.0}$ \\
\bottomrule
\end{tabular}
\end{center}
% numeric columns right-aligned (r) so decimals line up
```

14.4 Style I Cross-dataset / OOD transfer matrix

Train cohort as rows, test cohort as columns. The shaded diagonal is in-domain; off-diagonal cells are your robustness / generalization evidence.

Train \ Test	TCGA	CPTAC	In-house
TCGA	95.9	91.2	88.4
CPTAC	90.7	94.1	87.0
In-house	89.1	88.6	92.3

*AUC (%) for **Ours** trained on the row cohort, tested on the column cohort (NSCLC, CONCH). Shaded = in-domain; off-diagonal = cross-cohort generalization.*

LaTeX SOURCE

```
% requires: \usepackage{diagbox}  (omit width= so the diagonal never crosses the labels)
\begin{center}\footnotesize
\setlength{\tabcolsep}{8pt}
\begin{tabular}{l ccc}
\toprule
\diagbox{\textbf{Train}}{\textbf{Test}} & \textbf{TCGA} & \textbf{CPTAC} & \textbf{In-house} \\
\midrule
TCGA      & \cellcolor{soft}$95.9$ & $91.2$ & $88.4$ \\
CPTAC     & $90.7$ & \cellcolor{soft}$94.1$ & $87.0$ \\
In-house  & $89.1$ & $88.6$ & \cellcolor{soft}$92.3$ \\
\bottomrule
\end{tabular}
\end{center}
```

14.5 Style J Compact vs.-SOTA with relative gain

A tight main-paper table that foregrounds the improvement: venue, headline metric, and Δ over the previous best.

Method	Venue	AUC	Δ
ViLa-MIL	CVPR'24	92.8	
FOCUS	CVPR'25	93.4	+0.6
Ours	—	**94.0**	**+0.6**

TCGA-NSCLC AUC (%), 16-shot, CONCH. Δ = gain over the previous best.

LaTeX SOURCE

```
\begin{center}\footnotesize
\setlength{\tabcolsep}{7pt}
\begin{tabular}{l l c c}
\toprule
\textbf{Method} & \textbf{Venue} & \textbf{AUC} & \textbf{$\Delta$} \\
\midrule
ViLa-MIL   & CVPR'24    & $92.8$ &    \\
FOCUS      & CVPR'25    & $93.4$ & \textcolor{tealx}{+0.6} \\
\rowcolor{soft}
\textbf{Ours} & ---      & $\mathbf{94.0}$ & \textcolor{tealx}{\textbf{+0.6}} \\
\bottomrule
\end{tabular}
\end{center}
```

15 Defining a Custom Evaluation Metric

Sometimes the standard metrics (accuracy, AUC, F1) do not capture the specific behavior your method is designed to improve. A well-motivated *custom metric* can become a contribution in its own right but only if it is defined precisely, justified, and shown to correlate with something the field already trusts. A metric that measures only your own method's pet quantity, with no link to downstream performance, is a rejection trap.

15.1 Rules for introducing a metric

- ☐ State **what the metric measures** and **why existing metrics miss it** before the formula.
- ☐ Give a **precise definition**: a formula with every symbol defined, or a numbered procedure.
- ☐ Specify the **range, direction** (higher/lower is better), and any **degenerate cases**.
- ☐ **Validate** it: show it correlates with a trusted metric (downstream accuracy/F1), so it is not measuring noise.
- ☐ Report it **alongside** standard metrics, never instead of them.
- ☐ Name it memorably and use the name consistently (it is easier for reviewers to cite a named quantity).

REJECTION TRAP A custom metric on which only your method scores well, with no demonstrated link to any standard metric, reads as "measuring the thing we happened to optimize." Always show the correlation with an accepted metric that is what turns a self-serving number into evidence.

15.2 Define it in a boxed definition

Set the metric apart so a reviewer can find and check it instantly.

Definition: Hit Ratio@K (hierarchical alignment)

For each low-scale (parent) patch p, retrieve its top-K most similar low-scale texts. A *hit* is recorded if the most similar high-scale (child) text of a linked high-scale patch matches any child text of those retrieved parents. Over all N low-scale patches in the test WSIs,

$$\text{Hit Ratio@}K \;=\; \frac{1}{N}\sum_{p=1}^{N}\mathbb{1}\!\left[\text{hit}(p,K)\right]\in[0,1].$$

Higher is better; it measures how well parent–child textual semantics are preserved across scales, which accuracy alone does not capture.

LaTeX SOURCE

```
\begin{tcolorbox}[enhanced,colback=soft,colframe=accentdk,boxrule=0.5pt,arc=2pt,
  left=10pt,right=10pt,top=7pt,bottom=7pt,
  title={\sffamily\bfseries\small Definition: Hit Ratio@$K$ (hierarchical alignment)}]
For each low-scale (parent) patch $p$, retrieve its top-$K$ most similar low-scale
texts. A \emph{hit} is recorded if the most similar high-scale (child) text of a
linked high-scale patch matches any child text of those retrieved parents. Over all
$N$ low-scale patches in the test WSIs,
\[
\text{Hit Ratio@}K \;=\; \frac{1}{N}\sum_{p=1}^{N}\mathbb{1}\!\left[\text{hit}(p,K)\right]\in[0,1].
\]
Higher is better; it measures how well parent--child textual semantics are preserved
across scales, which accuracy alone does not capture.
\end{tcolorbox}
```

15.3 Validate it: correlate the custom metric with a trusted one

The single most convincing way to justify a metric is a scatter of your metric (x) against a standard metric (y). A clear positive trend says "optimizing this improves what you already care about."

LaTeX SOURCE

```
% needs: \usepackage{pgfplots} \pgfplotsset{compat=1.18}
\begin{center}
\begin{tikzpicture}
```

```
\begin{axis}[
  width=7.2cm,height=5.4cm,
  xlabel={\footnotesize Hit Ratio@2 (\%)},ylabel={\footnotesize Macro F1 (\%)},
  xmin=60,xmax=95,ymin=68,ymax=88,
  tick label style={font=\scriptsize},
  title={\footnotesize\bfseries PLIP encoder, 16-shot},
]
\addplot[only marks,mark=*,mark size=2.5pt]      coordinates {(83,86)(90,80)(87,75)}; % Bi.
\addplot[only marks,mark=square*,mark size=2.3pt] coordinates {(77,85)(82,78)(80,73)}; % Uni.
\addplot[only marks,mark=x,mark size=3pt,thick]   coordinates {(64,86)(65,77)(74,72)}; % No.
\legend{\scriptsize Bi.,\scriptsize Uni.,\scriptsize No.}
\end{axis}
\end{tikzpicture}
\end{center}
```

15.4 Report it alongside standard metrics

Variant	Hit Ratio@2 (%)	Macro F1 (%)
No interaction	67.7	78.3
Unidirectional	79.7	78.7
Bidirectional (**Ours**)	**86.7**	**80.3**

The custom metric (Hit Ratio@2) and the standard metric (Macro F1) move together; the gain on the custom metric is the larger, isolating the mechanism.

15.5 More custom-metric ideas (and what each defends)

- ☐ **Hit Ratio@K** — semantic/retrieval consistency; defends a cross-scale or retrieval-alignment claim. Correlate with F1.
- ☐ **Attention Concordance** — overlap (IoU / rank correlation) between the model's high-attention patches and pathologist annotations; defends interpretability.
- ☐ **Few-Shot Efficiency** — area under the accuracy-vs-shots curve (AU-SC); defends a label-efficiency claim with one number instead of a table.
- ☐ **Cross-Scale Consistency** — agreement between predictions made from low- vs high-magnification features; defends a multi-scale design.
- ☐ **Calibration gap** — ECE alongside accuracy; defends a clinical-trust claim.

REVIEWER'S LENS A reviewer's first question about any new metric is "why should I trust this number?" Answer it in the same subsection you define it: one sentence on what it captures that standard metrics don't, and one figure showing it correlates with a metric they already accept.

16 Result Visualization & Qualitative Results

Tables prove; visuals persuade. You need both.

- ☐ Add **qualitative results** that visually show *how* your method works don't rely on numeric tables alone.
  - ☐ Show representative successes side by side with the baseline's output.
  - ☐ Annotate what the reader should notice (boxes, arrows, zoom-ins).
- ☐ Turn key numeric results into clean plots (curves, Pareto fronts) where they read faster than tables.
- ☐ Keep visualization styling consistent with the rest of the paper's figures.
- ☐ Make qualitative figures honest representative examples, not only cherry-picked best cases.

17 Effectively Explaining Results

A table or figure on its own is data. The *write-up* is what turns it into evidence. Reviewers do not extract conclusions from your numbers; they verify the conclusions you state. A weak result paragraph leaves them to write their own reading usually a less favorable one. This section is the discipline of writing those paragraphs.

17.1 The four moves of a result paragraph

Every effective result paragraph makes the same four moves, in this order:

- ☐ **Claim** — one sentence stating the takeaway. Lead with it. The reviewer should be able to stop reading after this sentence and still get the point.
- ☐ **Evidence** — the specific number(s) from the table/figure that back the claim. Include the comparison: "**X.X** vs. Y.Y, +Z.Z."
- ☐ **Mechanism / Why** — one sentence on *why* this result happened, tying it to a design choice from the methodology.
- ☐ **Scope / Caveat** — one sentence on where the claim holds (and where it doesn't), so the reader trusts you to police your own claims.

Skip any move and the paragraph leaks: no claim → reviewer writes their own; no mechanism → reads as luck; no scope → reads as overreach.

17.2 Worked example: explaining a main-results table

Take Style A's main-results table (CAMELYON16, TCGA-NSCLC, TCGA-RCC; Ours vs. ABMIL, DSMIL, TransMIL).

REVIEWER'S LENS **Anti-example: data dump.** *"Table 1 shows the results on three datasets. ABMIL achieves 90.1 AUC on CAMELYON16, 94.0 on TCGA-NSCLC, and 97.1 on TCGA-RCC. DSMIL achieves 91.4, 94.6, and 97.4. TransMIL achieves 92.8, 95.1, and 97.6. Our method achieves 94.0, 95.9, and 98.2. We can see that our method performs better."*

This paragraph reads every number aloud, never states a claim, never explains *why*, and forces the reviewer to compute the deltas themselves. Reviewers who do that work draw their own conclusion — often "+1.2 AUC over TransMIL is within noise; weak contribution."

PRO TIP **Effective version (four moves, marked).** *"Our method achieves the best AUC on all three datasets, with the largest gain where prior methods struggle most.[claim] On CAMELYON16 we reach* **94.0** *AUC vs. TransMIL's* 92.8 *(+1.2, paired Wilcoxon $p < 0.01$), and the gain widens at lower shot counts (Table 3).[evidence] We attribute this to the text-guided prototype decoder, which preserves diagnostic patches under weak-label noise — a regime where attention-only baselines spread mass over background tissue (Sec. 4.3).[mechanism] The improvement is smallest on TCGA-RCC (+0.6), where all methods saturate above 97; we therefore foreground the low-shot regime for the headline claim.[scope]"*

The effective version is roughly the same length. It states a conclusion, picks the *one* number that supports it (not all twelve), names the mechanism so the reviewer can connect it to the methodology, and explicitly bounds the claim before the reviewer does. The data dump is the same words rearranged from worst to best use.

17.3 Worked example: explaining an ablation

Take a leave-one-out ablation (Format 2): full model at 87.1, −Attention 84.3, −Gate 85.8, −Aux loss 85.9.

REVIEWER'S LENS **Anti-example.** *"Removing attention drops accuracy to 84.3, removing the gate drops it to 85.8, removing the auxiliary loss drops it to 85.9. All three components contribute."*

The takeaway is buried, every drop sounds equally important, and the reader has no story for *why* attention matters most.

PRO TIP **Effective version.** *"Every component earns its place, but* ***attention is the primary driver****: removing it costs −2.8 accuracy, roughly twice the drop from removing the gate (−1.3) or the auxiliary loss (−1.2).*[claim+evidence] *This is consistent with the methodology's claim that instance weighting — not feature gating — is what recovers the rare diagnostic patches under few-shot supervision (Sec. 3.2).*[mechanism] *The gate and auxiliary loss contribute similar, smaller gains; we keep both because their improvements compose without overlap (Table 5).*[scope]*"*

17.4 Universal scaffolds (fill the slots)

- ☐ **Headline (claim):** "*{Ours} achieves the best {metric} on {N} of {M} datasets, with the largest gains in the {regime} regime.*"
- ☐ **Pointed comparison (evidence):** "*On {dataset}, {Ours} reaches {X.X} vs. {best baseline}'s {Y.Y} ({Δ}, {significance test} $p < \{value\}$).*"
- ☐ **Mechanism:** "*We attribute this to {component}, which {behavior}, consistent with the design rationale in Sec. {ref}.*"
- ☐ **Scope:** "*The gain is smallest on {dataset}, where all methods saturate / the regime is easier; we therefore foreground {regime} for the headline claim.*"
- ☐ **Trend (label-efficiency / sensitivity):** "*The gap {widens / narrows} as {k} {decreases / increases} (Fig. {ref}), supporting our {label-efficiency / robustness} claim.*"
- ☐ **Cost–benefit:** "*{Ours} matches the lightest baseline in {cost} (***Z***M params) while topping AUC — gains do not come from extra capacity.*"

17.5 Universal rules

- ☐ **State the claim first.** Never make the reviewer infer it.
- ☐ **Quote one or two numbers,** not the whole row. Pick the most-comparable cell against the strongest baseline.
- ☐ **Always report the delta** (Δ). Absolute numbers don't tell the reader if +1.2 is a lot.
- ☐ **Name the mechanism.** A result without an explanation reads as luck.
- ☐ **Pre-empt the comparison** a hostile reviewer would draw: who is the strongest baseline, which dataset is closest, and is the gap significant?
- ☐ **Scope the claim** before the reviewer does. "On three pathology datasets" beats "in general."
- ☐ **Link to a section.** Every claim should point at the table/figure/section that backs it ("Table 3," "Sec. 4.2").
- ☐ **Match the metric to the venue.** AUC for binary medical tasks, balanced accuracy for class-imbalance, macro-F1 for multi-class — and say which you are leading with and why.
- ☐ **One paragraph per claim.** If you have three findings in one table, write three short paragraphs, not one long one.

17.6 Trap list: how result write-ups fail

- ☐ **The data dump** — recites every cell, states no claim.
- ☐ **The hedged hero** — "Our method achieves competitive results" when you actually won; under-claiming loses you the contribution.
- ☐ **Over-claim from one cell** — one dataset, one seed, no significance: a single number does not support "state-of-the-art."
- ☐ **Unsupported adjectives** — "substantial," "significant," "robust" with no number, no test, no condition.
- ☐ **Mechanism missing** — numbers improved, but the paragraph never connects them to a design choice; the reader cannot tell what to credit.
- ☐ **Cherry-picking by silence** — discuss only the datasets you win on; reviewers notice the missing ones and the trust cost is higher than admitting the loss.
- ☐ **Caption / text mismatch** — the table caption says one thing, the paragraph another. Lost goodwill on the easiest thing to fix.

REJECTION TRAP "Our method outperforms all baselines" with a 0.3-point lead and no significance test is the most common over-claim in benchmarking papers. State the magnitude, state the test, and if the result is in the noise, say it improves "comparably" or "matches at lower cost" — a defensible claim beats an indefensible one.

REVIEWER'S LENS A senior reviewer reads the prose first and the table second. If the prose makes falsifiable claims that the table backs exactly, the reviewer's job becomes *verification*, not *discovery*. That is the most favorable mode you can put a reviewer in.

18 Failure Analysis & Case Studies

Acknowledging failure builds the trust that wins borderline reviews.

- ☐ **Analyze at least one failure case**, or include a detailed case study at minimum.
  - ☐ Explain *why* the failure happens, not just that it does.
  - ☐ Connect the failure to the method's assumptions or the stated limitations.
- ☐ Use a case study to walk the reader through the method end-to-end on one concrete example.

REVIEWER'S LENS A paper that honestly dissects its own failures reads as scientific and confident. A paper that hides them invites reviewers to go hunting and they will find worse.

19 Abstract Anatomy

The abstract is the single most-read paragraph in your paper, and the only one a reviewer reads before deciding how charitably to read the rest. It is not a summary; it is an argument compressed to five sentences. Write it last, after you know the contributions and the headline numbers, and treat each sentence as a slot with a specific function and a length budget.

19.1 The five-sentence template

- ☐ **Hook ($\sim$1 sentence, $\leq$25 words).** State the high-stakes setting and why anyone should care *before* you mention your method.

- ☐ **Problem / Gap (∼1 sentence, ≤30 words).** Name what existing work cannot do, in concrete terms. "However," / "Yet" / "Despite this" is the natural pivot.
- ☐ **Method (1–2 sentences, ≤50 words).** One sentence of what you propose, one sentence of *how* it works. Name your method here, capitalized, so reviewers have something to call it.
- ☐ **Headline result (1–2 sentences, ≤45 words).** Quote the strongest *specific* number you can defend ("+4.2 AUC over the prior best on TCGA-NSCLC at 16 shots"). Vague claims kill the abstract.
- ☐ **Implication (∼1 sentence, ≤25 words).** What this enables, or the broader takeaway. Avoid "in conclusion"-style filler.

Total budget: roughly 150–200 words. NeurIPS allows ∼250; CVPR is tight. If you exceed budget, cut from the method, not from the result — reviewers forgive method compression, not vague results.

19.2 Worked example (CPath / FSWC)

REVIEWER'S LENS **Anti-example: vague abstract.**
"Few-shot weakly-supervised whole-slide image classification is an important problem. Many methods have been proposed but they have limitations. We propose a new method that improves over prior work. We evaluate it on several datasets and show that our approach achieves competitive results. We hope this work inspires future research."

No specific problem, no named method, no numbers, no implication. A reviewer arrives at the introduction with zero expectations and looks for reasons to reject.

PRO TIP **Effective version (slots marked).**
Diagnosing rare cancer subtypes from whole-slide images is bottlenecked by the cost of obtaining slide-level labels, often only a handful per class.[hook] Existing few-shot weakly-supervised methods rely on either learnable prompts or prototype aggregation, but no framework lets these contested design choices be compared under one pipeline.[gap] We introduce ***FSWC-Bench****, a unified, config-driven framework that reduces every recent prompt/VLM-based MIL method to two swappable slots — the instance-to-bag aggregator and the prompt engine — over a shared CLAM-style backbone.[method, 1] A single YAML file specifies each method, so the same encoder, splits, and seeds can be held fixed across all 13 implemented baselines.[method, 2] Under a fixed CONCH encoder, FSWC-Bench reveals that aggregation design accounts for only +1.3 macro-F1 on average across TCGA-NSCLC, TCGA-RCC, and CAMELYON16, while the prompt engine accounts for +3.8 — inverting the focus of recent literature.[result] The framework supports controlled ablation along ten design axes and is released as a reproducibility scaffold for the field.[implication]*

19.3 Universal rules

- ☐ **Write the abstract last.** You cannot summarize numbers you haven't computed or contributions you haven't sharpened.
- ☐ **Name your method in the method sentence,** capitalized. Reviewers cite by name — give them one.
- ☐ **Quote at least one specific number.** "Improves" is invisible; "+4.2 AUC" is a contribution.
- ☐ **Scope every claim.** "On three pathology datasets" beats "in general"; reviewers trust calibrated language.
- ☐ **Mirror the abstract in the conclusion** — same scope, same headline number. Discrepancies erode trust.
- ☐ **Read it aloud.** If you stumble, the reviewer will too.

REJECTION TRAP The single most common abstract failure is the "competitive results" hedge: an abstract that under-claims because the authors are insecure about the contribution. Reviewers read this as "even the authors aren't sure." If you beat the prior best by 4.2 on the metric the field uses, say so in the abstract.

20 Equation Craft

Equations are read more skeptically than prose. A reviewer who finds a mis-aligned variable or an undefined symbol assumes the rest of the math is similarly careless. Consistency matters more than elegance.

20.1 Numbering, referencing, alignment

- ☐ **Number an equation only if you reference it.** Unreferenced numbered equations clutter the right margin.
- ☐ Use `align` (not `eqnarray`) for multi-line equations. `eqnarray` is deprecated and adds spurious whitespace.
- ☐ **Align on the verb** (=, `\le`, `:=`) using `&` so the eye tracks the operator down the page.
- ☐ Use `\eqref{...}` for "Eq. (3)," not raw `\ref{...}` which prints "3."
- ☐ Reference equations as "Eq. (3)" with a non-breaking space (`~`) so the number doesn't wrap onto the next line.
- ☐ **Define every symbol on first use,** immediately after the equation.
- ☐ Keep one symbol meaning one thing throughout the paper. If θ is the prompt context, never reuse it for the temperature.

20.2 Typography inside equations

- ☐ **Variables are italic** (default math mode). Multi-letter *names* (not single variables) take `\text{}` or `\mathrm{}` so they don't render as a product of letters.
- ☐ **Operators and functions** (`sin`, `log`, `argmax`, `softmax`) take `\operatorname{}` or built-ins (`\sin`, `\log`).
- ☐ **Sets and reals** use `\mathbb{}`: $\mathbb{R}^d$, $\mathbb{E}$, $\mathbb{P}$.
- ☐ **Matrices in bold** (`\mathbf{X}` or `\bm{X}`), **vectors lowercase bold**, **scalars italic, not bold**.
- ☐ **Units belong in the caption,** not the equation. "Throughput T (slides/s)" in prose, never $T = 38.6$ slides/s inline.
- ☐ Use `\,` for a thin space between number and unit when units must appear in math: $4.2\,\mathrm{ms}$, not $4.2ms$.
- ☐ Prefer `\left( ... \right)` for delimiters that scale with content; size them manually only when `\left`/`\right` guesses wrong.

20.3 Common silent mistakes

- ☐ `loss = CrossEntropy(...)` written as $loss = CrossEntropy(\ldots)$ — this renders as a product of letters $l \cdot o \cdot s \cdot s$. Use `\mathcal{L} = \mathrm{CrossEntropy}(\cdot)`.
- ☐ $argmax$, $softmax$, log, sin written without `\operatorname{}` — same issue.
- ☐ Mixing θ (parameter) and Θ (parameter space) but never defining the case.

- ☐ Stray italic "i", "j" indices in subscripts when they should be roman (`ij` as an index pair).
- ☐ Re-defining N as "number of patches" in one section and "number of slides" in another. Pick one.
- ☐ Equations split across two columns or pages because `align` wasn't allowed to break (use `align` not `align*`, or split with `multline`).

20.4 Worked example: a clean equation block

REVIEWER'S LENS **Anti-example.** *The loss is loss = CrossEntropy(softmax(WX+b), y)+ lambda * InfoNCE where W is the weight matrix, X is the input, b is the bias, y is the label, and lambda controls regularization.*

Operators rendered as letter-products, undefined dimensions, λ written as `lambda`, no equation number, and the symbols are explained as a flat list that doesn't track the order in the formula.

PRO TIP **Effective version.**

Given a bag of patch features $\mathbf{X} \in \mathbb{R}^{N\times D}$ *and a slide label* $y \in \{1,\dots,C\}$*, we minimize*

$$\mathcal{L} \;=\; \mathrm{CE}(\mathrm{softmax}(\mathbf{W}\mathbf{X}+\mathbf{b}),\, y) \;+\; \lambda\,\mathcal{L}_{\mathrm{InfoNCE}}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{C\times D}$ *is the alignment projection,* $\mathbf{b} \in \mathbb{R}^{C}$ *is a bias,* CE *denotes cross-entropy, and* $\lambda > 0$ *trades classification against the contrastive term* $\mathcal{L}_{\mathrm{InfoNCE}}$ *defined in Eq.* (2)*.*

LaTeX SOURCE

```
Given a bag of patch features $\mathbf{X} \in \mathbb{R}^{N\times D}$ and a slide
label $y \in \{1,\ldots,C\}$, we minimize
\begin{equation}
\mathcal{L} \;=\; \mathrm{CE}\!\left(\mathrm{softmax}(\mathbf{W}\mathbf{X}+\mathbf{b}),\, y\right)
      \;+\; \lambda \, \mathcal{L}_{\mathrm{InfoNCE}},
\label{eq:loss}
\end{equation}
where $\mathbf{W} \in \mathbb{R}^{C\times D}$ is the alignment projection,
$\mathbf{b} \in \mathbb{R}^{C}$ is a bias, $\mathrm{CE}$ denotes cross-entropy, and
$\lambda > 0$ trades classification against the contrastive term
$\mathcal{L}_{\mathrm{InfoNCE}}$ defined in Eq.~\eqref{eq:loss}.
```

REJECTION TRAP Inconsistent notation between equations, figures, and pseudocode is the easiest flag for a methods reviewer. If your equation uses $\mathbf{X}$, your figure uses F, and your pseudocode uses `feats`, the reviewer will ask which is which — and your confidence cost is higher than the typography cost of fixing it.

21 Reproducibility as Positioning

Reproducibility is usually treated as a defensive checklist: release code, fix seeds, list hyperparameters. Done that way, it earns the paper nothing beyond not losing points. Done as *positioning* — made visible, audited, and named — it gives the AC a concrete reason to defend the paper in discussion. Every borderline paper benefits from looking maturer than its peers, and reproducibility is the cheapest signal of maturity available.

21.1 From present to visible

- ☐ **State reproducibility in the abstract or first paragraph,** not buried in §5: "*We release code, configs, and pre-extracted features; all 47 numbers in this paper are regenerated by* `python reproduce.py`*.*"

- ☐ Add a **single sentence on the headline figure** pointing at the reproduction command ("Figure 1 is generated by `scripts/fig1.py`").
- ☐ Cite your own **anonymous code repository** in the introduction footnote, not only in the supplementary.
- ☐ Include a **compute budget statement** (e.g. "All experiments ran on 1×H100, 80 GB; full sweep takes 14 GPU-hours") — this is the number ACs look for to evaluate fairness.

21.2 The four artifacts

A reproducibility package is taken seriously when it ships all four:

- ☐ **Code** that runs on a fresh machine with one command. Pin dependencies (`requirements.txt` with versions, or a Docker image).
- ☐ **Configs and seeds** in version control — one config file per table row, so a reviewer can map "Table 2, row 3" to "`configs/exp_03.yaml`."
- ☐ **Data or features.** If raw data has a DUA, release the *extracted features* (frozen-encoder outputs are usually sharable) plus the extraction script.
- ☐ **Splits.** Patient-level splits as JSON, generated deterministically from a seed. This is the artifact most papers omit and the one reviewers most often ask about.

21.3 The reproducibility checklist (NeurIPS / CVPR style)

Filling the checklist honestly is itself a positioning move — a maximally-Yes, maximally-specific checklist signals seriousness. Reviewers see it.

- ☐ **Did you state the contributions made by the paper?** (Sec. 1.) Tick if you list them as a bulleted list; mark N/A only if the paper is a survey.
- ☐ **Did you describe the limitations of your work?** (Sec. 6.) "Yes" must point at a specific section, not be a vague claim.
- ☐ **Did you discuss negative societal impact?** (Sec. 6 or appendix.) Required at NeurIPS; reviewers check.
- ☐ **Did you include the full set of assumptions and a complete (and correct) proof?** For theoretical claims — "No" here without justification is a soundness hit.
- ☐ **Did you specify all training details?** Optimizer, batch size, epochs, augmentations, hardware, wall-clock time, random seeds.
- ☐ **Did you report error bars / significance?** "Yes, mean±std over 5 seeds, paired Wilcoxon for headline comparisons."
- ☐ **Did you cite the creators of every dataset / model / asset?** An uncited asset is a desk-reject risk under recent venue rules.

21.4 Make reproducibility a contribution, not a tax

- ☐ **If your code base is unusually clean,** say so — "unified framework with 13 baselines reproducible from 13 config files" is a contribution sentence, not a logistics detail.
- ☐ **Release a benchmark harness** (configs + features + splits + a one-line "reproduce" command) alongside the paper. The community will cite it, which becomes its own contribution stream.
- ☐ **Use the anonymous repo creatively**: a README that walks a reviewer through the headline result in five commands shows depth without de-anonymizing you.
- ☐ **Mention what would change if reproduced exactly** — "CONCH features were extracted

on an H100; CPU extraction matches within ±0.3 macro-F1." This level of detail is rare and disproportionately trust-building.

REVIEWER'S LENS ACs read paper-level reproducibility as a proxy for *authorial rigor*. A paper whose checklist is mostly "Yes, see §X" with specific section pointers reads as written by an author who knows what they did and could defend any number on demand. That impression carries into every other axis of the review.

PRO TIP The single highest-leverage reproducibility move is one figure caption that says: "*Generated by* `scripts/figN.py`*; all numbers are reproducible from the released code.*" It costs ten words and converts an opaque figure into an auditable artifact in the reviewer's mind.

REJECTION TRAP "Code will be released upon acceptance" is the single phrase most associated with later-broken promises in the field, and reviewers know it. Release anonymous code at submission — it costs you nothing and gains you the entire reviewer benefit of doubt that the conditional promise costs you.

22 Limitations & Future Work

- ☐ **Explicitly state the limitations** of the method and the evaluation.
- ☐ Identify concrete **places for improvement** and plausible future directions.
- ☐ Frame limitations as scoping, not apology show you understand the boundaries of your claims.
- ☐ Make sure stated limitations are consistent with (and not contradicted by) your claims.

23 Internal Consistency & Cross-Referencing

A paper that points at itself reads as one coherent argument.

- ☐ **Cross-reference sections throughout**: forward-reference evidence, back-reference motivation.
- ☐ Verify every claim is supported *somewhere*, and that every result is referenced from the text.
- ☐ Ensure numbers in the abstract, intro, tables, and conclusion all agree.
- ☐ Confirm notation, terminology, and method names are identical across all sections.
- ☐ Check that contributions promised in the intro are each delivered and pointed to explicitly.

24 Organizing the Supplementary Material

The supplementary is not a dump. It is the place where a curious reviewer becomes an advocate, and that conversion happens only if the supplementary has its own structure — a table of contents, named appendices, captions that stand alone, and a clear pointer from every main-paper claim to the supplementary section that extends it. Treat the supplementary as a second, deeper paper that mirrors the structure of the first. A reviewer who scrolls through it should come out the other side more impressed than when they went in.

24.1 The four jobs of a supplementary

- ☐ **Depth**: the extended versions of main-paper tables, full ablation grids, additional datasets or encoders. Backs every claim the main paper had to compress.
- ☐ **Honesty**: failure-case galleries, negative results, sensitivity to seeds, the experiments that didn't work. Builds the trust that makes the successes credible.
- ☐ **Interpretability**: qualitative evidence the main paper couldn't fit — attention maps, prototype visualizations, retrieval examples, mistake patterns.
- ☐ **Reproducibility**: protocol details, data preprocessing, hyperparameter sweeps, compute breakdown, library versions, license/DUA disclosures.

Each main-paper claim should map to at least one supplementary section. Each supplementary section should map back to a specific main-paper claim it extends, and that mapping should be stated in the section's first sentence.

24.2 Universal rules

- ☐ **Start with a table of contents.** A two-line TOC at the top of the supplementary is the single move that separates a structured supplementary from a dump.
- ☐ **Use Appendix A–Z labels** (not Sec. 7, 8, 9). Reviewers expect "Appendix B," and the naming makes cross-references readable.
- ☐ **Every appendix opens with a back-pointer:** "*This appendix extends the ⌈main-paper section⌋ by...*" — one sentence, mandatory.
- ☐ **Captions stand alone.** The reviewer is scrolling; assume they haven't read the main paper recently.
- ☐ **Repeat the figure / table styling** from the main paper (same fonts, same palette, same bolding convention). Inconsistency reads as careless.
- ☐ **Don't put method improvements in the supplementary.** At CVPR this is a desk-reject risk. The supplementary is for depth, not for a better version of the paper.
- ☐ **Keep the running header / page numbering continuous** (or restarted with "A1, A2, ...") — pick one and apply it everywhere.

24.3 A worked supplementary skeleton (FSWC-Bench)

The following is a complete table of contents for the supplementary of a benchmarking paper in your domain. It is opinionated about what belongs where and why; copy and adapt. Each line names the appendix, gives its one-sentence back-pointer, and a short list of what it contains.

Supplementary Material – Table of Contents

Appendix A	**Implementation Details** Full training protocol, optimizer/scheduler per method, library versions, hardware, total compute.
Appendix B	**Data Preprocessing** WSI tiling parameters, patch extraction, magnification handling, feature-store format, license/DUA notes.
Appendix C	**Few-Shot Split Construction** The exact k-shot sampling procedure; patient-level controls; per-seed split JSONs released.
Appendix D	**Hyperparameter Sweep Analysis** Per-method sensitivity to learning rate, temperature, prototype count, λ_{InfoNCE}; chosen values defended.
Appendix E	**Extended Main-Results Tables** Per-class precision/recall/F1; per-fold breakdown; results on encoders not in the main paper.
Appendix F	**Full Ablation Grid** Every aggregator × prompt × encoder cell; main paper showed the diagonals.
Appendix G	**Statistical Significance Tests** Paired Wilcoxon and Friedman tests per dataset; critical-difference diagram across all methods.
Appendix H	**Interpretability Gallery** Per-class attention maps; prototype-to-region attributions; multi-scale attention at 5×/20×.
Appendix I	**Failure-Case Gallery** Misclassified WSIs grouped by failure mode; pathologist commentary where available; what each failure implies.
Appendix J	**Negative Results** Variants we tried that did not help (e.g. deeper aggregators, larger prompt context); explanation per row.
Appendix K	**Compute and Carbon Budget** GPU-hours per method, total project compute, kWh / CO_2e estimate.
Appendix L	**Project Page & Interactive Results** Pointer to the anonymous interactive dashboard; what each view lets the reader explore.
Appendix M	**Licenses, DUAs, and Ethics** TCGA / CAMELYON16 / UBC-OCEAN access terms; consent; redistribution policy for released features.
Appendix N	**Reproducibility Checklist** Filled NeurIPS/CVPR checklist with section pointers.

24.4 What each appendix should actually contain

The skeleton above is the structure; the value is in what fills each appendix. This is where most supplementaries collapse to "additional results." The four in your prompt deserve specific guidance.

Appendix B — Data Preprocessing. The reviewer's test: can they reproduce your features from the raw `.svs` files? Each step has a numbered protocol entry. Include:

- ☐ Tissue segmentation parameters (e.g. Otsu threshold, tissue mask morphology).
- ☐ Patch coordinates: tile size at base magnification (e.g. 1024 × 1024 at 40×), stride, overlap.
- ☐ Magnification handling: how 5× patches were obtained — via slide pyramid (preferred) or via downsampling from 20× (state which).
- ☐ Color normalization (or explicit decision *not* to normalize, with rationale).
- ☐ Tissue-foreground filtering: minimum tissue fraction per patch, blur/artifact rejection.
- ☐ Feature-store format: `<feat_root>/<encoder>/<scale>/<slide_id>.pt`; tensor shape and dtype.

- ☐ One worked example: "slide TCGA-XX-YYYY yields $N = 11{,}832$ patches at $20\times$ and $N = 741$ at $5\times$."

Appendix D — Hyperparameter Sweep Analysis. The main paper shows one operating point per method; this appendix shows the *neighborhood* and defends the choice. For each method:

- ☐ A sensitivity *plot* (metric vs. hyperparameter), not a table — curves read faster.
- ☐ Marked: chosen value, the sensitivity around it, and the value used in the original paper.
- ☐ One paragraph per method: "*Performance is flat for* $\tau \in [0.05, 0.10]$*; we use* $\tau = 0.07$ *to match the original CoOp setting.*"
- ☐ A *cross-method* summary table at the end: which methods are robust, which are brittle. This is itself a finding.

Appendix H — Interpretability Gallery. The reviewer's test: do your attention maps actually align with diagnostic regions? Structure as a 3-column gallery:

- ☐ **Column 1:** the whole-slide thumbnail with class label and the model's prediction.
- ☐ **Column 2:** the attention map (heat overlay) at the operating magnification.
- ☐ **Column 3:** a high-resolution crop of the top-attention region, annotated with the morphological feature (if available, with pathologist annotation).
- ☐ Group the gallery by class so the reviewer can compare within-class consistency.
- ☐ Include *both* correct and incorrect cases — correct-only galleries read as cherry-picked.
- ☐ Caption each row with what it demonstrates ("*High attention concentrates on the lepidic-pattern region; consistent with LUAD subtyping*").

Appendix I — Failure-Case Gallery. This is the appendix that converts skeptics. Mature papers don't just show failures; they *taxonomize* them. For FSWC, the natural taxonomy is:

- ☐ **Class-confusion failures**: the model picked an adjacent class (e.g. chromophobe → clear cell). Show the attention map; explain why the regions look similar.
- ☐ **Stain / scanner-shift failures**: misclassifications correlated with cohort. Quantify the correlation.
- ☐ **Insufficient diagnostic tissue**: the slide is mostly necrosis or normal tissue; the model attends to the wrong thing because the right thing isn't there.
- ☐ **Annotation-disagreement failures**: cases where pathologists themselves disagree on the label.
- ☐ For each category: 2–3 example slides, the model's prediction, the model's attention, and a one-paragraph explanation.
- ☐ A final summary table: *what fraction of errors fall into each category*, and what that implies for future work.

REVIEWER'S LENS A reviewer who reads a structured failure analysis ends up with a more positive view of the paper than one who reads only successes. Honest failure analysis signals scientific maturity, and reviewers reward it. Cherry-picked galleries signal the opposite.

24.5 The reading rhythm of a great supplementary

- ☐ Start with the TOC and a one-paragraph reading map of the appendix structure.

- ☐ Open with the *lowest-risk, highest-trust* appendices: implementation details, data preprocessing, splits. Establish that the protocol is sound.
- ☐ Move to *depth*: extended results, full ablation grids, significance.
- ☐ Then *qualitative*: interpretability and failure galleries. By now the reviewer trusts the numbers; they read these for insight.
- ☐ Close with *honesty* (negative results) and *logistics* (compute, licenses, checklist). The last impression is "these people thought of everything."

PRO TIP Write a one-line "what's in this appendix" summary at the top of every appendix, even if the TOC already says it. A reviewer who jumps directly to Appendix H should know within five seconds what they're looking at without flipping back.

REJECTION TRAP The single most common supplementary failure is appendices that exist but are not pointed at from the main paper. If Appendix G has a significance test but the main paper never says "see Appendix G," the reviewer never finds it. For every appendix, find the main-paper sentence that should cite it and add the pointer. An unreferenced appendix is a wasted appendix.

24.6 The interactive project page (when you have one)

A linked, anonymous interactive page is a force multiplier — but only if it does something the PDF cannot. A page that re-renders the paper's tables is boring. A page that lets the reviewer:

- ☐ swap encoder × aggregator × prompt and watch a heatmap update in real time,
- ☐ click any cell of the main results table and see per-seed distributions,
- ☐ open any failure case and pan/zoom across the WSI with the attention overlay,
- ☐ download per-method config + log + seeded result with one click,

is a different artifact. It signals: *we are so confident in our numbers we invite you to interrogate them.* Even reviewers who don't click are influenced by the link's existence. Place the link in Appendix L and reference it from the main paper's introduction footnote.

25 Targeting Top-Tier Venues: NeurIPS & CVPR

At a ∼25% acceptance rate, a technically fine paper is rejected by default. Acceptance is won by clearing the venue's specific bar and by the rebuttal which recent large-scale analyses identify as the single most decisive phase.

25.1 What the scores actually reward

- ☐ Write for **two distinct readers**: the reviewer (depth, rigor) and the Area Chair (the one-paragraph story).
- ☐ Make the **soundness** of claims unimpeachable it is scored separately from novelty and is the easier axis to control.
- ☐ Give reviewers **quotable strengths**: a crisp contribution list and a headline result they can paste into their review.
- ☐ Pre-empt the **top 3 likely weaknesses** in the paper itself so reviewers cannot lead with them.
- ☐ Reviewers are asked to state *"what would raise my score"*; design the paper (and rebuttal) to hit those triggers.

25.2 Reproducibility & the paper checklist

- [ ] Fill the **NeurIPS Paper Checklist** honestly and completely a careless checklist invites scrutiny.
- [ ] Release **code, configs, seeds, and data splits**; state compute (GPU type, GPU-hours) for every experiment.
- [ ] Report **error bars and the significance test used** for headline comparisons.
- [ ] Address **limitations and societal impact** explicitly (NeurIPS reviewers are asked to check this).
- [ ] State all training details (optimizer, schedule, augmentations, hardware) so results are reproducible.

25.3 Supplementary material strategy

- [ ] Keep the **main paper self-contained**; the supplementary is for depth, not for the core argument.
- [ ] **Reference the supplementary** from the main text at the exact point each item supports.
- [ ] Put full ablations, extra qualitative results, proofs, and implementation details in the supplementary.
- [ ] Do **not** put an improved version of the method in the supplementary at CVPR this is explicitly disallowed.
- [ ] Keep everything **self-contained and anonymous**; no external links that subvert review (a desk-reject risk).

25.4 Rebuttal: the decisive phase

- [ ] Open by **thanking reviewers and listing the concerns** you will address (group common ones).
- [ ] Answer the **highest-impact objections first**; lead with new evidence, not with prose.
- [ ] Quote each reviewer's wording, then respond directly make it trivial to map answer to question.
- [ ] Be **concrete and concise**; promise specific revisions and say where they will go.
- [ ] Stay **civil and non-defensive** even with a harsh review; the AC is reading the tone.
- [ ] Explicitly state **"this addresses the concern that would raise your score."**
  - [ ] *NeurIPS:* new experiments **are** allowed in the rebuttal run the small studies reviewers ask for.
  - [ ] *CVPR/ICCV:* rebuttal is **one page**; you generally **cannot** add substantial new experiments reuse and re-present existing results, tables, and figures.

25.5 NeurIPS vs. CVPR mechanics that change your strategy

	NeurIPS	CVPR / ICCV
Page limit	∼9 content pages, refs unlimited	∼8 content pages, refs unlimited
Rebuttal length	Free-form text response	**One page**, fixed template
New results in rebuttal	Allowed (small studies)	Discouraged; re-present existing results
Reproducibility	Mandatory **Paper Checklist**	Code via Anonymous GitHub; no improved-method supp.
Style file	NeurIPS LaTeX (Word discontinued)	CVPR LaTeX author kit
Anonymity	Double-blind	Double-blind (strict on links/identity)
Decisive phase	Author–reviewer discussion	The single one-page rebuttal

REJECTION TRAP Hidden text or instructions aimed at influencing reviewers or LLM tools is treated as an **ethics violation and desk-rejected** at CVPR. Never embed anything in the PDF that is not meant for a human reader.

PRO TIP Post a clean arXiv preprint and release code early analyses of modern review find that visible code/data and timely, substantive rebuttal engagement correlate with acceptance, often outweighing initial reviewer skepticism.

26 Final Pre-Submission & Rebuttal-Proofing Pass

Read the paper as a hostile reviewer looking for the fastest reason to reject.

- ☐ For every claim, ask: *is this backed by analysis in the paper?* If not, add it or soften the claim.
- ☐ For every design choice, ask: *did I justify this?* If not, justify or ablate it.
- ☐ List the three weaknesses a reviewer is most likely to raise; pre-empt each in the text.
- ☐ Confirm reproducibility: code, data, configs, seeds, and environment are described or released.
- ☐ Check formatting, page limits, anonymization, references, and the venue's specific requirements.
- ☐ Proofread captions, equations, and tables one last time typos here erode trust fastest.
- ☐ Have a colleague read only the figures and abstract; if they get the story, you are ready.

PRO TIP Keep a one-page "defense map": each likely reviewer objection on the left, the section/figure/table that answers it on the right. If a row is empty, you have work to do and if a reviewer raises it anyway, your rebuttal is already written.

Tick every box and the only thing left for a reviewer to evaluate is the merit of your idea which is exactly where you want the decision to be made.