INTERVENE OR PRESERVE? SELECTIVE TEST-TIME INTERVENTION FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Target supervision at test time is usually treated as an instruction to adapt. We show that even correct target labels can be unsafe adaptation signals. Under a fixed sparse-label budget, semantic intervention yields large gains when queried labels identify a substantial fraction of the target semantic state, but its benefit rapidly diminishes and can become negative as the same supervision is diluted across increasingly complex target populations. This exposes a fundamental distinction between *having target labels* and *having sufficient semantic authority* to justify changing a frozen model's predictions.

We introduce **Selective Test-Time Intervention (STTI)**, a test-time inference framework that treats sparse target labels as optional semantic evidence rather than mandatory adaptation signals. STTI first constructs a strong preserved hypothesis from complementary frozen vision–language views, then actively queries a tiny number of target labels, uses them to identify shared target semantic correspondences, and constructs a competing intervention hypothesis. A regime-blind, risk-constrained controller chooses between these hypotheses, intervening only when observable evidence indicates that the expected benefit is sufficient to justify the associated risk. Across 11 benchmarks spanning diverse semantic complexities and annotation budgets, STTI retains the large gains available when sparse supervision is informative while progressively reverting to the frozen hypothesis as its semantic coverage diminishes. Mechanistic analyses show that queried labels resolve *semantic populations rather than isolated examples*, while causal diagnostics identify semantic coverage as a principal bottleneck governing intervention utility. These findings support a broader principle for test-time learning: *preserve the frozen semantic prior by default, and intervene only when sparse supervision identifies enough of the target state to warrant changing it.*

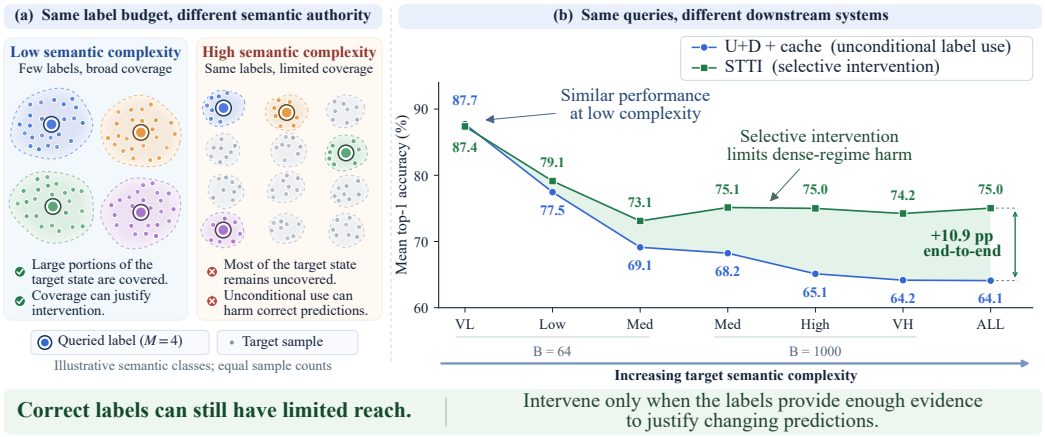


Figure 1: **Correct labels can still have limited reach. (a)** With the same query budget ($M = 4$), sparse labels cover far more of a low-complexity target state than a semantically diverse one; rings mark queried samples (schematic). **(b)** With identical queried supports, CLIP-start U+D+cache system degrades as semantic coverage drops, while STTI better preserves the frozen hypothesis, giving a **10.9 pp** end-to-end gap in ALL. Appendix F.3 gives the matched-P^0 decomposition.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Test-time adaptation (TTA) seeks to exploit target-domain evidence to improve a pretrained model at inference. For vision–language models (VLMs), this has led to prompt optimization Shu et al. (2022); Yoon et al. (2024); Sharifdeen et al. (2025), cache/prototype adaptation Karmanov et al. (2024); Zhang et al. (2024); Huang et al. (2026b), and transductive or distribution-aware inference Zanella et al. (2024; 2025); Huang et al. (2026a). Active TTA goes further by requesting limited target supervision Gui et al. (2024); Wang & Ding (2025); Lee et al. (2025); Lin et al. (2026). Despite their different mechanisms, these approaches share a common premise: *informative target evidence should be exploited to improve prediction*. We ask when that premise fails.

Our key observation is that *even correct target labels can be unsafe adaptation signals*. The issue is not label noise, but *semantic reach*. A queried label can identify a correspondence between a target semantic population and a class, allowing one annotation to repair many predictions. However, as the target state becomes more diverse, the same label budget represents an increasingly smaller fraction of that state. Because geometric proximity does not guarantee class-valid correspondence, propagating this evidence can overwrite predictions that were already correct. Figure 1 exposes this transition: intervention is highly effective when queried labels cover much of the target semantic state, but weakens and can become harmful as that coverage is diluted. We call the evidence that justifies changing the frozen prediction state its *semantic authority*. Correctness and authority differ: a label can be correct for the queried sample yet provide insufficient evidence for broader adaptation.

This failure mode differs from previously studied causes of TTA instability, including noisy or open-world inputs Cao et al. (2025), class imbalance Wang et al. (2026), biased selection Sun et al. (2026); Wu et al. (2026), and adaptation underspecification Shamsi et al. (2026). Recent label-free methods such as StatA Zanella et al. (2025), PTA Huang et al. (2026b), and MOON Huang et al. (2026a) ask how unlabeled target evidence should modify the predictor; they establish the contemporary $M = 0$ adaptation frontier, but are not equal-supervision competitors to our setting. Active TTA instead asks which target samples should be labeled and how the resulting supervision should guide adaptation Gui et al. (2024); Wang & Ding (2025); Lee et al. (2025); Lin et al. (2026). We pose a different decision: *once sparse target labels are available, do they identify enough of the target semantic state to justify changing the frozen predictor at all?* Our focus complements conventional corruption- or covariate-shift TTA by isolating changes in target semantic composition under clean image formation. These experiments do not cover all covariate shifts; they isolate a distinct failure mode caused by sparse supervision when target semantic composition changes.

We introduce **Selective Test-Time Intervention (STTI)**, a preserve-or-intervene framework for VLMs under sparse target supervision. Given a target batch, STTI first constructs a strong *preserved hypothesis* from complementary frozen VLM evidence. It then queries a tiny label budget, uses the revealed supervision to infer shared semantic correspondences, and forms a competing *intervention hypothesis*. A regime-blind, risk-constrained controller decides whether that intervention has sufficient authority to replace the preserved state. The controller has no access to dataset identity, target-complexity regime, or unqueried target labels.

Experiments across 11 benchmarks reveal a consistent mechanism. Sparse labels act on *semantic populations*: queried classes benefit strongly, whereas unconditional propagation can harm unqueried classes. Increasing the label budget extends, but does not eliminate, the regime in which intervention is beneficial. Causal diagnostics further identify semantic coverage as a principal bottleneck: in the densest setting, revealing only the aggregate mass of the population covered by queried classes recovers most of the deficit of unconditional intervention, without revealing sample identities or class assignments. Consistently, STTI intervenes most in concentrated target states and shifts toward preservation as coverage drops. This coverage–intervention pattern holds across frozen VLMs, though the current intervention parameterization is partly backbone-dependent.

We make three contributions: **(1)** we identify a failure mode in sparsely supervised TTA, where correct labels become unsafe when their semantic reach does not cover the population affected by intervention; **(2)** we introduce STTI, separating sparse-label availability from the authority to override a strong frozen hypothesis; and **(3)** we provide mechanistic and causal evidence that sparse labels resolve shared semantic populations, and that semantic coverage determines when intervention helps. These results support a simple principle: *preserve the frozen semantic prior by default, and intervene only when sparse supervision sufficiently identifies the target state to justify change*.

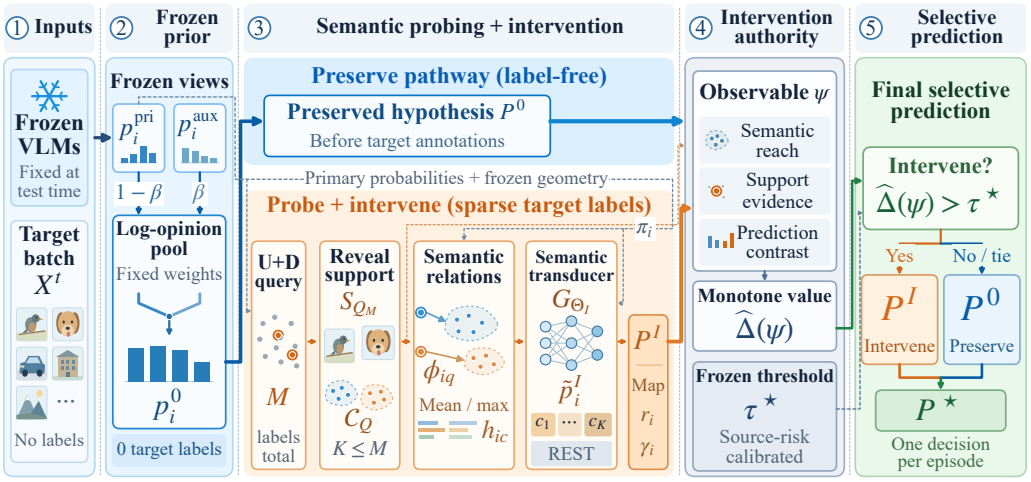


Figure 2: **Overview STTI.** STTI builds a label-free preservation hypothesis P^0, then queries a small target-label budget to infer semantic correspondences and form an intervention hypothesis P^I. Evidence for semantic reach and intervention reliability is summarized by ψ, and a source-calibrated risk controller chooses INTERVENE only if $\widehat{\Delta}(\psi) > \tau^\star$; otherwise it keeps P^0. The decision is made once per target episode, with all learned components frozen at deployment.

2 SELECTIVE TEST-TIME INTERVENTION

STTI explicitly disentangles the *availability* of sparse supervision from its *authority* to modify a frozen predictor: queried labels propose a semantic intervention, and an independent evidential criterion decides whether this intervention should be enacted.

2.1 INTERVENTION AS A TEST-TIME DECISION PROBLEM

Let $\mathcal{X}^t = \{x_i\}_{i=1}^B$ be a target batch from label space $\mathcal{Y} = \{1, \dots, C\}$. We assume frozen vision–language models and a small annotation budget $M \ll B$. At test time, an acquisition rule selects a query set $Q \subset \{1, \dots, B\}$ with $|Q| = M$, revealing only $\mathcal{S}_Q = \{(x_q, y_q) : q \in Q\}$. The remaining target labels are unavailable during inference.

Traditional sparse supervision focuses on how to use $\mathcal{S}_Q$ to adapt the predictor. We instead ask a preceding question: *does the information in $\mathcal{S}_Q$ warrant changing the frozen prediction state at all?* Explicitly, STTI forms two competing hypotheses. The *preserved hypothesis* $P^0 = \{p_i^0\}_{i=1}^B$ contains the strongest label-free semantic evidence available before annotation. The queried labels then induce a competing *intervention hypothesis*:

$$P^I = \mathcal{I}(\mathcal{X}^t, \mathcal{S}_Q; \Theta_I),$$

where Θ_I denotes the frozen intervention parameters and P^I represents the target-specific correction supported by the revealed semantic correspondences. Crucially, observing labels does not force STTI to use P^I. Instead, sparse supervision proposes an intervention whose value must be inferred from observable target evidence. STTI therefore introduces an episode-level decision.

$$a(\mathcal{X}^t, \mathcal{S}_Q) \in \{\text{PRESERVE}, \text{INTERVENE}\}, \quad \text{with final hypothesis} \quad P^\star = \begin{cases} P^I, & a = \text{INTERVENE}, \\ P^0, & a = \text{PRESERVE}. \end{cases}$$

Thus, sparse labels are treated as *evidence for intervention*, rather than as an instruction to adapt. The decision uses neither the identity of the dataset nor a predefined target-complexity regime and receives no supervision beyond the query budget M.

2.2 A STRONG FROZEN PRESERVATION PRIOR

Selective intervention matters only if there is a state worth preserving. We therefore build a strong, entirely label-free *preservation prior* before any target annotations are revealed. Rather than adding

another adaptation mechanism, it defines the semantic state that a label-conditioned intervention must justify overriding.

Let $p_i^{\mathrm{pri}} \in \Delta^{C-1}$ and $p_i^{\mathrm{aux}} \in \Delta^{C-1}$ denote the class posteriors of two complementary frozen VLM views for target sample x_i. Rather than aligning their representation spaces, we combine the views directly in the shared class simplex using a logarithmic opinion pool:

$$p_{ic}^0 = \frac{\left(p_{ic}^{\mathrm{pri}}\right)^{1-\beta} \left(p_{ic}^{\mathrm{aux}}\right)^{\beta}}{\sum_{k=1}^{C} \left(p_{ik}^{\mathrm{pri}}\right)^{1-\beta} \left(p_{ik}^{\mathrm{aux}}\right)^{\beta}}, \tag{1}$$

where probabilities are lower-clipped by a small numerical constant before taking logarithms. This yields the preserved hypothesis $P^0 = \{p_i^0\}_{i=1}^B$.

In our main setup, p^{pri} is the transductive StatA posterior Zanella et al. (2025), and p^{aux} is computed independently using frozen SigLIP2 Tschannen et al. (2025). We fix the source-selected weight $\beta = 0.65$ for all target evaluations; the VLM, fusion weight, and prompts remain fixed. Importantly, the preservation prior is computed *before* acquiring sparse supervision.

The role of P^0 is intentionally conservative. STTI does not ask whether sparse labels can improve a weak zero-shot baseline; it asks whether they provide enough target-specific information to override a strong label-free semantic hypothesis. The key is not building P^0, but gathering the evidence to replace it.

2.3 Sparse Labels as Semantic Probes

In STTI, the annotation budget is not a standard few-shot training set. Queried labels serve as *semantic probes*: each annotation reveals a correspondence between an observed target region and a known class, which may subsequently provide evidence about other target samples sharing that semantic state. Acquisition therefore aims to reveal informative, complementary parts of the target batch, while deferring whether to use those labels to the later intervention stage.

Let $\mathcal{E} \subseteq \{1, \ldots, B\}$ denote the indices of target samples eligible for annotation, and let $\bar{f}_i$ be the ℓ_2-normalized visual representation of x_i. Using only label-free target information, we build a nested query sequence $Q_M = \{q_1, \ldots, q_M\}$. We measure the uncertainty of the primary posterior p_i^{pri} using predictive entropy H_i. The first probe selects the most uncertain eligible sample,

$$q_1 = \arg\max_{i \in \mathcal{E}} H_i, \quad \text{where} \quad H_i = -\sum_{c=1}^{C} p_{ic}^{\mathrm{pri}} \log p_{ic}^{\mathrm{pri}}. \tag{2}$$

After selecting Q_{m-1}, we encourage subsequent probes to explore target regions not covered by the queried set. For each remaining candidate, we compute its distance to the nearest selected representation.

$$D_i^{(m)} = \min_{q \in Q_{m-1}} \left(1 - \bar{f}_i^{\top} \bar{f}_q\right). \tag{3}$$

Because entropy and geometric distance use different scales, STTI rank-normalizes each within the current candidate pool. Let $\mathcal{R}_m(\cdot) \in [0, 1]$ denote this operator; the next probe is selected as

$$q_m = \arg\max_{i \in \mathcal{E} \setminus Q_{m-1}} \left[\frac{1}{2}\mathcal{R}_m(H_i) + \frac{1}{2}\mathcal{R}_m(D_i^{(m)})\right], \qquad m > 1. \tag{4}$$

The revealed support is then $\mathcal{S}_{Q_M} = \{(x_q, y_q) : q \in Q_M\}$. We use a total episode budget $M \in \{1, 2, 4\}$ instead of a fixed number of labels per class; consequently, multiple probes may reveal the same semantic class while other target classes remain entirely unobserved.

Importantly, acquisition and intervention are intentionally decoupled. Eq. 2–4 decide *where to ask*, but do not prescribe how the resulting labels should alter the predictor: no target label affects the acquisition score before selection, and observation $\mathcal{S}_{Q_M}$ does not itself cause adaptation. Instead, queried labels propose a small set of candidate semantic correspondences whose coverage over the target population is estimated next. Thus, sparse supervision provides *evidence about the target state*, not automatic authority to change it.

2.4 QUERIED-CLASS SEMANTIC INTERVENTION

The queried support reveals only a small part of the target label space. Let $\mathcal{C}_Q = \{y_q : q \in Q_M\}$ be the set of unique queried classes, with $K = |\mathcal{C}_Q| \leq M$. Crucially, $\mathcal{C}_Q$ need not cover all classes present in the target batch: with a small label budget, probes can repeat classes while others remain unseen. Thus, we do not assume that sparse supervision identifies the full C-class target state. Instead, STTI constructs the intervention in the reduced semantic space $\mathcal{Z}_Q = \mathcal{C}_Q \cup \{\mathrm{REST}\}$, where REST represents all classes not revealed by the queried support.

Support–target semantic relations. For target x_i and queried support (x_q, y_q), we form a label-legal relational descriptor $\phi_{iq} = \Phi(x_i, x_q, y_q)$. Rather than using the query as a prototype, ϕ_{iq} measures how well the revealed class explains x_i using five evidence types: *visual affinity*, *cache evidence*, *text–class compatibility*, *competitor geometry*, and *local population context*. It encodes target–support similarity/overlap, label-conditioned cache effects, target/support compatibility with the revealed text class, margins to frozen competitors, and the prevalence/coherence of the revealed semantic region. See Appendix B.1 for details. ϕ_{iq} uses y_q but never the unknown label of x_i.

A queried class may be represented by more than one support. We therefore aggregate all support–target relations carrying the same revealed label.

$$\text{For} \quad Q_c = \{q \in Q_M : y_q = c\}, \quad c \in \mathcal{C}_Q, \quad \text{we define} \quad h_{ic} = \left[\operatorname*{mean}_{q \in Q_c} \phi_{iq} \parallel \max_{q \in Q_c} \phi_{iq}\right], \tag{5}$$

where $\parallel$ denotes concatenation. Thus repeated annotations of the same class strengthen the evidence for a single semantic hypothesis rather than creating additional competing states.

Queried-class/REST decomposition. The frozen primary posterior indicates how much probability mass already lies on the revealed classes. For each target, we collapse it into

$$\pi_{ic} = p_{ic}^{\mathrm{pri}}, \qquad c \in \mathcal{C}_Q, \qquad \pi_{iR} = 1 - \sum_{c \in \mathcal{C}_Q} \pi_{ic}, \tag{6}$$

where π_{iR} is the aggregate mass of all unqueried classes. Within this residual state, let r_i be the most likely unqueried class, and define

$$r_i = \arg\max_{c \notin \mathcal{C}_Q} p_{ic}^{\mathrm{pri}}, \qquad \text{and} \quad \gamma_i = \mathrm{clip}\left(\frac{p_{ir_i}^{\mathrm{pri}}}{\max(\pi_{iR}, \varepsilon)}, 0, 1\right) \tag{7}$$

as the fraction of the residual mass explained by that class. $\varepsilon > 0$ prevents division by near-zero REST mass, and clipping limits the redistribution factor to $[0, 1]$. The intervention does not assign new identities to unqueried classes; it only decides whether x_i belongs to a probe-exposed semantic population or remains in the unresolved REST population.

Permutation-equivariant semantic transduction. Because queried classes vary in number and order across episodes, a source-trained semantic transducer operates on the set of class-level relations and the collapsed primary prior.

$$\widetilde{p}_i^I = G_{\Theta_I}\left(\{h_{ic}\}_{c \in \mathcal{C}_Q}, \pi_i\right), \qquad \widetilde{p}_i^I \in \Delta^{|\mathcal{C}_Q|}, \tag{8}$$

The output is a posterior over $\mathcal{C}_Q \cup \{\mathrm{REST}\}$. The semantic transducer G_{Θ_I} is permutation-equivariant with respect to queried-class order and invariant to queried-support order (Proposition 1).

Queried states map to revealed classes. REST keeps the strongest unqueried class $r_i = \arg\max_{c \notin \mathcal{C}_Q} p_{ic}^{\mathrm{pri}}$ and scales its aggregate posterior by γ_i from Eq. 7 before comparing it with queried states.

2.5 INTERVENTION AUTHORITY AND RISK-CONSTRAINED INFERENCE

Constructing P^I does not mean it should replace P^0. The true episode-level benefit of intervention, $\Delta = 100\big[\mathrm{Acc}(P^I) - \mathrm{Acc}(P^0)\big]$, is unavailable at test time because the remaining target labels are unknown. STTI therefore estimates whether the queried supervision has sufficient *intervention authority* using only observable, label-legal evidence.

We summarize each episode by $\psi = \left[\widehat{\rho},\, \widehat{d},\, e_{\mathrm{sup}},\, r_{\mathrm{rep}},\, r_{\mathrm{cor}},\, r_{\mathrm{dis}}\right]$, which captures three complementary signals: *semantic reach* $(\widehat{\rho}, \widehat{d})$, indicating how much of the target state is explained by the queried semantics; *evidence against preservation* (e_{sup}); and *intervention–preservation contrast* $(r_{\mathrm{rep}}, r_{\mathrm{cor}}, r_{\mathrm{dis}})$, measuring the confidence and extent with which P^I challenges P^0. All quantities use only queried labels and model predictions; their exact definitions are given in Appendix C.1.

A source-trained monotone value model predicts the benefit of intervention,

$$\widehat{\Delta}(\psi) = b + \sum_{j=1}^{6} w_j \frac{o_j \psi_j - \mu_j}{\sigma_j}, \qquad w_j \geq 0, \qquad o = (+1, +1, +1, +1, -1, -1), \tag{9}$$

where (μ_j, σ_j) are source statistics and nonnegative coefficients are ridge-fit on source episodes. Fixed orientations encode evidence direction: stronger semantic reach, preservation error, and positive repair evidence cannot lower estimated intervention value, while stronger corruption or unqualified disagreement cannot raise it.

A positive $\widehat{\Delta}$ alone does not justify intervention. We set a source-only threshold $\tau^\star$ to maximize mean intervention utility, subject to the mean utility of the worst 10% of source validation conditions staying above -0.5 percentage points. These validation groups are used only for source-side risk calibration and are never provided to the value estimator; the full objective appears in Appendix C.2.

At target time, all parameters and $\tau^\star$ are frozen, and STTI makes the episode-level decision

$$P^\star = \begin{cases} P^I, & \widehat{\Delta}(\psi) > \tau^\star, \\ P^0, & \text{otherwise.} \end{cases} \tag{10}$$

Thus, queried labels propose an intervention; observable evidence estimates its value; lower-tail risk determines whether it is allowed.

3 Experiments and Analysis

3.1 Experimental Setup

We evaluate 11 VLM classification benchmarks using OpenAI CLIP ViT-B/16, with SigLIP2 as an independent frozen semantic view. Episodes cover seven regimes of increasing semantic complexity with annotation budgets $M \in \{1, 2, 4\}$ labels. All $M > 0$ comparisons use the same budget; acquisition baselines share a common frozen cache consumer, while STTI uses its selective preserve/intervene rule. Additional protocol details are in Appendix A.

3.2 Equal-Budget Test-Time Supervision

Table 1 compares methods with the same total target-label budget as target semantic complexity increases. The $M = 0$ block gives label-free VLM adaptation baselines; our controlled comparisons focus on $M > 0$, where all methods receive the same number of revealed target labels.

Acquisition quality is not intervention authority. Sparse supervision is most useful when a few queried labels cover much of the target state. At $M = 4$, Active-VLM with the common cache consumer achieves 90.4% (Very-Low) and 79.8% (Low), outperforming STTI. But as those four labels spread across a more diverse target population, the advantage vanishes. From B64 Medium onward, STTI leads, scoring 75.1%, 75.0%, 74.2%, and 75.0% in B1000 Medium, High, Very-High, and ALL. Overall, STTI attains 77.0% seven-regime macro accuracy versus 71.2% for the best cache-based control. Strong acquisition can make intervention effective when coverage is concentrated, but it does not guarantee labels remain safe to propagate as the target state grows.

The crossover persists under identical queries, but the standard comparison does not fully isolate label use. U+D+cache applies the same U+D supports to a CLIP-start cache, whereas STTI may retain the stronger P^0. A stricter ten-dataset matched-P^0 control (Appendix F.3) explains much of the dense-regime gap, but STTI still exceeds matched caching by 2.79 pp in Very-High and 3.18 pp in ALL, with paired 95% CIs excluding zero. Thus, P^0 supplies the strong floor; selective authorization protects it when coverage is lowest.

Selectivity is most valuable when supervision is scarce. The same pattern holds at smaller budgets. With one queried label, STTI leads in all regimes, reaching 75.0% macro accuracy versus 66.7% for the best cache control. At $M = 2$, Active-VLM is best only in Very-Low complexity, while STTI leads in the other six regimes and achieves 75.9% versus 68.3%. Thus, a larger budget broadens where direct adaptation helps but does not eliminate the underlying transition. This motivates our next question: *what semantic structure does a queried label actually resolve?*

Table 1: **Test-time adaptation across semantic complexity and supervision budgets.** The $M = 0$ block lists recent label-free VLM adaptation methods as contextual references; they are not equal-supervision competitors for STTI. For $M > 0$, all methods query the same number of target labels. Acquisition controls use a frozen Tip-Adapter-style cache, while STTI intervenes only when semantic evidence warrants it. The table reports mean accuracy; mean $\pm$ standard deviation is in Appendix A.1.

	Small-batch / sparse regimes				Higher semantic complexity				
Method	**B64 VL**	**B64 Low**	**B64 Med.**	**Sparse Avg.**	**B1000 Med.**	**B1000 High**	**B1000 VH**	**ALL**	**Macro**
Recent label-free VLM adaptation references ($M = 0$; contextual)									
CLIP	65.2	65.2	65.2	65.2	65.2	65.2	65.2	65.2	65.2
MTA	66.6	66.6	66.6	66.6	66.6	66.6	66.6	66.6	66.6
Dirichlet	68.5	70.3	67.5	68.8	64.4	45.3	33.6	29.5	54.2
ZLaP	27.5	35.2	44.6	35.8	41.5	52.2	58.4	66.4	46.5
TransCLIP	38.9	40.4	42.7	40.7	56.5	62.0	64.4	**70.3**	53.6
StatA	70.4	69.3	67.4	69.0	69.7	69.7	69.0	69.9	69.3
MOON	**74.6**	**73.1**	**69.4**	**72.4**	**72.9**	**71.3**	**69.2**	—	**—**
Equal-budget sparse-supervision controls ($M = 1$)									
Random + cache	76.3	68.4	64.9	69.8	64.6	63.6	63.3	63.3	66.3
Entropy + cache	77.7	69.5	65.2	70.8	65.0	63.8	63.5	62.3	66.7
Margin + cache	77.7	69.4	65.2	70.8	64.8	63.5	63.1	62.0	66.5
k-center + cache	76.6	68.6	64.7	70.0	64.1	63.1	62.8	62.8	66.1
Active-VLM acquisition + cache†	77.2	69.1	65.1	70.5	64.5	63.3	63.0	63.5	66.5
U+D + cache‡	77.7	69.5	65.2	70.8	65.0	63.8	63.5	62.3	66.7
Selective Test-Time Intervention (Ours)	**78.8**	**74.6**	**72.4**	**75.3**	**75.1**	**75.1**	**74.3**	**75.0**	**75.0**
Equal-budget sparse-supervision controls ($M = 2$)									
Random + cache	80.8	71.1	65.6	72.5	65.1	63.4	62.9	62.7	67.4
Entropy + cache	81.5	71.7	65.8	73.0	65.1	63.4	62.9	62.4	67.5
Margin + cache	81.6	71.7	65.8	73.0	65.4	63.4	62.7	62.1	67.5
k-center + cache	82.3	72.1	66.0	73.5	65.1	63.4	63.0	62.7	67.8
Active-VLM acquisition + cache†	**84.3**	73.8	66.7	75.0	65.5	63.2	62.6	60.9	68.2
U+D + cache‡	82.7	72.8	66.4	73.9	65.8	64.0	63.4	62.8	68.3
Selective Test-Time Intervention (Ours)	83.1	**76.5**	**72.5**	**77.4**	**75.1**	**75.1**	**74.3**	**75.0**	**75.9**
Equal-budget sparse-supervision controls ($M = 4$)									
Random + cache	86.1	75.5	67.9	76.5	67.0	64.2	63.3	63.0	69.6
Entropy + cache	84.7	74.3	67.3	75.4	65.9	63.5	62.7	62.4	68.7
Margin + cache	84.7	74.4	67.4	75.5	66.8	63.9	62.9	61.9	68.9
k-center + cache	88.0	77.1	68.8	78.0	67.2	64.8	64.1	64.1	70.6
Active-VLM acquisition + cache†	**90.4**	**79.8**	70.4	**80.2**	68.6	64.2	62.9	62.2	71.2
U+D + cache‡	87.7	77.5	69.1	78.1	68.2	65.1	64.2	64.1	70.8
Selective Test-Time Intervention (Ours)	87.4	79.1	**73.1**	79.9	**75.1**	**75.0**	**74.2**	**75.0**	**77.0**

†Active-VLM refers to its acquisition rule under our fixed-budget transductive protocol; queried labels use the same frozen cache as all acquisition baselines.
‡U+D+cache uses STTI's U+D supports with the CLIP-start cache; Appendix F.3 reports the stricter matched-P^0 control.

3.3 Sparse Labels Resolve Semantic Populations

Why does the same label budget vary so much with target complexity? We first ask what queried labels resolve. If sparse supervision were merely a small training set, the benefits would stay local to the labeled examples. Instead, its impact is shaped by the semantic classes uncovered by the queries.

For analysis, a target is *covered* if its ground-truth class is in the queried set $\mathcal{C}_Q$, and *uncovered* otherwise; these labels are diagnostic only. At $M = 4$, intervention increases accuracy on covered semantic populations by 21.54 **pp**, but decreases accuracy on fully unqueried classes by 9.96 **pp** with unconditional propagation (Fig. 3(a)). This shows queried-label influence extends beyond the annotated example, but primarily within the semantic population it identifies. A similar pattern holds at the sample level. A diagnostic classifier detects queried-population membership with AUROC 0.991, and distinguishes beneficial repairs from harmful overrides with AUROC 0.883 (95% CI $[0.862, 0.903]$) and average precision 0.953. Sparse labels thus act as *semantic probes*, revealing shared population–class correspondences rather than isolated supervision points.

This also explains why stronger acquisition cannot fully address Table 1. Acquisition decides *which* semantic populations to reveal, but with a fixed budget it can cover only a limited share of an

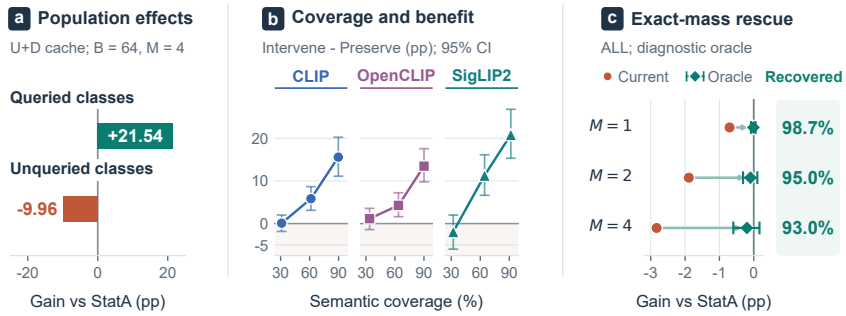


Figure 3: **Sparse labels separate semantic populations, but semantic coverage determines intervention utility. (a)** At $M = 4$, queried classes improve by 21.54 pp, while unconditional propagation lowers fully unqueried-class accuracy by 9.96 pp. **(b)** For CLIP, DataComp OpenCLIP, and SigLIP2, higher queried-class coverage yields larger gains under the same cache intervention ($B = 64$, $M = 4$); error bars show paired-task 95% CIs. **(c)** In the 10-dataset ALL diagnostic, revealing only aggregate queried-class mass recovers 98.7%, 95.0%, and 93.0% of the deficit at $M = 1, 2, 4$, respectively. Hidden target labels are used only for diagnosis.

increasingly diverse target state. Propagating evidence beyond those populations introduces collateral errors. Thus, the key factor is not only the number or informativeness of queries, but the fraction of the target population for which they provide valid semantic evidence.

3.4 Semantic Coverage Governs Intervention Utility

The above analysis suggests that the key factor is not how many labels are queried but how much of the target population they semantically cover. We therefore define queried-class coverage $\rho = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \mathbb{1}[y_i \in \mathcal{C}_Q]$ for analysis: the fraction of evaluation targets whose ground-truth class appears in the queried support. This hidden-label metric is used only for diagnosis; STTI does not observe ρ or unqueried target labels in inference.

The coverage–benefit relation transfers across VLMs. Fig. 3(b) shows that with $B = 64$ and $M = 4$, the intervention benefit increases with the queried-class coverage for OpenAI CLIP, DataComp OpenCLIP and SigLIP2. Because the same cache intervention is applied to all three VLMs rather than the final STTI operator, this crossover is not specific to STTI's controller or to a single primary VLM backbone. When the set of queried classes is broad, sparse supervision works well; but as that coverage narrows, the same label budget no longer sustains wide-reaching intervention. The complete results of the cross-VLM appear in Appendix F.

Controlling semantic mass recovers the dense-regime failure. Correlation does not prove coverage is the bottleneck. We therefore run a stricter oracle diagnostic in the 10-dataset ALL setting. The oracle reveals only the *aggregate number* of targets in queried classes, without revealing identities or labels or changing the ranking, queried class identities, or REST predictions. Using this scalar to set the intervention quota recovers **98.7%, 95.0%, and 93.0%** of the intervention deficit at $M = 1, 2, 4$, respectively (Fig. 3(c)). At $M = 4$, the residual is -0.199 pp, and its dataset-bootstrap 95% CI includes zero. Thus, we can remove most dense-regime failure without changing the semantic predictor; we just need to provide the correct target mass over which to trust its evidence.

These results show that semantic coverage is the main information bottleneck for sparse test-time intervention. Better acquisition can improve coverage, but with a fixed label budget, adequate coverage is not guaranteed as the target state grows more diverse. In this setting, intervention should lose authority and preservation should be preferred. Additional coverage anatomy and oracle diagnostics appear in Appendix E.3.

3.5 Does STTI Intervene When It Should?

Semantic coverage drives intervention utility but is unavailable at inference, so STTI must infer intervention authority from observable evidence alone. We test whether the source-calibrated controller behaves as expected without access to dataset or regime identity, configured K_{eff}, or unqueried

target labels. At $M = 4$, the intervention drops as the complexity of the target increases: from 55.0% (Very-Low) and 24.6% (Low) to 7.3% in $B = 64$ Medium, and to 1.9%, 1.2%, 1.1% and 0% in $B = 1000$ Medium, High, Very-High, and ALL. The controller therefore does not apply fixed adaptation: it intervenes often when few labels capture a shared semantic state, but increasingly reverts to the preserved hypothesis as labels become less representative. To assess the role of the stronger preservation state, Appendix F.3 applies the same U+D cache intervention to the same P^0. Matching the starting state narrows much of the absolute gap, but unconditional use of the same sparse labels still harms performance in Very-High and ALL.

This differs from a conservative "always preserve" strategy. A larger annotation budget makes intervention worthwhile in more episodes, without changing the frozen preservation hypothesis. STTI uses extra supervision only when it provides clear corrective evidence, rather than universally preferring adaptation or preservation. Finally, this trade-off persists even if the decision threshold is relaxed. Source-side risk calibration shows a benefit–safety frontier: more permissive intervention retains more sparse-regime gains but increases worst-case dense-regime losses. Before target evaluation, STTI selects an operating point by maximizing intervention utility under the given lower-tail risk constraint (Appendix C.2). The resulting policy is thus a regime-blind choice between two hypotheses: *intervene when sparse supervision earns sufficient semantic authority; otherwise preserve*.

4 Related Work

Test-time adaptation of vision–language models. VLM test-time adaptation improves zero-shot prediction on unlabeled target data via prompt optimization Shu et al. (2022), transductive inference Zanella et al. (2024), and adaptation of target representations or statistics. Recent backpropagation-free methods include PTA Huang et al. (2026b), which accumulates class-specific target prototypes, and MOON Huang et al. (2026a), which improves realistic transduction through dynamically regularized mixture estimation. The latter is particularly relevant to our setting because it explicitly studies changing target class composition. StatA Zanella et al. (2025) similarly exposes the sensitivity of transductive methods to target composition, while related work considers noisy/open-world targets Cao et al. (2025), adaptation reliability Sheng et al. (2026), and robust adaptation objectives Li et al. (2026); Wu et al. (2026). These methods principally ask *how* unlabeled target evidence should adapt a VLM. STTI addresses a different question: when sparse labeled target evidence is sufficiently representative to justify overriding a strong frozen hypothesis at all.

Sparse supervision, active querying, and selective adaptation. Active TTA introduces limited human supervision to improve reliability or label efficiency Gui et al. (2024); Wang & Ding (2025). BiTTA Lee et al. (2025) uses sparse binary feedback, while the closely related Active-VLM Lin et al. (2026) actively selects informative target examples before adapting the VLM. Selective TTA Chen et al. (2025) instead routes adaptation toward modalities affected by heterogeneous shift. These approaches decide *where to obtain or apply adaptation evidence*. STTI addresses a complementary decision: after sparse labels have been acquired, *is their semantic evidence sufficient to justify intervention rather than preservation?* Correct queried labels are therefore optional intervention evidence, not mandatory adaptation signals.

5 Discussion and Conclusion

Our results reveal a test-time learning gap: *target supervision can be correct yet too unrepresentative to warrant adaptation*. Sparse labels can identify shared semantic populations and deliver large gains when they cover much of the target state, but forced propagation increasingly harms unsupported predictions as coverage declines. STTI addresses this by decoupling the availability of target supervision from its authority to override a strong frozen hypothesis.

The empirical coverage–intervention relationship holds across CLIP, DataComp OpenCLIP, and SigLIP2. Yet the current intervention parameterization is partly backbone-dependent, and preservation fidelity varies with the frozen views. STTI intervenes per episode rather than per sample, motivating finer-grained interventions. Under preservation, queried labels guide the intervention decision but are not propagated to final predictions, trading annotation efficiency for safety. Collectively, our findings support a simple principle: *sparse labels should serve as evidence for intervention, not directives for adaptation; preserve the frozen semantic prior unless target supervision justifies changing it.*

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pp. 446–461. Springer, 2014.

Chentao Cao, Zhun Zhong, Tongliang Liu, Yang Liu, Kun Zhang, Bo Han, et al. Noisy test-time adaptation in vision-language models. In *ICLR*, pp. 59008–59039, 2025.

MingCai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In *ICML*, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pp. 178–178. IEEE, 2004.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. *arXiv preprint arXiv:2404.05094*, 2024.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Jiazhen Huang, Zhiming Liu, Changhu Wang, Wei Ju, Ziyue Qiao, and Xiao Luo. Von mises-fisher mixture model with dynamic shrinkage for realistic test-time transduction. In *ICML*, 2026a.

Zhaohong Huang, Yuxin Zhang, Wenjing Liu, Fei Chao, and Rongrong Ji. Prototype-based test-time adaptation of vision-language models. In *ICML*, 2026b.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, pp. 14162–14171, 2024.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pp. 554–561, 2013.

Taeckyung Lee, Sorn Chottananurak, Junsu Kim, Jinwoo Shin, Taesik Gong, and Sung-Ju Lee. Test-time adaptation with binary feedback. *arXiv preprint arXiv:2505.18514*, 2025.

Aodi Li, Liansheng Zhuang, Xiao Long, Houqiang Li, and Shafei Wang. Flatness guided test-time adaptation for vision-language models. In *ICLR*, pp. 25162–25175, 2026.

Jiaqi Lin, Chaoqi Chen, Xiasi Wang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Wenming Yang, and Qingmin Liao. Probe, anchor, and amend: Active test-time adaptation of vision-language models. In *ECCV*, 2026.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pp. 3498–3505. IEEE, 2012.

Afshar Shamsi, Xiao-Yu Guo, Hamid Alinejad-Rokny, Arash Mohammadi, Damien Teney, and Ehsan Abbasnejad. Multi-hypothesis test-time adaptation to mitigate underspecification. *arXiv preprint arXiv:2607.00259*, 2026.

Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. O-tpt: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *CVPR*, pp. 19942–19951, 2025.

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. *NeurIPS*, 38, 2026.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 35:14274–14289, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Jingwei Sun, Jianing Zhu, Jiangchao Yao, Gang Niu, Masashi Sugiyama, and Bo Han. Bilateral information-aware test-time adaptation for vision-language models. In *ICLR*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Guowei Wang and Changxing Ding. Effortless active labeling for long-term test-time adaptation. In *CVPR*, pp. 25633–25642. IEEE, 2025.

Xucong Wang, Zhe Zhao, Zekun Wang, Xiaofeng Cao, Xu Wang, Di Wu, Pengkun Wang, and Yang Wang. Long-tailed test-time adaptation for vision-language models. In *ICLR*, pp. 46492–46530, 2026.

Xiangyu Wu, Dongming Jiang, Feng Yu, Yueying Tian, Jiaqi Tang, Qing-Guo Chen, Yang Yang, and Jianfeng Lu. Adaptive debiasing tsallis entropy for test-time adaptation. *arXiv preprint arXiv:2602.11743*, 2026.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pp. 3485–3492. IEEE, 2010.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *ICLR*, 2024.

Maxime Zanella, Benoît Gérin, and Ismail Ben Ayed. Boosting vision-language models with transduction. *NeurIPS*, 37:62223–62256, 2024.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *CVPR*, pp. 25103–25112, 2025.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, pp. 493–510. Springer, 2022.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *NeurIPS*, 37:67795–67825, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.

AI USAGE DISCLOSURE

AI tools were used solely for editorial assistance, including language refinement, manuscript organization, LaTeX/table formatting, and consistency checking of author-generated results. They were not used to generate data, run experiments, implement the research method, or make final scientific decisions. All AI-assisted material was reviewed and verified by the authors, who take full responsibility for the final manuscript.

A EXPERIMENTAL PROTOCOL

This section describes the evaluation protocol in Sec. 3. It disentangles three factors often confounded in sparsely supervised test-time learning: (i) target-state semantic complexity, (ii) which target samples are queried, and (iii) how the resulting labels are allowed to affect prediction. This matters because STTI addresses (iii): *whether sparse supervision has enough semantic authority to alter a strong frozen prediction state*.

Benchmarks and frozen VLMs. We evaluate on 11 vision–language classification benchmarks: IMAGENET Deng et al. (2009), SUN397 Xiao et al. (2010), FGVC-AIRCRAFT Maji et al. (2013), EUROSAT Helber et al. (2019), STANFORDCARS Krause et al. (2013), FOOD101 Bossard et al. (2014), OXFORDPETS Parkhi et al. (2012), FLOWERS102 Nilsback & Zisserman (2008), CALTECH101 Fei-Fei et al. (2004), DTD Cimpoi et al. (2014), and UCF101 Soomro et al. (2012). Our main setup uses OpenAI CLIP ViT-B/16 as the primary VLM and SigLIP2 as an independent frozen semantic view. We compute the transductive posterior with StatA Zanella et al. (2025) and combine the two views via the preservation hypothesis (Sec. 2.2). All VLM parameters, prompts, fusion weights, and source-chosen hyperparameters are fixed during target evaluation. Appendix F also reports transfer to DataComp OpenCLIP ViT-L/14 and to SigLIP2 as the primary VLM.

Target episodes and semantic complexity. We construct target episodes by varying batch size B and the effective number of active semantic classes K_{eff}, producing seven evaluation regimes (Table 2). These range from concentrated target states, where few queried labels cover much of the target population, to increasingly diverse states where the same annotation budget covers only a small semantic fraction.

Table 2: **Target-complexity regimes.** K_{eff} denotes the approximate effective number of active semantic classes in a target episode.

Regime	**B**	$\mathbf{K}_{\text{eff}}$
Very-Low	64	$\approx$ 1–4
Low	64	$\approx$ 2–10
Medium	64	$\approx$ 5–25
Medium	1000	$\approx$ 5–25
High	1000	$\approx$ 25–50
Very-High	1000	$\approx$ 50–100
ALL	full dataset	all semantic classes

The repeated Medium semantic band at $B = 64$ and $B = 1000$ is intentional: it disentangles semantic diversity from batch-size dilution. Two target states can have similar effective numbers of active classes, yet the same label count is a much smaller share of a larger target population. The ALL setting instead uses a single full-dataset target task per benchmark covering the entire label space, not repeated samples of a synthetic high-complexity episode.

Sparse target supervision. For each label-based method, $M \in \{1, 2, 4\}$ is the total annotation budget per target episode (not shots per class). The acquisition rule queries exactly M eligible target samples and reveals only their ground-truth labels. Because queries may repeat classes, the number of distinct revealed classes satisfies $|\mathcal{C}_Q| \leq M$, and some target classes may remain unobserved.

All remaining target labels remain hidden until evaluation. STTI does not see configured K_{eff}, dataset identity, semantic-complexity regime, oracle semantic coverage, or any unqueried label when forming the intervention hypothesis or deciding to preserve vs. intervene. If the protocol separates query and evaluation sets, queried samples are excluded from evaluation, and their labels are used only as allowed sparse test-time supervision. Therefore, gains cannot come from scoring the annotated support.

Equal-budget acquisition controls. To separate acquisition quality from sparse-label use, all label-based acquisition controls use the same per-episode budget M. We compare Random, Entropy, Margin, k-center, the acquisition strategy of Lin et al. (2026), and our uncertainty–diversity (U+D) rule.

In these acquisition controls, the downstream label consumer is fixed: queried labels fill the same frozen Tip-Adapter-style cache readout Zhang et al. (2022). Thus, the differences reflect only which target samples are selected, not how their labels are used. Random samples uniformly from the query-eligible pool; Entropy and Margin favor predictive uncertainty; k-center emphasizes geometric diversity. The acquisition rule of Lin et al. (2026) is adapted to the finite transductive protocol and restricted to exactly M target queries, so the results measure its acquisition strategy followed by our common cache consumer, not its original adaptable version. These rows therefore evaluate acquisition policies under our common fixed-budget consumer rather than reproducing the methods' original end-to-end online adaptation pipelines.

Same-query intervention control. U+D follows Sec. 2.3's acquisition rule. U+D + cache queries the same samples but sends the revealed labels to the common frozen cache consumer rather than STTI; therefore, it fixes acquisition but not the frozen reference state (Appendix F.3).

STTI receives the same annotation budget, but does not need to apply queried labels via the cache. Instead, the revealed support forms a candidate intervention hypothesis, and the source-calibrated controller independently decides whether the evidence justifies replacing the preservation hypothesis. Thus, all $M > 0$ comparisons match the annotation cost, while differing in how much *authority* they grant the supervision.

Evaluation and aggregation. We report the accuracy of the top-1 classification. Within each target regime, sparse-supervision methods use matched target-task streams to avoid confounding from episode composition. Regime accuracy is the unweighted mean of the same 11 benchmark datasets. *Sparse Avg.* is the unweighted mean over the three $B = 64$ regimes, and *Macro* is the unweighted mean over all seven target regimes. Dataset-level results are reported in Appendix D.

Published $M = 0$ methods are included as contextual label-free adaptation references and follow their original protocols. They are not treated as equal-supervision baselines against $M > 0$ methods; all sparse-supervision claims use methods evaluated under the matched protocol $M > 0$.

A.1 Aggregate Results with Seed Variability

The compact main-paper table reports mean accuracy only. Table 3 preserves the corresponding seed standard deviations without changing any point estimate.

Table 3: **Full aggregate results with seed variability.** The $M = 0$ block reports recent label-free VLM adaptation methods as contextual references; they are not equal-supervision competitors to STTI. For $M > 0$, all methods receive exactly the same number of queried target labels. Acquisition controls use a shared frozen Tip-Adapter-style cache, while STTI chooses whether the semantic evidence justifies intervention or preservation. For stochastic $M > 0$ settings, entries report mean $\pm$ standard deviation over seeds. Regime values are unweighted means over the 11 datasets; "Sparse Avg." averages the three $B = 64$ regimes, and "Macro" averages all seven.

	Small-batch / sparse regimes				Higher semantic complexity				
Method	**B64 VL**	**B64 Low**	**B64 Med.**	**Sparse Avg.**	**B1000 Med.**	**B1000 High**	**B1000 VH**	**ALL**	**Macro**
Recent label-free VLM adaptation references ($M = 0$; contextual)									
CLIP	65.2	65.2	65.2	65.2	65.2	65.2	65.2	65.2	65.2
MTA	66.6	66.6	66.6	66.6	66.6	66.6	66.6	66.6	66.6
Dirichlet	68.5	70.3	67.5	68.8	64.4	45.3	33.6	29.5	54.2
ZLaP	27.5	35.2	44.6	35.8	41.5	52.2	58.4	66.4	46.5
TransCLIP	38.9	40.4	42.7	40.7	56.5	62.0	64.4	**70.3**	53.6
StatA	70.4	69.3	67.4	69.0	69.7	69.7	69.0	69.9	69.3
MOON	**74.6**	**73.1**	**69.4**	**72.4**	**72.9**	**71.3**	**69.2**	—	**—**
Equal-budget sparse-supervision controls ($M = 1$)									
Random + cache	76.3±0.11	68.4±0.11	64.9±0.02	69.8±0.04	64.6±0.05	63.6±0.01	63.3±0.02	63.3±0.07	66.3±0.03
Entropy + cache	77.7±0.13	69.5±0.08	65.2±0.01	70.8±0.04	65.0±0.04	63.8±0.03	63.5±0.01	62.3±0.09	66.7±0.04
Margin + cache	77.7±0.12	69.4±0.06	65.2±0.04	70.8±0.06	64.8±0.04	63.5±0.05	63.1±0.03	62.0±0.28	66.5±0.07
k-center + cache	76.6±0.10	68.6±0.11	64.7±0.08	70.0±0.07	64.1±0.04	63.1±0.01	62.8±0.04	62.8±0.22	66.1±0.03
Active-VLM acquisition + cache†	77.2±0.11	69.1±0.10	65.1±0.05	70.5±0.05	64.5±0.05	63.3±0.02	63.0±0.03	63.5±0.20	66.5±0.05
U+D + cache‡	77.7±0.13	69.5±0.08	65.2±0.01	70.8±0.04	65.0±0.04	63.8±0.03	63.5±0.01	62.3±0.09	66.7±0.04
Selective Test-Time Intervention (Ours)	**78.8±0.14**	**74.6±0.05**	**72.4±0.06**	**75.3±0.04**	**75.1±0.05**	**75.1±0.02**	**74.3±0.01**	**75.0±0.00**	**75.0±0.02**
Equal-budget sparse-supervision controls ($M = 2$)									
Random + cache	80.8±0.06	71.1±0.05	65.6±0.02	72.5±0.03	65.1±0.06	63.4±0.00	62.9±0.02	62.7±0.08	67.4±0.02
Entropy + cache	81.5±0.09	71.7±0.10	65.8±0.06	73.0±0.08	65.1±0.10	63.4±0.02	62.9±0.01	62.4±0.10	67.5±0.05
Margin + cache	81.6±0.09	71.7±0.08	65.8±0.06	73.0±0.07	65.4±0.04	63.4±0.04	62.7±0.03	62.1±0.18	67.5±0.05
k-center + cache	82.3±0.10	72.1±0.07	66.0±0.08	73.5±0.07	65.1±0.05	63.4±0.01	63.0±0.04	62.7±0.07	67.8±0.03
Active-VLM acquisition + cache†	**84.3±0.02**	73.8±0.10	66.7±0.05	75.0±0.05	65.5±0.09	63.2±0.03	62.6±0.05	60.9±0.07	68.2±0.04
U+D + cache‡	82.7±0.06	72.8±0.09	66.4±0.04	73.9±0.05	65.8±0.05	64.0±0.02	63.4±0.02	62.8±0.09	68.3±0.03
Selective Test-Time Intervention (Ours)	83.1±0.16	**76.5±0.03**	**72.5±0.10**	**77.4±0.09**	**75.1±0.04**	**75.1±0.01**	**74.3±0.02**	**75.0±0.00**	**75.9±0.04**
Equal-budget sparse-supervision controls ($M = 4$)									
Random + cache	86.1±0.00	75.5±0.04	67.9±0.03	76.5±0.03	67.0±0.07	64.2±0.01	63.3±0.02	63.0±0.12	69.6±0.01
Entropy + cache	84.7±0.03	74.3±0.10	67.3±0.08	75.4±0.04	65.9±0.07	63.5±0.01	62.7±0.02	62.4±0.21	68.7±0.03
Margin + cache	84.7±0.06	74.4±0.09	67.4±0.08	75.5±0.03	66.8±0.04	63.9±0.04	62.9±0.02	61.9±0.32	68.9±0.03
k-center + cache	88.0±0.11	77.1±0.01	68.8±0.11	78.0±0.07	67.2±0.07	64.8±0.03	64.1±0.04	64.1±0.10	70.6±0.06
Active-VLM acquisition + cache†	**90.4±0.04**	**79.8±0.13**	70.4±0.05	**80.2±0.05**	68.6±0.10	64.2±0.01	62.9±0.04	62.2±0.13	71.2±0.03
U+D + cache‡	87.7±0.01	77.5±0.07	69.1±0.06	78.1±0.04	68.2±0.05	65.1±0.02	64.2±0.02	64.1±0.10	70.8±0.01
Selective Test-Time Intervention (Ours)	87.4±0.08	79.1±0.03	**73.1±0.10**	79.9±0.07	**75.1±0.04**	**75.0±0.01**	**74.2±0.02**	**75.0±0.00**	**77.0±0.03**

† Active-VLM refers to its acquisition rule under our fixed-budget transductive protocol; queried labels use the same frozen cache as all acquisition baselines.
‡ U+D+cache uses STTI's uncertainty–diversity acquisition rule but replaces selective intervention with a shared cache consumer, separating acquisition from intervention.

Why vary semantic complexity? Conventional few-shot protocols typically increase the amount of supervision as the number of classes increases, confounding label quantity with semantic coverage. Our protocol keeps the annotation budget fixed while expanding the semantic state that those labels may affect, enabling a controlled test of STTI's core hypothesis: a label may be accurate and informative for its identified semantic population, yet become too unrepresentative to justify population-wide intervention as the unresolved target state expands.

The resulting regimes should therefore be interpreted as controlled stress tests of *sparse-supervision reach*, rather than as separate benchmark-specific few-shot settings. In particular, the protocol allows us to distinguish improved query acquisition from the more fundamental question of whether the acquired evidence has sufficient semantic authority to modify the frozen hypothesis.

Source training and model selection. All STTI parameters are fit only on source episodes and then frozen for held-out target evaluation: the semantic transducer G_{Θ_I}, descriptor-normalization statistics, the monotone intervention-value estimator, the preservation-fusion weight β, and the deployment threshold $\tau^\star$. No labels, dataset identity, complexity regime, or other validation signal from the target dataset is used for fitting or model selection.

Source data and outer splits. The source pool includes FGVC-AIRCRAFT, CALTECH101, STANFORDCARS, DTD, EUROSAT, FLOWERS102, FOOD101, OXFORDPETS, SUN397, and UCF101, using the official test partitions from the standard CoOp/Zhou splits Zhou et al. (2022). For each non-ImageNet target dataset d, we apply strict leave-one-dataset-out fitting: d is excluded from transducer training, descriptor normalization, epoch selection, value estimation, and risk calibration,

and the other nine datasets form the source pool. ImageNet is never used as source data; it is evaluated as a strictly unseen eleventh target on its canonical 50,000-image validation set, using all ten non-ImageNet datasets as sources.

Semantic-transducer training. The registered transducer-development stream includes 300 $B = 64$ tasks: 10 datasets, 3 semantic-complexity regimes, and 10 task indices per dataset–regime pair. It also includes fixed uncertainty–diversity and information-gain support sets for $M \in \{1, 2, 4\}$ across 4 cross-fitting folds, totaling 111,708 sample-level training states.

In each non-ImageNet outer fold, the other nine datasets provide 270 source tasks. The task indices 0–7 (216 tasks) are used for optimization, and the indices 8–9 (54 tasks) for epoch selection. After choosing the epoch count, the transducer is reinitialized and refitted for that many epochs using all 270 tasks.

The transducer minimizes categorical NLL over queried-class states and an explicit REST state, using group-DRO across source datasets. We train with AdamW (lr 10^{-3}, weight decay 10^{-4}), batch size 8192, dropout 0.1, group-DRO step size 0.1, and gradient-norm clipping at 5. Training lasts up to 80 epochs (minimum 12) with patience 12. We select the epoch that minimizes worst-source validation NLL. The deployed checkpoint follows the preregistered protocol and uses validation task indices 8 and 9.

Intervention-value fitting and risk calibration. The intervention-value estimator is fitted to a separate source episode stream using the realized intervention gain

$$\Delta = 100\big[\mathrm{Acc}(P^I) - \mathrm{Acc}(P^0)\big].$$

This fixed source stream includes 1,000 tasks in each of six episodic regimes, plus one ALL task per dataset. With $M \in \{1, 2, 4\}$ and four folds, this yields 72,012 episode–budget–fold records per dataset and 720,120 across the ten source datasets.

We fit a six-feature signed-monotone ridge model with equal total weight per dataset–regime–budget cell and ℓ_2 regularization $\lambda = 1$. We obtained nonnegative coefficients via deterministic projected gradient descent with inverse-Lipschitz step size, tolerance 10^{-11}, and a 20,000-iteration cap.

We select the operating threshold $\tau^\star$ on source validation cells by maximizing mean intervention utility under the lower-tail risk constraint (Appendix C.2). For each held-out target evaluation, we then freeze the transducer, value estimator, normalization statistics, fusion weight, and $\tau^\star$ across all annotation budgets and semantic-complexity regimes.

A.2 Computational Cost and Two-VLM Overhead

STTI decouples frozen feature extraction from lightweight post-encoding inference. The primary and auxiliary VLMs each encode each target image once, adding only one extra frozen encoder pass plus image–text scoring and fixed posterior fusion. All features and probabilities are cached; neither VLM is updated or backpropagated.

After encoding, exact pairwise target geometry needs $O(B^2 d)$ similarity computation for feature dimension d. A naive implementation would materialize $O(B^2)$ affinities; we instead compute the matrix in row blocks to reduce peak memory while keeping exact neighbours. With the fixed-k neighbourhood graph, support–target descriptor construction and the frozen intervention model scale as $O(BM)$ for fixed architecture and annotation budget M.

Table 4 reports measured inference cost on SUN397, the largest class vocabulary among the ten non-ImageNet benchmarks ($C = 397$). Encoder timings include image encoding, class-logit evaluation, and softmax. Cached-STTI timings include target-state construction and one full U+D query/evaluation pass at $M = 4$.

Table 4: **Measured inference cost on SUN397.** Median wall-clock time and peak allocated CUDA memory on one NVIDIA RTX 4090 Laptop GPU. Both encoders are resident for the two-view measurement.

Computation	B	C	Time (s)	Peak GPU (GiB)
Frozen image encoding and class scoring				
CLIP ViT-B/16 (primary)	64	397	0.063	1.22
CLIP ViT-B/16 (primary)	1000	397	1.069	2.10
CLIP + SigLIP2	64	397	1.135	3.51
CLIP + SigLIP2	1000	397	23.970	4.21
Complete STTI after features are cached				
Target state + $M = 4$ decision	64	397	0.146	0.12
Target state + $M = 4$ decision	1000	397	1.004	0.04

Protocol note. Encoder measurements use two warm-up passes and seven timed repetitions; cached-STTI uses five warm-ups and fifteen runs. Inputs are preallocated at native resolution (224 for CLIP, 384 for SigLIP2). Timings exclude disk I/O, image decoding, CPU preprocessing, text-prompt encoding, and one-time model loading. For cached runs, $B = 64$ and $B = 1000$ use registered Medium and High episodes, respectively; $B = 1000$ uses the memory-efficient StatA solver.

Result. Most added cost comes from the auxiliary frozen encoder, not selective intervention. With cached representations, the full STTI decision takes 0.146 s at $B = 64$ and 1.004 s at $B = 1000$, while two-view frozen encoding takes 1.135 s and 23.970 s, respectively (Table 4). With both VLMs loaded, peak GPU memory is 4.21 GiB. Since the encoders are independent and frozen, they can also run sequentially and be cached without affecting predictions.

B Details of Queried-Class Semantic Intervention

B.1 Support–Target Relational Evidence

This section details the relational evidence operator Φ used in Sec. 2.4. For each unlabelled target x_i and queried support (x_q, y_q), $\Phi(x_i, x_q, y_q)$ is constructed entirely from frozen model outputs, target–support geometry, and the revealed support label y_q. The ground-truth label of x_i is never used.

The deployed descriptor contains 44 observable statistics organized into five complementary families: visual affinity, cache evidence, text–class relation, competitor geometry, and local population context. Table 5 summarizes their roles.

Table 5: Support–target relational evidence used by the semantic intervention model. All quantities are label-legal: they may depend on the revealed support label y_q, but never on the unknown label of target x_i.

Family	**#**	**Information captured**
Visual affinity	5	Target–support cosine affinity, its target-batch percentile rank and standardized score, whether the support is the nearest queried sample, and whether its revealed class matches the nearest-support class.
Cache evidence	8	Kernel affinity and normalized weight, support-conditioned cache mass for the revealed class, support-class prevalence, the corresponding cache increment, and current cache prediction/margin.
Text–class relation	9	Compatibility of the target and support with the revealed text class, CLIP/StatA probability assigned to that class, and whether frozen predictions already select the revealed class.
Competitor geometry	9	Relative alignment of the revealed class against the current StatA/CLIP competitors, text-anchor similarity, frozen prediction margins, and target uncertainty.
Population context	13	Reciprocal neighbourhood structure, neighbourhood overlap and density, local prevalence of the revealed class, predicted class mass, the size, similarity, and coherence of the corresponding semantic region, and the prediction-derived effective class count.
Total	**44**	

The deployed intervention model uses exactly the 44 statistics in Table 5. Additional task-normalized quantities computed by the analysis pipeline are used only for diagnostics and ablations and are not supplied to the semantic transducer at target inference.

When several queried supports reveal the same class c, their pairwise descriptors are aggregated using element-wise mean and maximum statistics, as defined in Eq. 5. Since each pairwise descriptor contains 44 statistics, this produces an 88-dimensional class-level descriptor h_{ic}. The mean captures the typical relation to the queried class, while the maximum retains the strongest available support–target correspondence; both operations are invariant to support ordering.

Exact relational statistics. The descriptor in Table 5 is the fixed, label-legal map

$$\phi_{iq} = \Phi(x_i, x_q, y_q) \in \mathbb{R}^{44},$$

used by the deployed semantic transducer. Let $c = y_q$ be the revealed class of queried support x_q, let f_i and f_q denote ℓ_2-normalized frozen visual features, and let t_c denote the normalized text embedding of class c. We write p_i^{S} for the StatA posterior and

$$\ell_{ic}^{\mathrm{C}} = 100 f_i^{\top} t_c, \qquad p_i^{\mathrm{C}} = \mathrm{softmax}(\ell_i^{\mathrm{C}})$$

for the frozen CLIP logits and posterior. Their predictions are $\hat{y}_i^{\mathrm{S}}$ and $\hat{y}_i^{\mathrm{C}}$, respectively. Every statistic below depends only on frozen target-batch quantities and the revealed support labels; the unknown label of x_i is never used.

Visual affinity (5). For a target–support pair, define

$$s_{iq} = f_i^\top f_q.$$

For fixed support q, let $\mathrm{rank}_{\mathcal{T}}(s_{\cdot q})_i \in [0,1]$ denote the percentile rank of s_{iq} across the target batch and

$$z_{iq} = \frac{s_{iq} - \mu_q}{\max(\sigma_q, \epsilon)}$$

its batch-standardized affinity. If $q_i^\star = \arg\max_{r \in Q} s_{ir}$ is the nearest queried support, the five statistics are

$$\left[s_{iq},\ \mathrm{rank}_{\mathcal{T}}(s_{\cdot q})_i,\ z_{iq},\ \mathbb{1}[q = q_i^\star],\ \mathbb{1}[y_q = y_{q_i^\star}]\right].$$

They encode both the strength of the pairwise relation and whether the revealed class agrees with the strongest available support evidence.

Cache evidence (8). Using the frozen Tip-Adapter-style kernel,

$$\kappa_{iq} = \exp[-\beta(1 - s_{iq})],$$

define the total support weight $Z_i = \sum_{r \in Q} \kappa_{ir}$ and the class-conditioned cache mass

$$a_{ic} = \sum_{r \in Q:\, y_r = c} \kappa_{ir}.$$

Let $A_i = \sum_d a_{id}$, $Q_c = \{r \in Q : y_r = c\}$, and

$$z_i^{\mathrm{tip}} = \ell_i^{\mathrm{C}} + \alpha a_i, \qquad \hat{y}_i^{\mathrm{tip}} = \arg\max_d z_{id}^{\mathrm{tip}}.$$

Writing $\mathrm{margin}(v)$ for the difference between its largest and second-largest entries, the eight cache statistics are

$$\left[\kappa_{iq},\ \frac{\kappa_{iq}}{Z_i},\ a_{ic},\ \frac{a_{ic}}{A_i},\ \frac{|Q_c|}{|Q|},\ \alpha\kappa_{iq},\ \mathbb{1}[\hat{y}_i^{\mathrm{tip}} = c],\ \mathrm{margin}(z_i^{\mathrm{tip}})\right].$$

Thus, cache evidence distinguishes raw proximity from the actual label-conditioned influence exerted by the revealed support.

Text–class relation (9). The revealed label supplies a semantic anchor through t_c. The nine statistics are

$$\begin{aligned}&\big[f_i^\top t_c,\ f_q^\top t_c,\ p_q^{\mathrm{C}}(c),\ p_i^{\mathrm{C}}(c),\ p_i^{\mathrm{S}}(c),\\ &\ \mathbb{1}[\hat{y}_i^{\mathrm{S}} = c],\ \mathbb{1}[\hat{y}_i^{\mathrm{C}} = c],\ \mathbb{1}[\hat{y}_q^{\mathrm{S}} = c],\ \mathbb{1}[\hat{y}_q^{\mathrm{C}} = c]\big].\end{aligned}$$

These quantities distinguish visual proximity from explicit agreement between the target, support, revealed text class, and the two frozen predictive views.

Competitor geometry (9). We next measure the revealed class relative to the frozen alternatives currently preferred for the target and support. The corresponding statistics are

$$\begin{aligned}&\big[f_i^\top (t_c - t_{\hat{y}_i^{\mathrm{S}}}),\ f_i^\top (t_c - t_{\hat{y}_i^{\mathrm{C}}}),\\ &\ f_q^\top (t_c - t_{\hat{y}_q^{\mathrm{S}}}),\ f_q^\top (t_c - t_{\hat{y}_q^{\mathrm{C}}}),\\ &\ t_c^\top t_{\hat{y}_i^{\mathrm{S}}},\ t_c^\top t_{\hat{y}_i^{\mathrm{C}}},\\ &\ \mathrm{margin}(p_i^{\mathrm{S}}),\ \mathrm{margin}(\ell_i^{\mathrm{C}}),\ \mathcal{H}(p_i^{\mathrm{S}})\big],\end{aligned}$$

where the normalized StatA Zanella et al. (2025) entropy is

$$\mathcal{H}(p_i^{\mathrm{S}}) = -\frac{1}{\log C} \sum_{d=1}^{C} p_i^{\mathrm{S}}(d) \log p_i^{\mathrm{S}}(d).$$

This family asks whether the revealed class has sufficient semantic support to challenge the target's existing frozen explanation.

Population context (13). Let $\mathcal{N}_k(i)$ be the k nearest target neighbors of x_i in the frozen visual space, with $k = \min(5, B-1)$ in the implementation. We define reciprocal neighborhood support and neighborhood overlap as

$$r_{iq} = \mathbb{1}\left[q \in \mathcal{N}_k(i) \wedge i \in \mathcal{N}_k(q)\right],$$

and

$$J_{iq} = \frac{|\mathcal{N}_k(i) \cap \mathcal{N}_k(q)|}{|\mathcal{N}_k(i) \cup \mathcal{N}_k(q)|}.$$

Local density is the mean cosine affinity for the same neighborhood,

$$d_i = \frac{1}{|\mathcal{N}_k(i)|} \sum_{j \in \mathcal{N}_k(i)} f_i^\top f_j.$$

For the revealed class c, its local StatA and CLIP prevalence around x_i are

$$\nu_{ic}^{\mathrm{S}} = \frac{1}{|\mathcal{N}_k(i)|} \sum_{j \in \mathcal{N}_k(i)} \mathbb{1}[\hat{y}_j^{\mathrm{S}} = c],$$

$$\nu_{ic}^{\mathrm{C}} = \frac{1}{|\mathcal{N}_k(i)|} \sum_{j \in \mathcal{N}_k(i)} \mathbb{1}[\hat{y}_j^{\mathrm{C}} = c].$$

We further define the global hard-prediction masses

$$m_c^{\mathrm{S}} = \frac{1}{B} \sum_j \mathbb{1}[\hat{y}_j^{\mathrm{S}} = c], \qquad m_c^{\mathrm{C}} = \frac{1}{B} \sum_j \mathbb{1}[\hat{y}_j^{\mathrm{C}} = c].$$

Let

$$\mathcal{R}_c = \{j : \hat{y}_j^{\mathrm{S}} = c\}$$

denote the StatA-predicted semantic region of class c. When $\mathcal{R}_c \neq \varnothing$, its normalized visual center is

$$\mu_c = \mathrm{norm}\left(\frac{1}{|\mathcal{R}_c|} \sum_{j \in \mathcal{R}_c} f_j \right),$$

with target–region similarity and region coherence

$$\mathrm{sim}_{ic}^{\mathrm{reg}} = f_i^\top \mu_c, \qquad \mathrm{coh}_c^{\mathrm{reg}} = \frac{1}{|\mathcal{R}_c|} \sum_{j \in \mathcal{R}_c} f_j^\top \mu_c.$$

Both are set to zero when $\mathcal{R}_c$ is empty.

Finally, let

$$\bar{p}_c^{\mathrm{S}} = \frac{1}{B} \sum_{i=1}^{B} p_i^{\mathrm{S}}(c).$$

The prediction-derived effective number of semantic classes is

$$K_{\mathrm{eff}}^{\mathrm{S}} = \exp\left(-\sum_{c=1}^{C} \bar{p}_c^{\mathrm{S}} \log \bar{p}_c^{\mathrm{S}} \right).$$

The 13 population statistics are therefore

$$\big[r_{iq},\ J_{iq},\ d_i,\ d_q,\ d_i - d_q,\ \nu_{ic}^{\mathrm{S}},\ \nu_{ic}^{\mathrm{C}},\ m_c^{\mathrm{S}},\ m_c^{\mathrm{C}},\ |\mathcal{R}_c|/B,\ \mathrm{sim}_{ic}^{\mathrm{reg}},\ \mathrm{coh}_c^{\mathrm{reg}},\ K_{\mathrm{eff}}^{\mathrm{S}} \big].$$

These quantities determine whether the pairwise relation is isolated or supported by a coherent target population associated with the revealed semantics.

Class-level aggregation. The five families provide $5 + 8 + 9 + 9 + 13 = 44$ pair-level statistics. For each unique queried class c, repeated supports are collapsed exactly as in Eq. 5,

$$h_{ic} = \left[\operatorname*{mean}_{q \in Q_c} \phi_{iq} \parallel \max_{q \in Q_c} \phi_{iq} \right] \in \mathbb{R}^{88}.$$

The resulting 88-dimensional class descriptor is standardized using source-training means and scales, clipped to the fixed numerical range used by the frozen checkpoint, and then passed to the shared relation encoder. No target-domain statistic is fitted or updated during deployment.

B.2 PERMUTATION-EQUIVARIANT SEMANTIC TRANSDUCER

This section details the semantic transducer G_{Θ_I} introduced in Sec. 2.4. Given the class-level relational descriptors produced in Appendix B.1, the transducer predicts over the episode-specific state space $\mathcal{Z}_Q = \mathcal{C}_Q \cup \{\text{REST}\}$. Because both the number and ordering of queried classes vary across episodes, its construction is permutation equivariant with respect to $\mathcal{C}_Q$. All parameters and normalization statistics are learned from source episodes and remain fixed during target inference.

Shared relation encoding. For each target x_i and queried class $c \in \mathcal{C}_Q$, Sec. B.1 produces a class-level descriptor h_{ic} by aggregating all queried supports carrying label c. Before encoding, these descriptors are standardized using source-derived feature statistics stored with the frozen checkpoint. A shared relation encoder then maps each class independently,

$$e_{ic} = E_{\theta_E}(h_{ic}), \qquad c \in \mathcal{C}_Q. \tag{11}$$

The same encoder parameters are used for every queried class, preventing the model from assigning meaning to an arbitrary candidate ordering.

Permutation-invariant semantic context. A queried class should be interpreted relative to the other semantics revealed in the same episode. We therefore summarize the set of encoded class relations using element-wise mean and maximum pooling,

$$\bar{e}_i = \frac{1}{K} \sum_{c \in \mathcal{C}_Q} e_{ic}, \qquad e_i^{\max} = \max_{c \in \mathcal{C}_Q} e_{ic}, \qquad K = |\mathcal{C}_Q|. \tag{12}$$

Both summaries are invariant to permutations of $\mathcal{C}_Q$ and provide episode-level context for interpreting each candidate semantic population.

Collapsed-prior context. The relational evidence is complemented by observable information from the primary frozen posterior. Recall from Sec. 2.4 that

$$\pi_{ic} = p_{ic}^{\text{pri}}, \qquad \pi_{iR} = 1 - \sum_{c \in \mathcal{C}_Q} \pi_{ic}, \quad \text{and let} \quad q_i = \sum_{c \in \mathcal{C}_Q} \pi_{ic} = 1 - \pi_{iR}$$

denote the total primary-posterior mass assigned to queried classes. For each queried class, we construct a compact scalar descriptor s_{ic} containing its prior mass, relative mass and rank within the queried set, together with global information about the number of revealed semantics and the residual REST state. A corresponding descriptor s_{iR} summarizes the unresolved residual state, including its aggregate mass and the concentration of that mass on its strongest unqueried class. All of these quantities are derived from observable frozen predictions; no unqueried target label is used.

Queried-class and REST prediction. A shared queried-class head combines the class-specific relation embedding, the permutation-invariant set context, and the corresponding collapsed-prior context,

$$\ell_{ic}^I = g_{\theta_Q}\left(e_{ic} \,\|\, \bar{e}_i \,\|\, e_i^{\max} \,\|\, s_{ic}\right), \qquad c \in \mathcal{C}_Q, \tag{13}$$

where $\|$ denotes concatenation. A separate REST head operates only on permutation-invariant episode context and residual-state statistics,

$$\ell_{iR}^I = g_{\theta_R}\left(\bar{e}_i \,\|\, e_i^{\max} \,\|\, s_{iR}\right). \tag{14}$$

The intervention posterior is then

$$\widetilde{p}_i^I = \text{softmax}\left(\{\ell_{ic}^I\}_{c \in \mathcal{C}_Q}, \ell_{iR}^I\right). \tag{15}$$

Importantly, the deployed transducer does not add the collapsed primary log-posterior directly to these logits. The primary predictor supplies semantic context through π_i, but the intervention posterior itself is determined by the learned support–target relational evidence. Consequently, G_{Θ_I} is not a fixed residual correction to the primary predictor.

Proposition 1 (Permutation equivariance of semantic intervention)**.** *For any permutation σ of the queried classes $\mathcal{C}_Q$, if the class-level relational descriptors $\{h_{ic}\}_{c\in\mathcal{C}_Q}$ and the corresponding collapsed-prior probabilities $\{\pi_{ic}\}_{c\in\mathcal{C}_Q}$ are permuted consistently by σ, while π_{iR} remains unchanged, then the intervention posterior $\widetilde{p}_i^I$ permutes its queried-class probabilities by σ while leaving the REST probability unchanged. Consequently, after remapping the queried-class states to their revealed labels, the intervention prediction is invariant to the ordering of queried classes and queried supports.*

Proof. The relation encoder E_{θ_E} and queried-class head g_{θ_Q} share parameters across queried classes, so permuting $\mathcal{C}_Q$ permutes the corresponding class embeddings $\{e_{ic}\}$ accordingly. The episode summaries $\bar{e}_i$ and $e_i^{\max}$ use permutation-invariant mean/max pooling and therefore remain unchanged. The prior descriptors s_{ic} are constructed from the corresponding queried-class probabilities π_{ic} together with permutation-invariant episode statistics, and hence are permuted in the same way. Because g_{θ_Q} is shared, the queried-class logits are therefore permuted accordingly.

The REST descriptor s_{iR} depends only on aggregate residual-state statistics and permutation-invariant context, so the REST logit produced by g_{θ_R} remains unchanged. Softmax preserves the permutation of the queried-class logits while leaving the REST probability unchanged. Remapping the queried-class probabilities to their revealed labels therefore leaves the final intervention prediction unchanged. Mean/max aggregation over repeated supports is also invariant to support ordering. Hence, the intervention is equivariant to queried-class permutations and invariant to queried-support ordering. □

Mapping to the original label space. Queried-class coordinates map directly to their revealed labels. For the unresolved REST state, let

$$r_i = \arg\max_{c\notin\mathcal{C}_Q} p_{ic}^{\mathrm{pri}}, \qquad \gamma_i = \mathrm{clip}\left(\frac{p_{ir_i}^{\mathrm{pri}}}{\max(\pi_{iR},\varepsilon)}, 0, 1\right), \tag{16}$$

where $\varepsilon > 0$ prevents division by vanishing REST mass, clipping keeps the redistribution factor in $[0, 1]$, and γ_i quantifies how much residual mass is concentrated in the strongest unqueried class. Because $\widetilde{p}_{iR}^I$ aggregates unresolved classes, the score for representative r_i is scaled by γ_i. The intervention prediction is therefore

$$\hat{y}_i^I = \begin{cases} c^\star, & c^\star = \arg\max_{c\in\mathcal{C}_Q} \widetilde{p}_{ic}^I, \quad \widetilde{p}_{ic^\star}^I \geq \gamma_i \widetilde{p}_{iR}^I, \\ r_i, & \text{otherwise.} \end{cases} \tag{17}$$

This mapping prevents the aggregate REST state from being treated as if it provided the same semantic certainty as an explicitly queried class.

Frozen target-time inference. The intervention parameters are

$$\Theta_I = \{\theta_E, \theta_Q, \theta_R\}.$$

They are learned exclusively from source episodes and frozen before target evaluation, together with all feature-normalization statistics. STTI therefore performs no target-time parameter optimization: target labels enter only through the explicitly queried supports and the relational evidence derived from them.

Implementation details. The deployed relation encoder maps each standardized 88-D class descriptor through an $88 \to 64 \to 64$ MLP. The queried-class and REST heads use width-64 and width-32 hidden layers before their scalar outputs, with GELU activations, LayerNorm, and dropout 0.1. Standardized relation descriptors are clipped to $[-12, 12]$, and numerical probability guards use $\varepsilon = 10^{-9}$. The deployed configuration uses permutation-invariant set context and a learned REST head, without an additive StatA log-posterior offset.

Table 6: **Summary of the deployed permutation-equivariant semantic transducer.** Structural components of the frozen intervention model.

Component	Configuration
Pairwise relational evidence	44 observable support–target statistics
Class-level relation	Element-wise mean and maximum aggregation over supports sharing the same queried class
Class-level descriptor	88-dimensional mean–maximum representation
Maximum queried-class states	Determined by the maximum annotation budget, $K_{\max} = M_{\max} = 4$
Relation encoder	Shared MLP applied independently to every queried-class descriptor
Episode-level set context	Permutation-invariant mean and maximum pooling across queried classes
Queried-class predictor	Shared prediction head combining class relation, set context, and collapsed-prior context
REST predictor	Separate head using permutation-invariant context and residual-state statistics
Primary-posterior usage	Used as contextual evidence; no additive log-posterior offset
Target-time optimization	None; all intervention parameters and normalization statistics remain frozen

C Intervention Authority and Risk Calibration

C.1 Observable Intervention-Authority Statistics

This section gives the exact definitions of the six observable statistics used by the intervention-authority estimator in Sec. 2.5. All statistics are computed using only frozen model predictions, the explicitly queried labels, and the intervention posterior. No unqueried target label, dataset identity, benchmark regime, or configured target complexity is used.

Let $\mathcal{T}$ denote the unlabelled evaluation targets, $\mathcal{C}_Q$ the set of unique classes revealed by the queried support, p_i^{pri} the primary frozen posterior, p_i^0 the preservation posterior, and $\widetilde{p}_i^I$ the queried-class/REST intervention posterior.

Estimated semantic coverage. The intervention posterior provides an observable estimate of how much target probability mass is explained by the semantic classes revealed by the probes. We define

$$\widehat{\rho} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \sum_{c \in \mathcal{C}_Q} \widetilde{p}_{ic}^I. \tag{18}$$

Importantly, $\widehat{\rho}$ is not the oracle fraction of targets whose ground-truth classes belong to $\mathcal{C}_Q$; it is estimated entirely from the intervention posterior.

Information density. Coverage also depends on how many distinct semantics are revealed relative to the apparent complexity of the target batch. We estimate the latter from the marginal primary posterior

$$\bar{p}_c^{\mathrm{pri}} = \frac{1}{B} \sum_{i=1}^{B} p_{ic}^{\mathrm{pri}}, \tag{19}$$

and its entropy-derived effective class count

$$K_{\mathrm{eff}} = \exp\left(- \sum_{c=1}^{C} \bar{p}_c^{\mathrm{pri}} \log \bar{p}_c^{\mathrm{pri}} \right). \tag{20}$$

The observable information-density statistic is then

$$\widehat{d} = \min\left(\frac{|\mathcal{C}_Q|}{K_{\mathrm{eff}}}, 1 \right). \tag{21}$$

This quantity depends only on predicted target semantics and the number of distinct classes revealed by the query budget; it does not use the configured K_{eff} range defining an experimental regime.

Evidence against preservation. Because queried labels are the only target labels available to STTI, they provide direct evidence about whether the preserved hypothesis is locally consistent with the target domain. We measure

$$e_{\text{sup}} = \frac{1}{M} \sum_{q \in Q_M} \left(1 - p^0_{qy_q}\right). \tag{22}$$

A large e_{sup} therefore indicates that the newly revealed target semantics are poorly explained by the frozen preservation state.

Intervention–preservation contrast. The remaining statistics quantify how strongly the proposed intervention departs from preservation in the unlabeled evaluation targets.

For the preserved hypothesis, define

$$\hat{y}^0_i = \arg\max_c p^0_{ic}, \qquad u^0_i = \max_c p^0_{ic}. \tag{23}$$

For the intervention hypothesis, queried class states retain their posterior scores $\widetilde{p}^I_{ic}$, while the aggregate REST state is assigned to its strongest unqueried class and moderated by the fraction γ_i defined in Eq. 7. Its prediction confidence is therefore

$$u^I_i = \max\left(\left\{\widetilde{p}^I_{ic} : c \in \mathcal{C}_Q\right\} \cup \left\{\gamma_i \widetilde{p}^I_{iR}\right\}\right), \tag{24}$$

with $\hat{y}^I_i$ denoting the prediction of the corresponding mapped class.

Let

$$d_i = \mathbb{1}\left[\hat{y}^I_i \neq \hat{y}^0_i\right], \qquad \delta_i = u^I_i - u^0_i. \tag{25}$$

Source normalization of cross-hypothesis scores. The numerical scales of u^I_i and u^0_i need not coincide because they are produced by different predictive constructions. We therefore do not interpret δ_i as a calibrated difference in probabilities of correctness. Instead, r_{rep} and r_{cor} are observable signed score-contrast proxies. In every outer leave-one-dataset-out fold, these features are oriented and standardized using means and scales estimated only from the source episodes. Their coefficients are then fitted against realized source intervention utility, and the deployment threshold is selected by nested source-only calibration. The resulting downstream mapping is therefore calibrated against source intervention utility; the procedure does not probability-calibrate P^I and P^0 or assume that their raw scores are directly comparable. No target labels are used for normalization, fitting, or threshold selection.

We then define the positive confidence margin associated with the changed predictions as

$$r_{\text{rep}} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} d_i [\delta_i]_+, \tag{26}$$

and the corresponding negative margin as

$$r_{\text{cor}} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} d_i [-\delta_i]_+, \tag{27}$$

where $[z]_+ = \max(z, 0)$. Finally, the overall extent to which the intervention changes the preserved prediction state is

$$r_{\text{dis}} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} d_i. \tag{28}$$

Together, the intervention-authority descriptor is

$$\psi = \left[\widehat{\rho}, \widehat{d}, e_{\text{sup}}, r_{\text{rep}}, r_{\text{cor}}, r_{\text{dis}}\right]. \tag{29}$$

All six statistics are in $[0, 1]$. Their fixed semantic orientations are

$$o = (+1, +1, +1, +1, -1, -1), \tag{30}$$

Table 7: Observable statistics used to estimate intervention authority. The first four statistics provide positive evidence for intervention, whereas the last two penalize potentially unsafe departures from the preserved hypothesis. All quantities are computed without access to unqueried target labels, dataset identity, or target-regime information.

Statistic	Observable evidence	Orientation
$\widehat{\rho}$	Estimated target probability mass explained by the semantic classes revealed through the queried support.	$+$
$\widehat{d}$	Density of revealed semantics relative to the prediction-derived effective complexity of the target batch.	$+$
e_{sup}	Evidence that the preserved hypothesis is inconsistent with the explicitly queried target labels.	$+$
r_{rep}	Positive confidence margin on predictions changed by the intervention; an observable proxy for potentially beneficial repair.	$+$
r_{cor}	Negative confidence margin on changed predictions; an observable proxy for potentially harmful intervention.	$-$
r_{dis}	Fraction of target predictions on which the intervention departs from the preserved hypothesis.	$-$

Table 8: **Ablation of cross-hypothesis confidence-margin features at $M = 4$.** Both rows use identical registered target episodes, U+D supports, intervention hypotheses, and outer folds. The full row reproduces the registered six-feature STTI operating point reported in Table 27; the reduced controller removes r_{rep} and r_{cor} and is independently refitted and recalibrated on the corresponding source folds. Accuracy (%) is averaged over the ten non-ImageNet datasets, and Macro averages all seven regimes.

	Small-batch / sparse			Higher semantic complexity				
Controller	**B64 VL**	**B64 Low**	**B64 Med.**	**B1000 Med.**	**B1000 High**	**B1000 VH**	**ALL**	**Macro**
Full STTI controller (registered)	**87.21**	**78.66**	**72.42**	**74.66**	**74.46**	**73.68**	**74.76**	**76.55**
Without $r_{\mathrm{rep}}, r_{\mathrm{cor}}$	80.17	75.10	71.53	74.25	73.91	73.07	73.55	74.51
Change	-7.04	-3.56	-0.89	-0.41	-0.55	-0.61	-1.21	-2.04

Protocol note. Both variants share the same target episodes, U+D supports, preservation and intervention predictions, and outer folds. Only the features supplied to the intervention-value estimator and the resulting source-fitted coefficients and threshold differ. Every displayed dataset is excluded from feature normalization, value-model fitting, and threshold calibration in its corresponding outer fold. Differences are computed from paired full-precision dataset accuracies before rounding.

corresponding respectively to greater semantic reach, greater information density, stronger evidence against preservation, stronger positive intervention margin, stronger negative intervention margin, and greater unqualified departure from the preserved state. These oriented statistics are standardized using source-episode statistics and supplied to the monotone value estimator in Eq. 9.

Sensitivity to heterogeneous confidence scales. We test whether the controller's decisions depend materially on directly comparing the numerical confidence scales of P^I and P^0. Both controller variants are evaluated on the identical registered target episodes and U+D supports used for the primary $M = 4$ STTI results. In each outer fold, the reduced controller removes the signed confidence-margin features r_{rep} and r_{cor}, retains $\widehat{\rho}$, $\widehat{d}$, e_{sup}, and r_{dis}, and is refitted and recalibrated from scratch using only the remaining source datasets. Table 8 reports the resulting strict outer-LODO comparison.

Result. Removing the two cross-hypothesis confidence-margin features reduces seven-regime macro accuracy from 76.55% to 74.51%, a decrease of $2.04\,\mathrm{pp}$. The largest reductions occur in the Very-Low and Low regimes, indicating that the margins contain material information for deciding when sparse supervision warrants intervention. Nevertheless, these quantities are used only as source-

standardized predictors of intervention utility; their numerical values are not interpreted as calibrated probability differences between P^I and P^0.

C.2 Source-Side Risk Calibration

The intervention-value estimator in Sec. 2.5 produces a point estimate $\widehat{\Delta}(\psi)$ of the expected gain from replacing the preserved hypothesis P^0 with the intervention hypothesis P^I. Because estimation error can make a positive point estimate unreliable, the deployment threshold is calibrated entirely on source episodes using an explicit lower-tail risk constraint.

Source validation cells. Let $\mathcal{V}$ denote the set of source validation cells, where each $v \in \mathcal{V}$ corresponds to one dataset–complexity–annotation-budget configuration. These groupings are used only to evaluate the robustness of candidate thresholds during source-side calibration; neither the cell identity nor the underlying complexity regime is provided to the intervention-value estimator.

Nested calibration and threshold search. For each outer held-out non-ImageNet target, calibration uses only the other nine datasets. We obtain out-of-source calibration scores by nested leave-one-source-dataset-out prediction: each source dataset is scored by a value model fitted on the remaining eight, and the resulting predictions are concatenated. This yields $9 \times 7 \times 3 = 189$ dataset–regime–budget validation cells. For ImageNet, the corresponding procedure uses all ten non-ImageNet sources and 210 validation cells. Candidate thresholds are the unique values from 201 equally spaced empirical quantiles of these nested scores, augmented with $-\infty$, 0, and $+\infty$. After $\tau^\star$ is selected using the risk objective below, the value estimator is refitted on all available outer-source datasets and both the estimator and threshold are frozen.

For source episode e, let

$$\Delta_e = 100\left[\mathrm{Acc}(P_e^I) - \mathrm{Acc}(P_e^0)\right]$$

denote the realized intervention gain in percentage points. For a candidate threshold τ, the corresponding preserve-or-intervene action is

$$a_\tau(\psi_e) = \mathbb{1}\left[\widehat{\Delta}(\psi_e) > \tau\right]. \tag{31}$$

Because preservation leaves P^0 unchanged, its gain relative to the reference hypothesis is zero. The utility of the selective policy on episode e is therefore

$$u_e(\tau) = a_\tau(\psi_e)\,\Delta_e. \tag{32}$$

For validation cell v, containing source episodes $\mathcal{E}_v$, we compute the mean policy utility

$$U_v(\tau) = \frac{1}{|\mathcal{E}_v|}\sum_{e\in\mathcal{E}_v} u_e(\tau). \tag{33}$$

Thus, $U_v(\tau) > 0$ indicates that selective intervention improves upon preservation in that source condition, whereas $U_v(\tau) < 0$ measures the collateral cost of authorizing harmful interventions.

Lower-tail risk constraint. Optimizing only the average of $U_v(\tau)$ can favour thresholds that obtain large gains in easy source conditions while incurring substantial losses in poorly represented ones. We therefore explicitly constrain lower-tail utility. For a collection $\mathcal{U}(\tau) = \{U_v(\tau) : v \in \mathcal{V}\}$, define

$$\mathrm{Tail}_\alpha\big(\mathcal{U}(\tau)\big) = \frac{1}{|\mathcal{V}_\alpha(\tau)|}\sum_{v\in\mathcal{V}_\alpha(\tau)} U_v(\tau), \tag{34}$$

where $\mathcal{V}_\alpha(\tau)$ contains the lowest-performing α fraction of validation cells under threshold τ.

The deployed threshold is selected as

$$\tau^\star = \arg\max_\tau \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} U_v(\tau) \quad \text{s.t.} \quad \mathrm{Tail}_{0.10}\big(\mathcal{U}(\tau)\big) \geq -\epsilon, \tag{35}$$

with $\epsilon = 0.5$ percentage points. Hence, threshold selection seeks the largest average source-side benefit while requiring the mean utility of the worst 10% of validation conditions to remain within 0.5 percentage points of preservation.

Target-time use. Once selected, $\tau^\star$ is frozen together with the intervention transducer and value estimator. No target dataset identity, target-complexity regime, unqueried target label, or target-side threshold tuning is used. For a target episode, intervention is authorized only when

$$\widehat{\Delta}(\psi) > \tau^\star;$$

otherwise STTI retains P^0. The calibration therefore treats preservation as the zero-risk reference action and uses source evidence only to determine how much predicted intervention value is required before departing from it.

D Detailed Per-Dataset Results

Tables 9–11 report the dataset-level results underlying the aggregate comparisons in Sec. 3. We show the $M = 4$ setting, where every label-using method receives exactly four target labels in total per episode.

For Random, Entropy, Margin, k-center, Active-VLM, and U+D, queried labels feed the same frozen Tip-Adapter-style Zhang et al. (2022) cache readout, isolating acquisition differences while keeping the label consumer fixed. STTI uses the same budget but builds a semantic intervention hypothesis from the queried support and decides whether to replace the preserved hypothesis. *Avg* is the unweighted mean over 11 datasets.

Per-dataset transition across semantic complexity. The dataset-level results replicate the main paper's regime-level transition. In Very-Low and Low regimes, strong acquisition followed by the common cache consumer is highly competitive because four queried labels can cover a substantial fraction of the target semantic state. As semantic complexity rises, this advantage steadily declines. From $B = 64$ Medium onward, STTI has the best 11-dataset average among equal-budget methods and stays stable through $B = 1000$ Medium, High, Very-High, and ALL. This shift is not driven by any single benchmark; while effects vary by dataset, the aggregate weakness of mandatory cache adaptation becomes clearer as the unresolved semantic population grows.

Acquisition quality does not determine intervention authority. The U+D+cache control provides the most direct separation between acquisition and intervention because it uses the same uncertainty–diversity query rule as STTI while replacing selective intervention with the common cache consumer. The two approaches are relatively close when the target state is sparse, but their gap grows systematically with semantic complexity. Therefore, selecting informative target samples is not sufficient to explain the robustness of STTI. The critical additional decision is whether the semantic evidence revealed by those same labels is representative enough to justify modifying predictions outside the queried support. These per-dataset results thus support the paper's central distinction between *obtaining informative supervision* and *granting that supervision authority to alter the frozen hypothesis*.

D.1 Acquisition-Policy Transfer with Active-VLM

The $M = 4$ results in Table 1 raise a natural factorization question. Active-VLM followed by the common cache is slightly stronger than STTI in the lowest-complexity sparse regimes, but the two rows differ in both the acquisition rule and the downstream use of the queried labels. We therefore test whether Active-VLM's acquisition rule and STTI's selective intervention mechanism are complementary. We replace only STTI's uncertainty–diversity (U+D) acquisition rule with Active-VLM acquisition; the semantic transducer, intervention-value certificate, source-calibrated threshold, preservation state, annotation budget, and target-task streams remain frozen.

Acquisition-policy robustness rather than a new accuracy gain. Active-VLM+STTI attains 76.85% macro accuracy, compared with 76.55% for U+D+STTI. The paired difference is $+0.30\,\mathrm{pp}$ (dataset-bootstrap 95% CI $[-0.04, +0.77]\,\mathrm{pp}$), with a positive mean difference on 7 of 10 datasets. Thus, changing the acquisition policy does not produce a statistically decisive improvement, but it also does not materially degrade the frozen STTI policy. In contrast, Active-VLM+STTI exceeds Active-VLM+cache by $+5.90\,\mathrm{pp}$ (95% CI $[+2.68, +9.20]\,\mathrm{pp}$), with a positive difference on 9 of 10 datasets.

Table 9: **Detailed per-dataset results for $B = 64$.** Very-Low, Low, and Medium correspond to approximately 1–4, 2–10, and 5–25 effective classes, respectively. Published label-free references use $M = 0$. All label-using methods receive exactly $M = 4$ target labels total. Avg is the unweighted mean over all 11 datasets.

Regime	Method	Avg	ImageNet	SUN397	Aircraft	EuroSAT	Cars	Food101	Pets	Flowers	Caltech	DTD	UCF101
Shared frozen references													
	CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
	MTA	66.6	69.3	64.8	27.4	46.9	68.0	87.2	89.4	71.7	94.0	44.4	69.0
$B = 64$ **Very-Low** ($K_{\text{eff}} \approx$ 1–4)													
Published label-free adaptation ($M = 0$)													
	Dirichlet	68.5	79.2	75.7	28.2	47.2	68.2	88.1	87.5	71.2	88.8	50.3	69.0
	ZLaP	27.5	14.5	13.0	8.4	36.6	23.7	31.9	57.0	22.4	52.4	13.0	29.2
	TransCLIP	38.9	21.6	21.1	11.6	45.1	34.7	59.2	72.4	36.4	62.3	26.1	37.7
	StatA	70.4	72.9	66.0	29.3	56.8	76.2	90.3	95.5	77.6	93.0	46.1	70.2
$M = 4$ *acquisition controls — common frozen Tip-Adapter-style cache readout*													
	Random + cache	86.1	86.5	83.8	73.7	72.1	91.3	94.9	95.7	88.1	97.0	75.4	89.1
	Entropy + cache	84.7	87.5	85.2	68.7	64.5	90.2	95.2	94.3	84.7	98.0	73.5	89.9
	Margin + cache	84.7	87.8	85.4	69.5	64.1	90.1	95.2	94.3	84.5	98.0	73.7	89.7
	k-center + cache	88.0	86.4	83.2	76.6	76.6	93.7	95.5	97.1	90.7	97.7	78.8	91.3
	Active-VLM acquisition + cache†	**90.4**	89.5	88.7	80.2	80.4	95.7	96.4	97.7	93.0	98.1	82.3	93.0
	U+D + cache‡	87.7	87.3	84.6	74.5	71.6	94.1	95.7	97.1	90.9	98.2	78.6	92.4
$M = 4$ *proposed selective intervention — no common cache readout*													
	Selective Test-Time Intervention (Ours)	87.4	89.4	85.6	70.2	77.0	95.7	96.6	98.7	87.3	97.8	71.6	91.3
$B = 64$ **Low** ($K_{\text{eff}} \approx$ 2–10)													
Published label-free adaptation ($M = 0$)													
	Dirichlet	70.3	80.1	78.0	28.1	43.5	71.5	92.3	92.7	74.7	93.0	48.9	70.9
	ZLaP	35.2	19.1	19.0	12.0	46.4	27.9	43.5	66.6	31.3	60.8	22.4	38.7
	TransCLIP	40.4	20.3	22.4	14.3	53.9	30.8	55.6	69.4	40.9	64.6	31.6	40.9
	StatA	69.3	72.8	66.9	27.7	51.3	73.5	89.5	93.7	76.6	93.6	46.9	69.6
$M = 4$ *acquisition controls — common frozen Tip-Adapter-style cache readout*													
	Random + cache	75.5	77.6	74.2	52.8	48.8	81.5	90.5	91.4	80.3	95.1	58.8	79.6
	Entropy + cache	74.3	79.0	76.0	49.0	40.9	81.4	90.9	88.8	76.5	96.3	57.7	81.2
	Margin + cache	74.4	79.3	76.3	49.9	40.8	81.3	91.0	88.8	76.2	96.3	57.7	81.3
	k-center + cache	77.1	77.6	73.7	54.8	55.0	83.1	90.8	93.0	82.1	95.4	61.7	81.2
	Active-VLM acquisition + cache†	**79.8**	80.6	79.0	58.7	58.1	86.0	91.9	93.4	85.6	96.0	65.2	83.3
	U+D + cache‡	77.5	78.8	75.6	54.2	48.7	85.1	91.5	93.4	82.8	96.5	61.9	83.6
$M = 4$ *proposed selective intervention — no common cache readout*													
	Selective Test-Time Intervention (Ours)	79.1	83.5	78.5	49.5	56.4	84.9	95.0	97.3	83.5	96.5	65.2	80.1
$B = 64$ **Medium** ($K_{\text{eff}} \approx$ 5–25)													
Published label-free adaptation ($M = 0$)													
	Dirichlet	67.5	77.7	72.9	26.1	38.6	71.6	90.8	88.4	71.5	93.7	42.9	67.8
	ZLaP	44.6	29.0	27.9	16.5	49.0	36.0	59.1	76.4	42.9	72.0	32.0	50.3
	TransCLIP	42.7	15.5	22.8	17.0	58.2	32.9	56.3	72.6	45.0	65.6	37.5	46.5
	StatA	67.4	70.7	65.3	26.0	45.0	71.1	88.2	90.8	73.7	93.9	47.5	69.1
$M = 4$ *acquisition controls — common frozen Tip-Adapter-style cache readout*													
	Random + cache	67.9	71.2	67.3	35.2	37.8	72.1	87.3	88.0	73.7	93.7	47.9	72.2
	Entropy + cache	67.3	72.5	69.0	33.7	33.0	72.4	87.7	86.3	70.5	94.5	47.1	73.6
	Margin + cache	67.4	72.7	69.1	34.0	32.9	72.4	87.6	86.2	70.5	94.5	47.1	73.6
	k-center + cache	68.8	71.1	66.8	35.6	43.3	72.5	87.4	89.0	74.8	93.8	49.8	72.6
	Active-VLM acquisition + cache†	70.4	73.2	70.1	37.8	45.3	74.7	88.0	88.4	77.7	94.2	50.7	74.5
	U+D + cache‡	69.1	72.2	68.3	36.1	37.7	74.4	88.1	89.2	74.9	94.9	49.9	74.7
$M = 4$ *proposed selective intervention — no common cache readout*													
	Selective Test-Time Intervention (Ours)	**73.1**	79.2	73.4	34.9	43.3	74.3	93.7	95.4	82.0	96.1	63.7	68.1

†Active-VLM denotes its acquisition rule adapted to select exactly four target samples in the finite transductive protocol. ‡U+D denotes uncertainty–diversity acquisition followed by the common cache consumer rather than the STTI intervention mechanism. Avg is computed from full-precision accuracies before rounding.

The regime-wise pattern clarifies the source of this robustness. Across the three $B = 64$ regimes, Active-VLM+STTI and Active-VLM+cache are effectively tied on average ($+0.07$ pp), and the hybrid remains 2.41 pp below the cache in Very-Low. As semantic complexity increases, the frozen controller rapidly reduces intervention, from 60.7% in Very-Low to approximately 1–2% in the $B = 1000$ regimes and 0% in ALL. The large dense-regime gains over Active-VLM+cache therefore reflect STTI's ability to withhold a harmful sparse-label intervention and retain the preservation hypothesis, rather than a uniformly stronger acquisition rule.

Takeaway. The Active-VLM swap supports STTI's modularity: its source-frozen preserve–intervene decision transfers to a different acquisition policy and maintains the accuracy of the original U+D configuration. The experiment is evidence of acquisition-policy robustness, not evidence that Active-VLM acquisition significantly improves STTI.

Table 10: **Detailed per-dataset results for $B = 1000$.** Medium, High, and Very-High correspond to approximately 5–25, 25–50, and 50–100 effective classes, respectively. Published label-free references use $M = 0$. All label-using methods receive exactly $M = 4$ target labels total. Avg is the unweighted mean over all 11 datasets.

Regime	Method	Avg	ImageNet	SUN397	Aircraft	EuroSAT	Cars	Food101	Pets	Flowers	Caltech	DTD	UCF101
Shared frozen references													
	CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
	MTA	66.6	69.3	64.8	27.4	46.9	68.0	87.2	89.4	71.7	94.0	44.4	69.0
$B = 1000$ **Medium** ($K_{\text{eff}} \approx 5$–25)													
Published label-free adaptation ($M = 0$)													
	Dirichlet	64.4	60.9	75.4	26.7	38.8	74.1	76.2	91.0	71.6	92.4	36.2	65.4
	ZLaP	41.5	16.6	20.1	16.4	49.0	32.2	55.5	76.4	40.6	67.7	34.2	48.1
	TransCLIP	56.5	39.9	42.7	22.0	63.1	49.9	80.6	87.9	58.7	79.1	42.9	55.0
	StatA	69.7	70.8	64.5	28.4	60.4	74.0	87.5	93.1	77.5	92.8	47.1	70.2
$M = 4$ *acquisition controls — common frozen Tip-Adapter-style cache readout*													
	Random + cache	67.0	70.3	66.6	32.2	37.9	70.6	87.4	87.7	72.9	94.1	45.9	71.2
	Entropy + cache	65.9	70.4	66.7	31.6	31.0	70.3	87.7	86.1	69.7	94.7	44.6	72.0
	Margin + cache	66.8	71.8	68.0	31.3	34.3	71.5	87.7	87.1	70.1	94.8	45.7	72.3
	k-center + cache	67.2	69.4	65.3	31.3	40.7	70.2	87.2	89.2	73.2	94.0	47.6	71.4
	Active-VLM acquisition + cache[†]	68.6	71.4	68.6	32.9	41.6	72.2	88.2	89.0	76.1	94.3	47.5	72.8
	U+D + cache[‡]	68.2	70.4	66.4	32.5	41.7	72.5	88.0	89.5	74.0	95.0	47.0	73.3
$M = 4$ *proposed selective intervention — no common cache readout*													
	Selective Test-Time Intervention (Ours)	**75.1**	79.8	74.8	35.0	59.4	75.8	94.0	97.1	82.1	96.6	65.9	65.9
$B = 1000$ **High** ($K_{\text{eff}} \approx 25$–50)													
Published label-free adaptation ($M = 0$)													
	Dirichlet	45.3	17.3	37.3	21.0	37.9	65.4	46.3	81.3	46.3	80.5	21.1	43.6
	ZLaP	52.2	23.8	32.2	22.2	49.3	45.4	74.9	86.5	56.2	79.7	43.6	60.8
	TransCLIP	62.0	43.9	49.6	24.8	64.0	57.3	83.0	91.4	69.1	85.5	47.5	65.4
	StatA	69.7	71.9	66.4	25.9	60.7	73.6	88.0	91.4	76.7	93.2	47.9	71.5
$M = 4$ *acquisition controls — common frozen Tip-Adapter-style cache readout*													
	Random + cache	64.2	68.3	64.0	25.5	36.3	67.0	85.8	86.0	69.7	93.3	42.0	68.3
	Entropy + cache	63.5	68.3	64.0	26.1	30.8	66.9	85.9	84.6	67.6	93.3	42.1	68.4
	Margin + cache	63.9	69.0	64.5	25.4	33.1	67.0	85.8	85.7	68.4	93.3	42.1	68.5
	k-center + cache	64.8	67.8	63.3	25.6	39.2	66.9	86.0	87.7	70.1	93.4	44.1	68.5
	Active-VLM acquisition + cache[†]	64.2	68.8	64.6	24.2	38.6	66.7	85.5	85.2	70.0	93.4	39.9	68.8
	U+D + cache[‡]	65.1	68.3	63.9	26.1	40.7	67.5	86.2	87.2	70.2	93.9	43.5	69.0
$M = 4$ *proposed selective intervention — no common cache readout*													
	Selective Test-Time Intervention (Ours)	**75.0**	80.3	75.0	34.2	59.5	74.5	94.2	95.8	81.6	96.5	67.3	66.1
$B = 1000$ **Very-High** ($K_{\text{eff}} \approx 50$–100)													
Published label-free adaptation ($M = 0$)													
	Dirichlet	33.6	10.8	15.7	17.5	37.8	51.2	29.1	79.3	24.3	59.1	19.0	26.1
	ZLaP	58.4	32.7	44.0	25.4	49.3	55.2	83.3	87.3	64.8	87.9	45.2	67.8
	TransCLIP	64.4	44.5	53.0	25.6	64.1	60.9	85.2	91.9	74.3	90.5	48.1	70.7
	StatA	69.0	71.8	67.1	23.9	60.7	70.2	87.1	91.1	74.3	93.7	48.0	70.7
$M = 4$ *acquisition controls — common frozen Tip-Adapter-style cache readout*													
	Random + cache	63.3	67.4	63.2	22.8	36.3	65.5	85.4	85.9	68.6	93.1	41.2	67.1
	Entropy + cache	62.7	67.4	63.3	23.5	30.6	65.5	85.5	84.6	67.6	92.8	41.8	66.9
	Margin + cache	62.9	67.6	63.4	22.6	33.2	65.2	85.3	85.6	67.8	92.6	41.5	66.8
	k-center + cache	64.1	67.1	62.9	23.7	39.2	65.6	85.6	87.7	69.0	93.4	43.4	67.5
	Active-VLM acquisition + cache[†]	62.9	67.5	63.4	21.2	38.6	64.8	84.8	84.9	67.3	93.3	38.7	66.9
	U+D + cache[‡]	64.2	67.4	63.2	23.4	40.5	65.6	85.6	87.0	69.1	93.5	43.0	67.4
$M = 4$ *proposed selective intervention — no common cache readout*													
	Selective Test-Time Intervention (Ours)	**74.2**	79.9	74.8	32.2	59.6	70.6	93.8	95.6	80.9	96.3	67.5	65.2

[†] Active-VLM denotes its acquisition rule adapted to select exactly four target samples in the finite transductive protocol. [‡] U+D denotes uncertainty–diversity acquisition followed by the common cache consumer rather than the STTI intervention mechanism. Avg is computed from full-precision accuracies before rounding.

Table 11: **Detailed per-dataset results in the ALL setting.** All semantic classes are present in the target task. Published label-free references use $M = 0$. All label-using methods receive exactly $M = 4$ target labels total. The ALL setting uses one full-dataset target task per dataset. Avg is the unweighted mean over all 11 datasets.

Regime	Method	Avg	ImageNet	SUN397	Aircraft	EuroSAT	Cars	Food101	Pets	Flowers	Caltech	DTD	UCF101
Shared frozen references													
	CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
	MTA	66.6	69.3	64.8	27.4	46.9	68.0	87.2	89.4	71.7	94.0	44.4	69.0
ALL classes													
Published label-free adaptation ($M = 0$)													
	Dirichlet	29.5	66.6	0.7	15.5	39.9	21.2	8.5	76.3	14.3	44.0	19.4	18.4
	ZLaP	66.4	68.8	67.2	26.9	49.1	67.1	86.9	87.8	68.7	90.8	45.5	71.4
	TransCLIP	70.3	70.4	68.9	26.9	66.1	69.5	87.1	92.5	76.5	92.7	48.6	74.1
	StatA	69.9	69.9	68.7	24.7	67.3	68.0	87.1	92.4	75.2	94.2	48.4	73.5
$M = 4$ acquisition controls — common frozen Tip-Adapter-style cache readout													
	Random + cache	63.0	66.6	62.5	22.1	37.6	64.4	85.3	85.9	68.1	93.2	41.1	66.7
	Entropy + cache	62.4	66.7	62.6	23.1	26.8	64.8	85.9	84.8	69.5	92.9	42.4	67.4
	Margin + cache	61.9	66.6	62.5	22.4	27.8	64.0	85.1	85.9	66.0	92.8	41.1	67.1
	k-center + cache	64.1	66.7	62.6	23.3	41.1	65.4	85.8	88.0	68.0	93.5	43.0	67.1
	Active-VLM acquisition + cache[†]	62.2	66.5	62.0	21.5	37.7	62.5	84.3	85.4	66.0	93.3	38.0	66.9
	U+D + cache[‡]	64.1	66.7	62.6	23.2	43.1	65.0	85.9	87.4	68.2	93.5	42.0	67.3
$M = 4$ proposed selective intervention — no common cache readout													
	Selective Test-Time Intervention (Ours)	**75.0**	77.8	76.5	33.7	67.5	66.9	94.0	96.7	81.0	96.5	68.5	66.2

[†] Active-VLM denotes its acquisition rule adapted to select exactly four target samples in the finite transductive protocol. [‡] U+D denotes uncertainty–diversity acquisition followed by the common cache consumer rather than the STTI intervention mechanism. The ALL setting contains one full-dataset target task per dataset rather than repeated sampled tasks. Avg is computed from full-precision accuracies before rounding.

Table 12: **Acquisition-policy transfer with Active-VLM at $M = 4$.** Active-VLM+cache uses Active-VLM acquisition followed by the common frozen cache consumer. U+D+STTI is the original frozen STTI configuration, whereas Active-VLM+STTI changes only the acquisition rule. Results are unweighted means over the same 10 non-ImageNet datasets. Sparse Avg. averages the three $B = 64$ regimes and Macro averages all seven regimes.

	Small-batch / sparse regimes				Higher semantic complexity				
Method	**B64 VL**	**B64 Low**	**B64 Med.**	**Sparse Avg.**	**B1000 Med.**	**B1000 High**	**B1000 VH**	**ALL**	**Macro**
Active-VLM acquisition with the common cache consumer									
Active-VLM + cache	**90.55**	**79.77**	70.12	80.15	68.19	63.70	62.41	61.91	70.95
Frozen STTI under alternative acquisition policies									
U+D + STTI	87.21	78.66	72.42	79.43	**74.66**	**74.46**	**73.68**	**74.76**	76.55
Active-VLM + STTI	88.14	79.61	**72.92**	**80.22**	74.60	74.38	73.58	**74.76**	**76.85**
Paired differences and STTI deployment behavior									
Δ vs. Active-VLM + cache (pp)	-2.41	-0.16	+2.79	+0.07	+6.40	+10.68	+11.17	+12.85	+5.90
Δ vs. U+D + STTI (pp)	+0.93	+0.95	+0.50	+0.79	-0.07	-0.08	-0.10	0.00	+0.30
Active-VLM+STTI intervention rate	60.7%	30.1%	11.0%	33.9%	2.0%	1.4%	1.5%	0.0%	15.2%

All methods use exactly four queried target labels per cross-fitted episode. The 10-dataset subset excludes ImageNet; consequently, values differ slightly from the 11-dataset aggregates in Table 1. Sparse Avg. and Macro are computed from full-precision accuracies before rounding. Positive deltas favour Active-VLM+STTI.

E MECHANISTIC EVIDENCE AND EXTENDED ANALYSIS

Scope and reporting conventions. This appendix analyzes how STTI works and when sparse target supervision enables or fails semantic intervention. It addresses four questions: (i) what semantic structure sparse labels reveal, (ii) why intervention utility varies with target complexity, (iii) whether semantic coverage explains the transition, and (iv) whether remaining failures can be resolved using observable target evidence.

Unless stated otherwise, gains are absolute top-1 accuracy differences relative to the frozen transductive StatA reference. The budget $M \in \{1, 2, 4\}$ is the *total* number of queried target labels per episode. Hidden-label quantities and oracles are for diagnosis only and are never available to the deployed STTI policy. The ALL setting includes one full-dataset target task per benchmark; sampled regimes vary target batch size B and effective semantic complexity K_{eff}.

E.1 SPARSE SUPERVISION RESOLVES SEMANTIC CORRESPONDENCE

We first ask what sparse target labels actually reveal. A useful correction may admit a compact response structure yet be unidentifiable from unlabeled target statistics. Table 13 separates these issues: the intervention response is highly structured and largely low-dimensional, but the label-free orientation decision is near chance. Thus, low-dimensional representability does not guarantee selecting the correct target-specific correction without semantic supervision.

The queried labels provide precisely this missing information. If they served only as a small conventional training set, their influence should remain local to annotated examples. Instead, the effect is population-structured: queried classes improve substantially, while unconditional propagation can harm fully unqueried classes (Table 14). Revealing queried semantic correspondences also makes support-to-target membership highly predictable. Thus, sparse labels uncover shared target population–class relations, not isolated supervision points.

Table 13: Low-dimensional response structure does not imply label-free identifiability. Oracle corrections are large in a compact subspace, while deployable orientation remains weak.

Diagnostic	**Result**
Exactly unimodal response curves	≈ 94–95%
Response-curve R^2	≈ 0.97–0.98
Rank-1 / 2 / 3 response energy	$69.2/84.0/86.5\%$
Rank-2 hidden-label oracle gain	$+7.32$ pp
Rank-3 hidden-label oracle gain	$+8.19$ pp
Deployable low-rank combination	$+0.28$ pp
Label-free orientation agreement	$\approx 48.5\%$

Table 14: Evidence that sparse target supervision identifies semantic populations rather than isolated samples. Hidden-label quantities are diagnostic only.

Diagnostic	**Result**
$M = 4$ gain on queried semantic classes	$+21.54$ pp
$M = 4$ gain on fully unqueried semantic classes	-9.96 pp
Changed-sample repair classifier AUROC	0.883 $[0.862, 0.903]$
Changed-sample average precision	0.953
Support-to-target membership AUROC	0.991
Queried-class/REST hidden-label oracle	$\approx +20.09$ pp

Diagnostic-classifier protocols. The learned diagnostics in Table 14 use the 300 registered $B = 64$ tasks and the strict leave-one-dataset-out (LODO) evaluation. For support-to-target membership, each example is a queried-support/target pair whose label is one iff the revealed support class equals the target's hidden class. We used the 44 observable affinity, cache-evidence, text–class,

competitor-geometry, and population-context features. The features are standardized on the nine source datasets, and a logistic regression ($C = 1$, `liblinear`, 500 iterations, tolerance 10^{-6}) is fitted before scoring the held-out dataset. For changed-sample repair, examples are restricted to targets for which the candidate intervention changes the StatA prediction. A positive example is StatA-wrong/intervention-correct and a negative example is StatA-correct/intervention-wrong; accuracy-neutral changes are excluded from AUROC computation. A source-standardized logistic regression model ($C = 1$, `lbfgs`, 2000 iterations) is fitted to the 58 observable post-query relational features, with inverse episode-size and class-balancing weights. These comprise support-affinity (8), cache-evidence (7), StatA-state (6), CLIP-state (7), reciprocal-neighborhood (7), local-density (7), support–text (7), and change-geometry (9) statistics. The model and threshold selection use nested source-only validation. Unqueried target labels define diagnostic outcomes only and are never model inputs or used for held-out dataset selection.

Takeaway. Sparse labels reveal shared semantic correspondences: a few annotations can identify target populations whose influence extends well beyond the queried examples.

E.2 The value of intervention undergoes a complexity-dependent transition

The same sparse intervention is not uniformly beneficial in all target states. Keeping the intervention mechanism fixed while increasing the target semantic complexity exposes a systematic crossover (Table 15). Intervention produces large positive gains in Very-Low and Low complexity, weakens through the Medium regimes, and becomes harmful once the same fixed annotation budget is diluted over a sufficiently diverse target state. Increasing M shifts the crossover towards more complex regimes but does not eliminate it.

The transition cannot be explained by K_{eff} alone. In particular, the two Medium regimes contain comparable semantic complexity but different batch sizes, and the intervention is substantially weaker at $B = 1000$. This indicates that the relevant quantity is not only how many semantic classes are active but how much of the target population can be represented by the fixed queried-label budget.

The repair–corruption decomposition in Table 16 further localizes the transition. From $B = 1000$ Medium to High, the exposure to corruption changes only modestly, whereas the available repair mass collapses. The intervention therefore becomes harmful primarily because the queried labels cease to identify enough correctable target population to offset collateral changes, rather than because the intervention suddenly becomes more aggressive.

Takeaway. The crossover is consistent with an identifiability bottleneck: as a fixed label budget covers less of the target semantic state, the available supervision identifies progressively less correctable population mass.

Table 15: Complexity-dependent value of sparse semantic intervention. Entries are gains relative to StatA (pp); StatA is reported as absolute accuracy for scale. The intervention mechanism is fixed across regimes.

Target regime	StatA	$\mathbf{M = 1}$	$\mathbf{M = 2}$	$\mathbf{M = 4}$
$B = 64$, Very-Low	69.628	+12.681	+17.825	**+22.141**
$B = 64$, Low	68.932	+4.413	+8.079	**+12.540**
$B = 64$, Medium	67.321	+0.889	+1.864	+4.068
$B = 1000$, Medium	69.507	+0.283	+0.736	+1.792
$B = 1000$, High	69.556	−0.746	−1.236	−1.674
ALL	69.976	−0.703	−1.892	−2.830

Protocol note. These rows evaluate the candidate intervention hypothesis P^I applied unconditionally on the cross-fitted mechanistic diagnostic stream. They therefore characterize intervention utility and are not the final selective STTI predictions $P^\star$ reported in Table 1. The outer-LODO preserve–intervene decomposition in Table 27 uses a separate registered evaluation stream, so its absolute P^I values are not expected to coincide.

Table 16: Repair–corruption anatomy at $M = 4$ around the Medium→High transition.

Regime	Net gain	Repair opp.	Corruption
$B = 1000$, Medium	+1.79 pp	≈ 4.52 pp	≈ 2.73 pp
$B = 1000$, High	−1.67 pp	≈ 1.21 pp	≈ 2.88 pp

Protocol note. At $M = 4$, the repair/corruption ratio falls from 24.3× (Very-Low) to 5.78× (Low), 2.19× ($B = 64$ Medium), 1.66× ($B = 1000$ Medium), and 0.42× (High).

E.3 Controlling semantic coverage isolates the dense-regime bottleneck

The preceding results point to a more specific variable: the fraction of the evaluation population whose true class appears among the queried classes. We call this diagnostic semantic coverage ρ. To test whether insufficient covered mass is a principal cause of dense-regime failure, we apply a restricted hidden-label diagnostic that keeps the learned ranking, queried-class identities, and REST predictions unchanged. It reveals only the total number of evaluation targets whose ground-truth classes are in $\mathcal{C}_Q$ and uses this count to set the intervention quota.

Exact-mass implementation. For the cross-fitted evaluation fold e, the oracle reveals only the scalar quota

$$q_e = \sum_{i \in \mathcal{E}_e} \mathbb{1}[y_i \in \mathcal{C}_Q], \qquad s_i = \sum_{c \in \mathcal{C}_Q} \widetilde{p}^I_{ic},$$

where $\mathcal{E}_e$ is the evaluation set and s_i is the frozen queried-class posterior mass. We rank the evaluation targets by decreasing s_i, break ties by the fixed sample index, and select exactly the top q_e targets. The selected targets retain the frozen queried-class prediction; all remaining targets retain the prediction of the frozen REST. Hidden labels therefore determine only the aggregate quota, not sample identities, ranking scores, queried-class identities, or either branch's class prediction.

This intervention changes only the available aggregate-mass information, yet recovers 98.7%, 95.0% and 93.0% of the ALL intervention deficit for $M = 1, 2, 4$, respectively (Table 17). At $M = 4$, the remaining deficit -0.199 pp has a dataset-bootstrap 95% confidence interval that includes zero. Thus, aggregate semantic coverage explains most of the dense-regime deficit without altering the semantic predictor, queried-class identities, or target ranking. The remaining question is whether the covered mass can be localized reliably from observable evidence.

We score queried-population membership using s_i above; localization AUROC measures its ranking quality, while exact-quota precision is the fraction of the top q_e ranked targets whose true classes belong to $\mathcal{C}_Q$.

The aggregate mass does not fully determine the intervention because that mass must still be localized within a low-prevalence target population. Table 18 isolates this residual gap. Once true queried-population membership is supplied, the existing queried-class predictor is already strong, and

Table 17: Effect of controlling aggregate semantic intervention mass on ALL. The diagnostic reveals only the number of evaluation targets whose true classes belong to the queried set; sample identities and class assignments remain hidden, while target ranking and the within-branch queried-class and REST predictions remain fixed.

M	True mass ρ	Current gain	Exact-mass gain	Deficit recovered	Repairs	Corruptions
1	0.0174	-0.7032	-0.0094	98.7%	0.2018	0.9050
2	0.0417	-1.8917	-0.0955	95.0%	0.7346	2.6263
4	0.0847	-2.8296	-0.1986	93.0%	1.0691	3.8986

Protocol note. At $M = 4$, the exact-mass gain has 95% dataset-bootstrap CI $[-0.5953, +0.1672]$ pp. The oracle is diagnostic and is not part of the deployable method.

Table 18: Coverage anatomy at $M = 4$ on ALL. Successively stronger label-revealed diagnostics isolate the remaining information gap after aggregate mass is fixed.

Diagnostic	Result
Current intervention	-2.830 pp
Exact aggregate mass only	-0.199 pp
Exact membership + frozen queried identity	$+2.954$ pp
Exact membership + oracle queried identity + frozen REST	$+3.806$ pp
Localization AUROC	0.9319
Exact-quota localization precision	0.6322
Queried identity accuracy on truly covered targets	0.9068
Frozen REST accuracy on uncovered targets	0.7196
StatA accuracy on uncovered targets	0.7056

intervention becomes beneficial. The residual dense-regime difficulty is therefore primarily *localizing* the covered population rather than identifying its queried semantic class.

Takeaway. Semantic coverage is a principal information bottleneck: supplying only the correct aggregate covered mass removes most of the dense intervention deficit without changing the learned semantic predictor; the remaining error is dominated by localizing that covered population.

E.4 ALTERNATIVE DENSE-REGIME EXPLANATIONS ARE INSUFFICIENT

We next test whether dense-regime failure stems from simpler implementation limits. We examine stronger or distribution-matched membership ranking, global low-rank correction, class-agnostic visual population recovery, posterior-weighted class moments, graph propagation, and a richer per-sample action space. Table 19 reports the key diagnostic for each hypothesis.

The results reject a single, simple capacity explanation. Some alternatives have low oracle ceilings even with hidden-label help; others show substantial oracle headroom but fail when the required state must be inferred from observable target evidence. Stronger ranking or richer actions still do not close the gap when queried-population membership is uncertain. Thus, the dense failure is not fixable simply by increasing localizer capacity or adding another geometric propagation mechanism.

Implementation of the falsification controls. All controls in Table 19 reuse frozen ALL features, StatA posteriors, cross-fitting folds, and registered supports; hidden labels are accessed only by rows explicitly marked as oracles. *Distribution-matched ranking* fits, under strict LODO, a $273 \to 64 \to 1$ scorer to 108 source ALL episodes per fold using source-standardized observable features, 30 epochs of AdamW (learning rate 10^{-3}, weight decay 10^{-4}), and binary-cross-entropy plus an equal-weight hard-negative ranking loss with unit margin. The exact target quota is supplied only after scoring. The *global low-rank correction* uses the first $r \in \{1, 2, 3\}$ singular modes of the observable Jacobian StatA

Table 19: Falsifying plausible alternative explanations for dense-regime failure. Hidden-label rows are diagnostic ceilings; observable rows remain strictly within the stated label budget.

Hypothesis	Decisive result	Implication
More / distribution-matched membership ranking	Distribution-matched $M = 4$ ranking: -0.408 pp (AUROC 0.898); exact membership: $+2.954$ pp	Generic ranking quality is not sufficient at the extreme low-prevalence decision tail.
Shared global low-rank correction	Rank-3 hidden oracle: $+1.149$ pp; four-label estimator: -0.137 pp	The dense correction is not dominated by one shared low-rank shift with meaningful ceiling.
Class-agnostic population recovery	Spherical k-means: -22.431 pp; source-trained Fisher k-means: -14.767 pp (ARI 0.523)	Better clustering does not recover the class-conditioned target populations required for semantic naming.
Posterior-weighted conditional moments	True class moment: $+14.660$ pp; observable soft moment: approximately -2.23 pp	The first moment is valuable, but the frozen posterior does not supply sufficiently clean class responsibilities.
Local graph propagation	-0.078 pp despite $66.32\times$ homophily enrichment; target-label graph-hyperparameter oracle: only $+0.347$ pp	Local semantic coherence is real but is not corrective information for the remaining errors.
Richer per-sample action set	Exact membership: $+2.954$ pp; safe three-action oracle: $+3.002$ pp	Adding a third action contributes only $+0.048$ pp beyond membership; the missing information is not primarily action richness.

response ; the hidden oracle selects their dataset-wide coefficients using the accuracy of the target, while the deployable control estimates them using hierarchical-prior MAP from the M queried labels. *Population recovery* applies spherical k-means with the known class count K to normalized CLIP features (four k-means++ restarts, 50 iterations, tolerance 10^{-5}) and names clusters by Hungarian matching to frozen text prototypes. Its Fisher control first learns a 256-dimensional LDA projection on the other nine datasets, with automatic covariance shrinkage and at most 16 samples per source class. *Conditional moments* use $w_{ic}(\tau) \propto \exp(\log p^0_{ic}/\tau)$, $m_c = \mathrm{norm}(\sum_i w_{ic} x_i)$, and prototype $\mathrm{norm}((1-\alpha)t_c + \alpha m_c)$; the observable row fixes $\tau = \alpha = 1$, while the true-moment oracle replaces w_{ic} by class indicators and selects one dataset-wide α. *Graph propagation* minimizes KL divergence to the frozen StatA field plus weighted ℓ_1 total variation on a mutual-kNN CLIP graph, with $k \in \{5, 10, 20, 40\}$ and $\lambda \in \{0.003, 0.01, 0.03, 0.1, 0.3\}$; queried supports are clamped to one-hot labels, and deployable parameters are selected by strict LODO. Finally, the *safe three-action oracle* chooses per target among the three already-frozen StatA, REST, and queried-class predictions, preferring StatA when correct, then REST, and using the queried-class prediction only when it uniquely repairs the remaining error. It fits no model and changes no prediction scores.

E.5 COMPLEMENTARY FROZEN EVIDENCE MOTIVATES THE PRESERVATION STATE

Selective intervention matters only if there is a strong state to preserve. We assess complementarity between independent frozen VLM views via equal-probability fusion (Table 20). This diagnostic is not the preservation rule implemented in Sect. 2.2; rather, it tests whether combining frozen semantic evidence yields a stronger label-free reference than either view alone.

Across the ten locally available ALL datasets, the equal-weight fusion reaches 73.35% macro accuracy, $+3.37$ pp above StatA. The result therefore establishes genuine frozen-view complementarity independently of the sparse-label intervention mechanism.

The existence of complementary frozen evidence does not mean global reliability can be safely inferred from four target labels. At $B = 1000$ High, a fixed source-based reliability rule outperforms StatA, and a hidden-label one-scalar oracle shows additional low-dimensional headroom. But estimating the same reliability state from four actively selected supports increases corruption and

Table 20: Frozen semantic complementarity on the ten locally available ALL datasets. No target labels are used.

Frozen prediction state	Accuracy	Δ StatA
StatA	69.97%	+0.00 pp
CLIP zero-shot	65.11%	−4.87 pp
SigLIP2 zero-shot	56.50%	−13.47 pp
Equal-probability CLIP–SigLIP2 fusion	73.35%	+3.37 pp

Protocol note. A best-expert-per-task hidden-label oracle reaches 76.70%, exposing +3.35 pp of selection headroom above the best fixed expert. ImageNet is not included in this local evaluation.

Table 21: Dense global reliability at $B = 1000$ High and $M = 4$. The first row is label-free; the second estimates task-wide reliability from four queried labels; the final row is a one-scalar hidden-label oracle.

Reliability rule	Δ StatA	Pos. data	Worst	Never-queried	Repairs	Corrupt.
Fixed source-derived prior	+5.014	8/10	−5.213	+4.956	6.963	1.949
Four-label global calibration	+3.885	8/10	−15.347	+3.370	7.393	3.508
Per-task one-scalar oracle	+7.424	10/10	+0.239	+7.211	9.138	1.714

Protocol note. On EuroSAT, StatA / fixed prior / four-label calibration / oracle obtain 60.93% / 60.57% / 45.58% / 61.17%, respectively. The preregistered sparse calibration hypothesis failed on High.

can fail badly (Table 21). This aligns with acquisition-biased identification of a task-wide reliability parameter, not with a lack of useful frozen-view complementarity.

Takeaway. Preservation is not an unrelated auxiliary mechanism. It provides the default state when sparse supervision does not contain enough information to identify a target-specific intervention safely.

E.6 Selective semantic intervention unifies preservation and adaptation

The preceding analyses reduce the problem to an episode-level decision: preserve the strongest frozen semantic hypothesis unless observable evidence indicates that the queried supervision identifies enough target structure to justify intervention. The controller is deliberately blind to dataset identity, experimental regime, configured K_{eff}, and unqueried target labels. It estimates only the expected value and risk of applying the candidate intervention to the current episode.

The behavior reported in Sec. 3 follows this mechanism: intervention is frequent in concentrated target states and progressively vanishes as the same annotation budget becomes less representative. Thus, STTI does not encode a regime-specific rule; preservation and intervention are two actions of the same source-calibrated inference policy.

The remaining gap cannot be removed simply by relaxing a single decision threshold. Table 22 shows a continuous trade-off between retaining sparse-regime intervention gains and protecting the frozen hypothesis in dense regimes. More permissive operating points recover additional sparse benefit but also admit larger worst-case losses relative to preservation. The selected operating point therefore reflects the lower-tail source-side risk constraint of Appendix C.2, rather than target-side threshold tuning.

Takeaway. STTI is one inference policy with two actions: preserve or intervene. Which action is selected depends on the observable value and risk of the same sparse semantic evidence, not on a predefined benchmark regime.

Table 22: Frozen safety–retention frontier. All operating points were registered before evaluation.

Floor	Macro	Gap	VL ret.	Low ret.	Dense min.
-0.25	$+6.81$	-1.42	61.0%	49.7%	-0.03
-0.50	$+7.26$	-0.97	73.6%	64.3%	-0.11
-0.75	$+7.41$	-0.82	79.8%	72.0%	-0.38
-1.00	$+7.54$	-0.69	83.5%	76.9%	-0.38
-1.50	$+7.61$	-0.62	87.2%	83.0%	-0.73
-2.00	$+7.72$	-0.51	89.7%	87.7%	-0.73

Protocol note. The highlighted -0.50 pp row is the registered primary. "Gap" is relative to the regime-wise global-action envelope; "Dense min." is the worst loss relative to the frozen prior across High, Very-High, and ALL.

E.7 Fine-grained harm identification remains domain dependent

Could episode-level selection be replaced by a per-sample veto? We test this on targets for which intervention and preservation predict different labels. The changed set is strongly asymmetric: 58.67% of disagreements are beneficial for intervention, 21.40% are harmful, and 19.93% are incorrect under both hypotheses. A generic "preserve on disagreement" rule would therefore discard substantially more repairs than corruptions.

The remaining question is whether harmful overrides can be recognized from observable frozen-view signals with sufficient cross-domain reliability. The additional StatA/SigLIP/fused-view features reach only 0.736 macro leave-one-dataset-out AUROC, below the existing confidence control at 0.829 (Table 23). The large dataset-to-dataset variation in Table 24 shows that these signals do not have a sufficiently invariant interpretation across unseen domains to support a transferable sample-level veto.

Table 23: Changed-sample harm identification at $M = 4$. A changed sample is one for which the sparse intervention and frozen prior predict different labels.

Quantity	Result
Changed samples	20,281
Beneficial intervention among changed samples	58.67%
Harmful intervention among changed samples	21.40%
Both hypotheses wrong	19.93%
New frozen-view harm features: macro LODO AUROC	$0.736\ [0.620, 0.851]$
Existing confidence control: macro LODO AUROC	0.829
New $-$ control AUROC	$-0.092\ [-0.195, +0.005]$
Datasets with new AUROC ≥ 0.78	5/10

Table 24: Cross-dataset transfer of changed-sample harm information. The same observable features range from near-random to excellent depending on the held-out domain.

Dataset	Harm	New AUROC	Control
Aircraft	13.53%	0.722	0.764
Caltech101	36.86%	0.551	0.857
Cars	9.15%	0.578	0.864
DTD	35.09%	0.868	0.846
EuroSAT	24.29%	0.510	0.662
Flowers	25.66%	0.974	0.905
Food101	47.55%	0.909	0.882
Oxford Pets	62.55%	0.925	0.774
SUN397	15.17%	0.875	0.932
UCF101	5.78%	0.454	0.801

Protocol note. The additional harm features did not satisfy the pre-specified cross-domain transfer criterion; consequently, no sample-level veto is included in STTI.

Limitation. The residual selective-policy gap is not explained by a single decision threshold. The risk frontier and cross-domain harm analysis indicate that fine-grained intervention harm remains insufficiently invariant under the current observable evidence.

E.8 Reporting boundaries and interpretation

The analyses in this appendix serve distinct scientific roles. Observable experiments evaluate mechanisms that could in principle support deployment; hidden-label diagnostics instead isolate information bottlenecks or quantify attainable ceilings and are never used by STTI at inference. Sensitivity and falsification analyses test whether the proposed explanation survives plausible alternative mechanisms. We therefore avoid interpreting diagnostic oracle performance as achievable test-time accuracy.

Taken together, the evidence supports a narrower but stronger conclusion than the claim that every episode admits a perfectly observable intervention decision. Sparse target labels are highly valuable when they identify a shared semantic population, their authority declines as the same budget covers less of the target state, and the remaining failure modes increasingly concern identification of the population over which intervention is valid. This motivates treating sparse supervision as evidence for an optional semantic intervention rather than as an unconditional instruction to adapt.

Table 25: **Cross-VLM sparse-side mechanism test at $M = 4$.** The same Tip-Adapter-style cache intervention is used for all three primary-VLM configurations, isolating whether sparse semantic intervention opportunity transfers independently of the final CLIP-trained STTI operator. Coverage is the fraction of evaluation targets whose ground-truth class belongs to the queried-class set. Confidence intervals are paired task-level 95% intervals for Intervene−Preserve.

Primary configuration	Regime	Coverage	Preserve	Cache Intervene	Δ	95% CI	Repairs / Corruptions
OpenAI CLIP	Very-Low	89.7%	72.90	88.48	+15.57	[+11.10, +20.23]	18.85 / 3.28
	Low	61.4%	74.64	80.47	+5.83	[+3.11, +8.66]	11.39 / 5.56
	Medium	31.0%	68.16	68.23	+0.08	[−1.84, +2.00]	7.95 / 7.88
DataComp OpenCLIP	Very-Low	90.0%	79.13	92.58	+13.46	[+9.79, +17.55]	15.98 / 2.52
	Low	63.6%	84.11	88.38	+4.27	[+1.61, +7.22]	8.06 / 3.80
	Medium	33.8%	79.14	80.30	+1.16	[−1.41, +3.56]	6.97 / 5.81
SigLIP2	Very-Low	91.6%	65.41	86.25	+20.83	[+15.31, +26.78]	23.71 / 2.88
	Low	64.4%	64.70	75.98	+11.28	[+6.62, +16.12]	17.48 / 6.20
	Medium	32.1%	64.81	62.89	-1.92	[−6.00, +2.00]	9.91 / 11.83

The gate uses the same total budget of four target labels for every configuration and a common cache-based intervention construction. It is a preregistered sparse-side mechanism test, not the final Selective TTI controller evaluation. Repairs and corruptions are measured relative to Preserve.

F Cross-VLM Transfer: General Principle, Backbone-Specific Intervention

The main experiments use OpenAI CLIP ViT-B/16 as the primary VLM and SigLIP2 as an independent frozen semantic view. We use additional frozen VLMs to separate two questions that should not be conflated: (i) whether the relationship between sparse semantic coverage and intervention utility persists across VLMs, and (ii) whether the complete intervention parameterization learned for the OpenAI-CLIP source configuration transfers unchanged to another primary backbone.

We evaluate three primary/auxiliary assignments: (i) OpenAI CLIP ViT-B/16 with SigLIP2, (ii) DataComp OpenCLIP ViT-L/14 with SigLIP2, and (iii) SigLIP2 with DataComp OpenCLIP ViT-L/14. The third assignment reverses the primary/auxiliary roles and therefore provides a stronger test of whether the observed intervention behavior is specific to CLIP occupying the primary position.

The main experiments evaluate the OpenAI-CLIP source configuration. Here, we report only the two primary alternative configurations under the shared ten-dataset ablation protocol, avoiding comparisons across aggregates from different benchmark sets.

F.1 The coverage–intervention relationship transfers across VLMs

A common preserve–intervene transition appears across VLMs. Despite substantial differences in their absolute prediction quality, all three primary VLMs exhibit the same qualitative dependence on queried semantic coverage. At approximately 90% coverage, intervention is strongly beneficial: $+15.57$ pp for OpenAI CLIP, $+13.46$ pp for DataComp OpenCLIP, and $+20.83$ pp for SigLIP2. At approximately 61–64% coverage, the advantage remains positive but smaller. By approximately 31–34% coverage, repairs and corruptions approach balance, and the net benefit collapses toward zero or becomes negative.

Importantly, this experiment uses the same cache-based intervention construction for all three primary VLMs rather than the final STTI transducer or controller. The resulting crossover therefore does not depend on the particular CLIP-trained intervention parameterization. It supports a more general empirical principle: sparse target labels are most useful when the semantic populations they reveal account for a substantial fraction of the target state, and their intervention value diminishes as that coverage is diluted.

F.2 The complete intervention parameterization is partly backbone dependent

Frozen transfer to DataComp OpenCLIP. The CLIP-trained parameterization transfers meaningfully to DataComp OpenCLIP. Always Intervene is substantially better than Preserve in Very-Low and Low complexity and remains mildly beneficial at $B = 64$ Medium, but loses its advantage as semantic complexity increases. The transferred selector responds conservatively to this crossover: it

Table 26: **Frozen transfer to alternative primary VLMs at $M = 4$.** The OpenAI-CLIP-trained intervention parameterization is transferred without backbone-specific refitting to DataComp OpenCLIP and to the SigLIP2-primary role-reversal configuration. Results are averaged over the ten non-ImageNet benchmarks used by the ablation protocol. Preserve is the frozen cross-view hypothesis, Intervene applies the semantic intervention branch to every episode, and Selective TTI applies the complete preserve–intervene policy. Δ_{sel} denotes Selective TTI minus Preserve.

Regime	Primary StatA	Tip	Preserve	Intervene	**Selective TTI**	Δ_{sel}
DataComp OpenCLIP ViT-L/14 primary + SigLIP2 auxiliary *(frozen transfer)*						
B=64 Very-Low	81.34	93.26	82.01	95.23	**88.12**	+6.11
B=64 Low	80.93	86.62	81.59	89.56	**83.69**	+2.09
B=64 Medium	80.33	80.61	80.91	83.20	**80.98**	+0.06
B=1000 Medium	80.49	79.74	81.58	82.17	**81.58**	+0.00
B=1000 High	81.19	77.36	82.03	80.63	**82.02**	-0.01
B=1000 Very-High	81.13	76.52	81.77	80.23	**81.76**	-0.01
ALL	82.22	77.12	82.85	81.55	**82.85**	+0.00
SigLIP2 primary + DataComp OpenCLIP ViT-L/14 auxiliary *(frozen role reversal)*						
B=64 Very-Low	56.65	86.39	74.83	71.12	**76.27**	+1.44
B=64 Low	56.58	73.68	74.74	65.86	**73.80**	-0.93
B=64 Medium	56.07	62.30	74.61	60.98	**72.71**	-1.90
B=1000 Medium	56.80	61.52	73.64	58.65	**73.24**	-0.39
B=1000 High	57.05	56.92	73.79	57.98	**73.51**	-0.28
B=1000 Very-High	56.91	55.51	73.68	57.69	**73.43**	-0.26
ALL	57.94	55.28	74.06	58.44	**74.06**	+0.00

All results use the same ten non-ImageNet benchmarks as the auxiliary ablation studies. The OpenAI-CLIP source configuration is intentionally omitted here to avoid mixing this ten-dataset aggregate with the 11-benchmark main evaluation that additionally includes ImageNet. Tip uses the same $M = 4$ queried-label budget, whereas Preserve consumes no target labels. All source-trained intervention and decision parameters remain frozen during transfer.

retains part of the sparse-regime intervention gain while becoming nearly identical to Preserve in the dense regimes. Thus, the complete parameterization exhibits useful transfer beyond its source backbone, although the sparse gains are attenuated.

SigLIP2 role reversal exposes a parameterization limitation. The SigLIP2-primary configuration is qualitatively different. Transferring the CLIP-trained intervention operator unchanged makes Always Intervene worse than Preserve across the evaluated regimes, while Selective TTI predominantly falls back toward the preserved hypothesis (Table 26). This rules out a claim of backbone-invariant parameter transfer.

Crucially, Table 25 shows that this failure cannot be explained by an absence of useful sparse-label information: under the common intervention test, SigLIP2 exhibits the strongest Very-Low and Low intervention gains of the three primary VLMs. The remaining dependence therefore lies in the learned mapping from support–target semantic correspondence to intervention, and in its associated source-side calibration, rather than in the underlying coverage–utility phenomenon itself.

What transfers and what does not. Together, the two experiments separate the general inference principle from the current learned parameterization. The relationship between queried semantic coverage and intervention utility appears across OpenAI CLIP, DataComp OpenCLIP, and SigLIP2 under a common intervention construction. The complete source-trained STTI parameterization also transfers usefully from OpenAI CLIP to DataComp OpenCLIP, but does not transfer unchanged under SigLIP2 role reversal. We therefore claim cross-VLM support for the *preserve–intervene principle*, while treating backbone-invariant semantic intervention and calibration as an open problem.

Transfer protocol. All transferred models and decision parameters are frozen before evaluation; no backbone-specific refitting, target-side calibration, or unqueried target labels are used. Cross-VLM

Table 27: **Matched preservation-state cache control and preserve–intervene decomposition at $M = 4$.** Always preserve outputs the frozen hypothesis P^0 without target labels, always intervene applies the candidate hypothesis P^I to every episode, and STTI selects between them. The matched cache applies the same U+D supervision to the same frozen preservation hypothesis P^0 used by STTI, providing a stricter comparison between unconditional cache-based label consumption and selective intervention from the same preservation state. Results are unweighted means over the same ten non-ImageNet component-ablation benchmarks.

	Small-batch / sparse regimes				Higher semantic complexity				
Method	**B64 VL**	**B64 Low**	**B64 Med.**	**Sparse Avg.**	**B1000 Med.**	**B1000 High**	**B1000 VH**	**ALL**	**Macro**
Fixed preserve/intervene endpoints									
Frozen preservation P^0 ($M = 0$)	74.63	73.54	71.97	73.38	74.68	74.58	73.78	74.76	73.99
Always intervene P^I ($M = 4$)	**91.32**	**81.51**	71.18	**81.34**	70.21	67.87	66.93	67.66	73.81
Same-query unconditional cache controls ($M = 4$)									
U+D + cache (CLIP start)	87.75	77.28	68.76	77.93	67.95	64.80	63.86	63.96	70.62
U+D + cache (matched P^0)	86.88	78.36	**72.46**	79.23	73.85	72.09	70.89	71.58	75.16
Selective preservation–intervention ($M = 4$)									
Selective Test-Time Intervention (Ours)	87.21	78.66	72.42	79.43	**74.66**	**74.46**	**73.68**	**74.76**	**76.55**
Paired differences									
Always intervene $P^I - P^0$ (pp)	+16.69	+7.97	-0.78	+7.96	-4.47	-6.70	-6.85	-7.10	-0.18
Matched cache $- P^0$ (pp)	+12.25	+4.82	+0.50	+5.86	-0.83	-2.49	-2.88	-3.18	+1.17
STTI − matched cache (pp)	+0.33	+0.30	-0.04	+0.20	+0.82	+2.37	**+2.79**	**+3.18**	+1.39

All $M = 4$ methods use identical registered target streams and U+D support sets. Always intervene outputs P^I for every episode; P^0 consumes no target labels. Sparse Avg. and Macro are computed from full-precision values before rounding. For STTI minus the matched cache, the 95% paired dataset-bootstrap confidence intervals are $[-0.47, +7.68]$ in High, $[+0.04, +8.00]$ in Very-High, and $[+0.21, +8.89]$ in ALL; all earlier regime intervals include zero. For matched cache minus P^0, the Very-High and ALL intervals are $[-8.29, -0.04]$ and $[-8.88, -0.21]$.

experiments are therefore interpreted as generalization analyses of the source-trained intervention mechanism rather than as independently trained variants.

F.3 Matched Preservation-State Cache Control

The standard same-query U+D control matches STTI's queried samples and annotation budget, but applies cached evidence to the original CLIP prediction state. STTI may instead retain the stronger dual-VLM hypothesis P^0. To control for the stronger starting state, we use a stricter comparator that applies the same U+D supervision directly to P^0.

For normalized frozen visual features $\bar{f}_i$ and $\bar{f}_q$, the matched cache predicts

$$\widehat{y}_i = \arg\max_c \left[\log p_i^0(c) + \alpha_{cache} \sum_{q:y_q=c} \exp\left(-\beta_{cache}\left[1 - \bar{f}_i^\top \bar{f}_q\right]\right) \right].$$

Thus, the matched control and STTI share the same frozen reference state, queried labels, and target streams. The matched control applies cache-based label evidence unconditionally, while STTI constructs its semantic intervention hypothesis and selectively decides whether it should replace P^0.

We evaluate on registered streams from the ten non-ImageNet benchmarks used in our ablations, with identical $M = 4$ U+D support sets. For each held-out dataset, we select cache parameters $(\alpha_{cache}, \beta_{cache})$ through strict leave-one-dataset-out validation on the other nine datasets; no labels of the evaluated dataset are used for parameter selection.

Result. Matching the preservation state explains much of the absolute gap to the single-VLM cache control, confirming that the stronger dual-VLM hypothesis P^0 is an important component of STTI's dense-regime floor. The fixed endpoints expose the intended trade-off: always intervening gains 16.69 pp over P^0 in Very-Low but loses 6.70–7.10 pp in High–ALL, whereas STTI sacrifices part of the sparse-regime maximum to retain the preservation floor as semantic coverage declines. From Very-Low to $B = 1000$ Medium, STTI and the matched cache differ by at most 0.82 pp, with paired confidence intervals including zero. At higher semantic complexity, unconditional use of the same four labels becomes harmful: relative to P^0, the matched cache loses 2.88 pp in Very-High and 3.18 pp in ALL. STTI avoids these losses, outperforming the matched cache by 2.79 pp (95% CI $[+0.04, +8.00]$) and 3.18 pp (95% CI $[+0.21, +8.89]$).

Table 28: **Localization of the UCF101 dense-regime degradation at $M = 4$.** We report the constituent frozen views, the preservation hypothesis, the final STTI prediction, and the number of evaluation folds in which intervention is selected.

Regime	StatA	CLIP	SigLIP2	Frozen Preserve	STTI ($M = 4$)	Intervention
$B = 1000$, High	71.25	67.51	5.66	66.04	66.04	0/4000
ALL	73.59	67.43	5.95	66.22	66.22	0/4

G ADDITIONAL PRESERVATION ANALYSIS

The main paper already reports the complete equal-budget comparison among standard acquisition rules, the same-query U+D control, and STTI across all annotation budgets and semantic-complexity regimes. Those results establish that stronger acquisition alone does not explain the robustness of selective intervention. Here we examine a complementary question exposed by the per-dataset results: what happens when the state selected for preservation is itself imperfect?

G.1 UCF101 PRESERVATION ANALYSIS

UCF101 is the clearest dense-regime case in which STTI remains below the strong label-free StatA reference. We therefore ask whether this degradation arises because the controller authorizes an inappropriate intervention or because the frozen preservation hypothesis itself is weak on this target domain. We separately examine the constituent frozen views, the resulting preservation hypothesis, and the final preserve–intervene decision in $B = 1000$ High and ALL.

The degradation originates in preservation, not intervention. Table 28 localizes the UCF101 degradation to the frozen preservation branch. In both dense settings, the controller selects intervention in zero evaluation folds, so the final STTI accuracy is identical to the preserved hypothesis: 66.04% at $B = 1000$ High and 66.22% in ALL. The observed degradation therefore cannot be attributed to harmful label-conditioned intervention; it is already present in the state that STTI elects to preserve.

Frozen-view reliability is strongly domain-dependent. The constituent views expose the source of this limitation. CLIP obtains 67.51% and 67.43% in the two regimes, whereas the auxiliary SigLIP2 view obtains only 5.66% and 5.95%. The resulting frozen fusion remains relatively close to CLIP, trailing it by 1.47 and 1.21 percentage points, but remains substantially below StatA. Thus, complementary frozen views that improve the preservation state on average need not remain equally informative on every target domain.

Implication for STTI. This failure does not invalidate the preserve–intervene principle, but it identifies a distinct limitation of the current implementation: preservation quality is bounded by the fixed, dataset-agnostic composition of its frozen views. The controller avoids further label-conditioned intervention when the available sparse evidence is insufficient, but it cannot recover from a preservation hypothesis that is itself suboptimal. We deliberately avoid UCF101-specific reweighting or target-dependent correction because such choices would compromise the dataset-agnostic evaluation protocol. Learning a transferable estimate of cross-view reliability while retaining the conservative preserve–intervene decision is therefore an important direction for future work.