WAS IT NEVER SHOWN, OR NEVER SEEN? ATTRIBUTING FAILURE IN OBSERVATION-LIMITED 3D VISION-LANGUAGE TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language systems for volumetric scans, dense 3D scenes, and long videos must construct a budgeted observation before inference. When a prediction fails, aggregate accuracy cannot reveal whether evidence was omitted, the model failed despite adequate evidence, or a useful observation was not selected. We introduce a matched-counterfactual framework for attributing these failures to input-stage (L1), model (L2), and observation-selection (L3) bottlenecks under a declared reference protocol. Cross-domain analysis reveals that observation vulnerability increases sharply under tight budgets and disproportionately affects compact targets. Guided by this diagnosis, we propose a **Joint Adaptive Observation Policy** that jointly allocates a fixed budget across support, sampling density, representation resolution, and modality-specific mapping. Given a query and a charged low-cost preview, the policy selects a feasible observation. Training combines task reward with dense evidence-adequacy rewards while keeping the task model frozen. It improves low-budget CT grounding, temporal video localisation, and 3D visual grounding over default fixed observations by 26.7, 15.2, and 14.8 points, respectively. Reference-recoverable misses decrease substantially, whereas joint system–reference misses change comparatively little, supporting adaptive observation as an intervention on failures originating before task inference.

1 INTRODUCTION

Volumetric scans, dense 3D scenes, and long videos often exceed the context, memory, and latency budgets of vision-language systems, which select slices, frames, points, regions, or representation resolutions before inference (Bai et al., 2025; Chen et al., 2020; Zhang et al., 2025). The task model receives only this compact observation; excluded evidence requires re-observation. We study evidence-localisable tasks whose retained visual support can be assessed, including grounding, detection, segmentation, temporal localisation, and annotated VQA.

This creates an attribution problem. Aggregate accuracy cannot distinguish failures in observation formation, model inference, and observation selection, although each calls for a different intervention. We introduce a matched-counterfactual framework separating input-stage (L1), model (L2), and selection (L3) failures (Figure 1). System and reference observations are evaluated on the same target, budget, frozen model, and decoding procedure; the classification accounts for whether the system uses a fixed observation or can select among feasible actions.

The analysis reveals substantial observation-side bottlenecks. Under a fixed 32-slice CT observation, sampling-vulnerable cases account for 61.5% of small-lesion misses but only 5.7% of large-lesion misses. Finer sampling remains associated with better localisation within target-extent strata, and geometric vulnerability grows sharply at low budgets across CT, video, and point clouds. Improved intensity visibility alone does not consistently improve CT localisation. Together, these findings motivate coordinating observation factors rather than optimising each independently.

We propose a **Joint Adaptive Observation Policy** that treats observation formation as joint resource allocation. Given a query and a charged low-cost preview, it selects support, sampling density, representation resolution, and modality-specific mapping from a finite architecture-compatible registry. Realised token and encoder-compute costs mask infeasible actions. After observation-randomised

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.

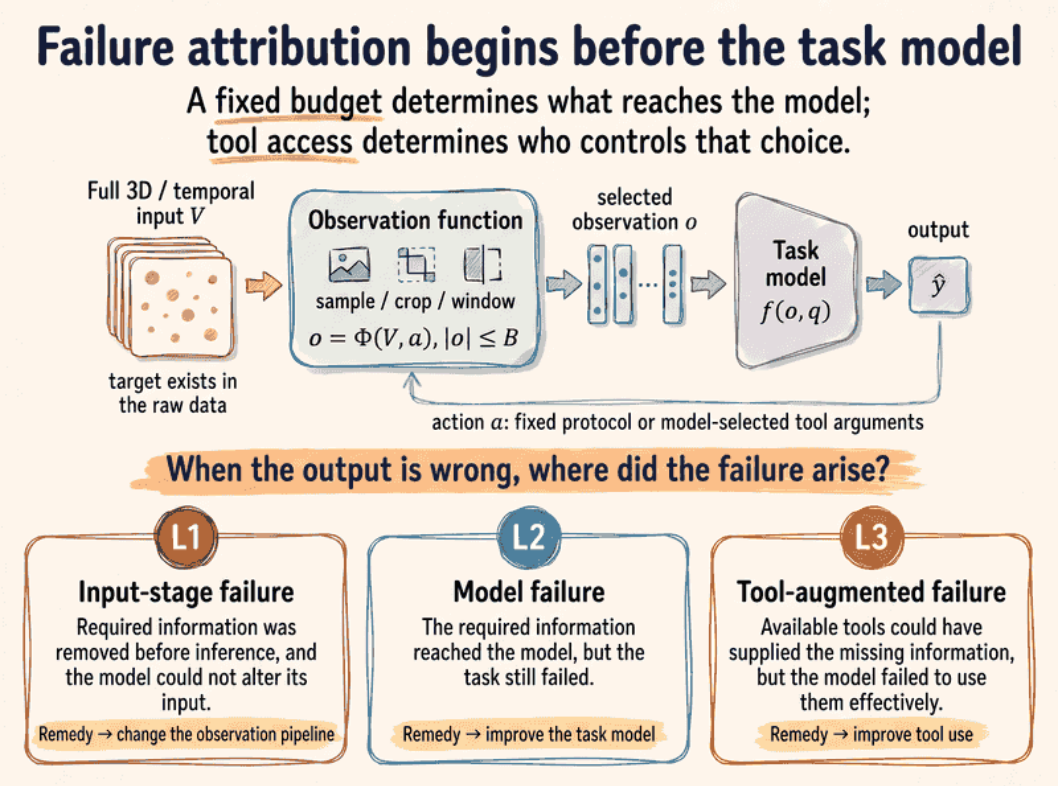

Figure 1: Conceptual overview of the framework. Each domain instantiates selectable observation support, sampling density, representation resolution, and modality-specific mapping. L1 means that adequate evidence was unavailable to the system input, L2 that the same frozen task model failed after adequate evidence was supplied, and L3 that a useful observation was available but the policy selected an unsuccessful one. The schematic's "tool-augmented failure" is this observation-selection class. All paired comparisons use the same target, model, and hard observation budget.

task-model adaptation, we freeze the model and train the policy with task reward augmented with dense evidence-adequacy rewards. The policy serves both as a method for improving budgeted observations and as a controlled intervention for evaluating the framework's diagnostic value.

Across CT, Charades-STA, and ScanRefer, the policy improves over default fixed observations by 26.7, 15.2, and 14.8 points at low budgets, versus 3.2, 1.4, and 2.6 at standard budgets. Paired outcomes show fewer reference-recoverable misses and comparatively small changes in joint system–reference misses. Our contributions are threefold:

- A matched-counterfactual framework separating input-stage, model, and selection failures through geometry screening, realised adequacy, and paired evaluation.
- Cross-domain evidence that observation bottlenecks depend strongly on target geometry and available budget.
- A joint adaptive observation policy that improves performance under hard budgets, with frozen-model controls connecting its gains to recoverable observation failures.

2 Related Work

Failure Diagnosis in Vision-Language Systems. Error taxonomies analyse failures under a supplied input (Hoiem et al., 2012; Bolya et al., 2020; Tong et al., 2024). Observation-limited systems introduce an earlier source of error: the supplied input may omit evidence available in the raw data. This issue arises through slice selection and resolution in medical volumes (Hamamci et al., 2026;

Wu et al., 2025; Bai et al., 2025), sampling and tokenisation in point clouds (Chen et al., 2020; Qi et al., 2017; Xu et al., 2024), and frame selection in video understanding (Sigurdsson et al., 2016; Ren et al., 2024; Zhang et al., 2025; Feng et al., 2026). Our framework evaluates alternative observations under the same budget and frozen model to distinguish recoverable observation failures from failures persisting under an adequate reference.

Adaptive Observation and Active Perception. Vote-MI selects CT slices (Wang et al., 2024); Adaptive Focus and LENS select video frames or spatio-temporal regions (Wang et al., 2021; Zhang et al., 2026); V*, DeepEyes, and ACTIVE-o3 support active visual search (Wu & Xie, 2023; Zheng et al., 2026; Zhu et al., 2026); and ActiveVLA studies 3D active perception (Liu et al., 2026). These methods establish the value of task-dependent observation. We build on this direction by jointly allocating four observation factors and testing the resulting policy as a matched-budget intervention on diagnosed failures. Token compression operates on content already inside the pipeline (Ryoo et al., 2021; Chen et al., 2024; Shao et al., 2025), whereas observation formation also determines which raw evidence enters encoding. Appendix Table 9 compares these axes.

3 Failure-Attribution Framework

3.1 Setting

Let V be a raw 3D or spatio-temporal input and y its task-relevant localisable evidence. Under budget B, action $a \in \mathcal{A}_B$ forms an observation for query q:

$$o = \Phi(V, a; \xi), \qquad \hat{y} = f(q, o), \tag{1}$$

where ξ is the declared sampling or rendering realisation and the frozen task model f receives (q, o) but not V. Actions may control support, density, representation, or modality-specific mapping. Attribution is therefore relative to B, $\mathcal{A}_B$, the tools, randomness, adequacy definition, and frozen task model. The framework applies when Φ is lossy and configurable and evidence adequacy can be evaluated; generic questions with distributed or unannotated evidence fall outside this scope.

3.2 Three classes of failure

L1 is a system miss with no adequate candidate, or a fixed-observation miss paired with a successful adequate reference that the model could not request. L2 means that both the system and the same frozen model under an adequate reference miss. L3 means that a policy-selected observation misses while its adequate reference succeeds. Hits without an adequate candidate and reverse discordance are not failures; all classes are relative to the declared reference set, adequacy criteria, budget, and model. Adequacy describes retained evidence, not task success: an adequate reference can still yield an incorrect prediction. Target annotations support policy training and offline auditing, but are unavailable to the deployed policy.

3.3 Failure Attribution Procedure

We screen vulnerability without task-model outcomes. For evidence extent $e(y)$ and nominal spacing $\Delta(a)$ under regular sampling,

$$G(a, y) = e(y)/\Delta(a). \tag{2}$$

Provided the support contains y, $G \geq 1$ guarantees an intersection for every sampling phase; $G < 1$ only marks vulnerability. Realised coverage instead counts

$$m(a, y; \xi) = \sum_{s \in S(a,\xi)} \mathbf{1}[s \cap y \neq \varnothing], \qquad m(a, y; \xi) \geq k, \tag{3}$$

with primary CT/video threshold $k = 1$ and sensitivities $k = 2, 4$. For irregular point clouds and incomplete video metadata, Δ is an effective proxy and G remains a screen.

For each target, a predeclared finite set $\mathcal{R}_B(y) \subseteq \mathcal{A}_B$ yields reference candidates. Its adequate subset $\mathcal{R}_B^+(y)$ must satisfy $m(a_j, y; \xi_j) \geq k$ and the support, representation, and mapping thresholds

in Section 5; intersection alone is insufficient. The geometry-guided reference uses fixed coverage/context tie-breaks, never task outcomes, and is not an accuracy oracle. If no adequate candidate exists, a miss is L1 and a hit is recorded separately. Otherwise the system and reference use the same target, frozen model, and decoding: fixed miss/reference hit is L1, policy miss/reference hit is L3, both miss is L2, and system-hit cases are successes or reverse discordance. Only reference-eligible targets support paired accuracy differences. Appendix Tables 8 and 17 give the quantity summary and exhaustive outcomes; threshold details are in Appendices A.3 and B.3.

4 Measurement and Findings

4.1 Finding 1: observation vulnerability concentrates in compact targets

CT organ and lesion grounding provide a contrast in target support: the same 32-slice screen flags 2.8% of organ queries but 9.7% of lesion targets, while $G < 1$ accounts for 5.7% of large-lesion misses and 61.5% of small-lesion misses (Appendix Table 10). Vulnerability therefore concentrates in compact evidence. These are candidates, not realised omissions or attributed failures, which require m and paired outcomes. Figure 2 illustrates the distinction; task definitions and sensitivity analyses are in Appendices A.1 and A.3.

4.2 Finding 2: sampling geometry predicts localisation beyond target scale

To separate sampling from the difficulty of small targets, we compare fine and coarse spacing within axial- extent strata and jointly regress hits on extent and observed spacing. Fine sampling is associated with 28.1- and 21.6-point higher hit rates for large and small lesions, and the spacing coefficient remains negative after controlling for extent (Table 1). For large lesions, G predicts localisation better than extent or spacing alone; for small lesions, spacing is slightly stronger. We therefore use G as a geometry screen, not a universal predictor. These associations are not causal, and spacing describes the formed observation rather than physical acquisition thickness.

Table 1: Predictive association after conditioning on lesion size. The hit-rate difference compares fine with coarse sampling within extent strata; b is the coefficient on $\log \Delta$ in a joint model with $\log$ extent. *** denotes $p < 10^{-3}$. Full fitting details are in Appendix A.4.

	Large lesions	Small lesions
Hit-rate difference, fine minus coarse	$+28.1$	$+21.6$
b, coefficient on $\log \Delta$	-1.90^{***}	-1.76^{***}
Pseudo-R^2: $\log G$ / extent / spacing	.097 / .022 / .036	.063 / .025 / .073

4.3 Finding 3: improved visibility does not imply improved task performance

A CT visibility screen flags 40.4% of lesion targets, but its default-adequate, alternative-recoverable, and unrecoverable strata do not consistently order localisation performance. We therefore treat intensity mapping as one action factor rather than an independent adequacy guarantee; complete results and search settings are in Appendix A.4.

4.4 Finding 4: geometric vulnerability increases as observation budgets shrink

Across all three modalities, reducing the sampling budget sharply increases geometric vulnerability: the $G < 1$ candidate rate rises from 9.7% to 66.2% in CT, from 3.0% to 44.9% in video, and from 4.6% to 39.3% in point clouds when moving from standard to low budgets. These regimes motivate adaptive observation, but recoverable headroom still requires paired interventions within baseline-defined G/m strata (Table 7). Figure 3 shows the continuous diagnostics; complete proxy values are in Appendix A.5.

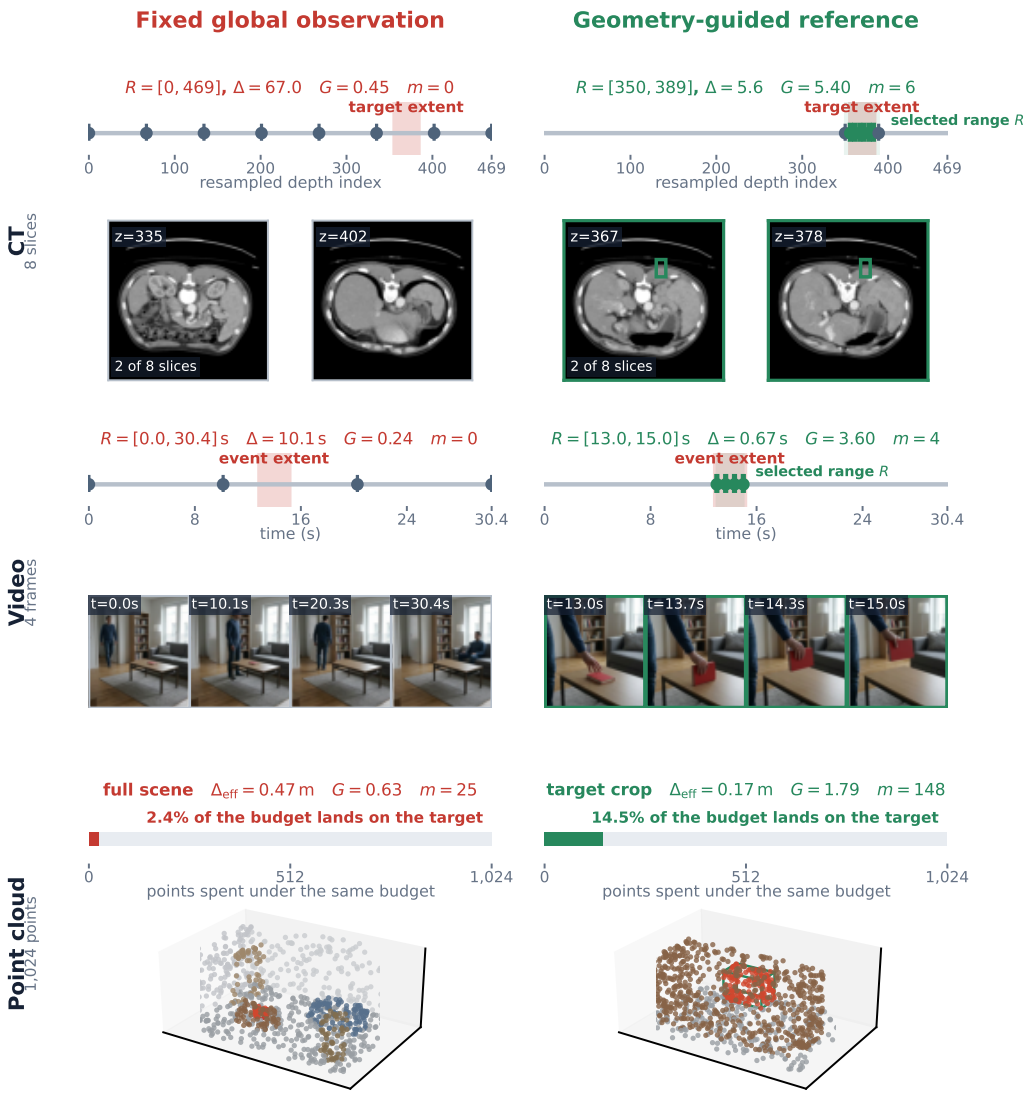

Figure 2: Matched-budget observations in three domains. Each cell reports the selected range R, spacing Δ, screen G, and realised count m above a sampling axis. Red marks the target extent; green marks samples that intersect it. The right column reallocates, rather than increases, the observation budget. The CT row is a real lesion-grounding case; the video and point-cloud rows are illustrative renderings. Their proxy quantities explain the analogous mechanism.

5 Joint Adaptive Observation Policy

The diagnosis motivates joint resource allocation: concentrating support can free resources for denser sampling or finer representation, while mapping affects the visibility of retained evidence. Our Joint Adaptive Observation Policy learns these coupled choices under a fixed budget and also serves as a controlled intervention on the diagnosed bottlenecks. For domain d, the query-conditioned policy selects $a_d = (s_d, \delta_d, \rho_d, \mu_d)$: support, density, representation, and modality-specific mapping. Its finite registry contains only architecture-compatible actions (Table 2). Given charged preview p_d and hard token–cost budget $B = (B_T, B_C)$,

$$\mathcal{A}_{d,B} = \{a \in \mathcal{A}_d : T_d(a) + T_d(p_d) \le B_T,\ C_d(a) + C_d(p_d) \le B_C\}, \tag{4}$$

where T is realised processor-token/group count and C encoder work in the declared domain-specific unit; infeasible actions are masked. Table 3 gives the calibrated caps.

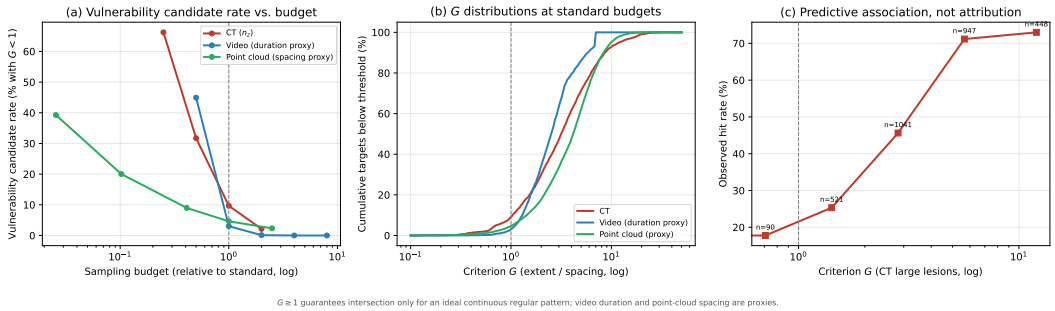

Figure 3: Continuous geometry diagnostics. Left: within-domain trajectories of the all-target $G < 1$ candidate rate. Centre: G distributions at the standard budgets. Right: association between G and localisation for large CT lesions. Video and point-cloud quantities are within-domain proxies.

Table 2: Factorized action interface. Every action is a complete, architecture-compatible tuple; feasible sets are obtained by filtering the same fixed registry using realised token and encoder costs.

Domain	Support s_d	Density δ_d	Resolution ρ_d	Mapping μ_d
CT	normalised xyz crop	slice count	rendered $H \times W$	WL/WW preset
Video	temporal range and xy crop	frame count	frame resolution	identity, contrast, gamma
Point cloud	3D crop	point cap	point groups and neighbourhood	voxel size and sampler

Adequacy is evaluated in the actual rendered coordinate frame along all four axes. Support is the fraction of annotated evidence inside the selected range or crop; sampling counts realised slice/frame intersections or retained instance points; representation measures target-bearing processor patches or point groups; and mapping measures target–shell contrast or retained voxel coverage. Thresholds, action realizations, and normalisation scales are fixed on development data. Missing or unreliable evidence annotations are marked unassessable rather than treated as inadequate.

At deployment, the policy sees only the query and charged preview; it never receives target annotations. After evaluation, the offline reference uses annotated evidence to choose among the same feasible action IDs, with fixed geometry and cost tie-breaks that never use task outcomes. It is therefore a diagnostic counterfactual, not a deployable oracle or a jointly achievable query-level upper bound. Crops outside the system registry cannot establish L3. Appendix A.6 gives the full information-access rules and tie-breaks.

Observation-randomised adaptation first exposes the task model to feasible actions; that model is then frozen across all observation strategies. We subsequently train the policy with task reward augmented with dense evidence-adequacy rewards: $q_{\text{obs}} = q_{\text{sup}}(q_{\text{samp}} + q_{\text{repr}} + q_{\text{map}})/3$ and $r = r_{\text{task}} + \lambda q_{\text{obs}} - r_{\text{format}}$. The support term gates the dense reward, whereas formal adequacy requires every component to meet its fixed threshold. Task reward favours correct predictions; the dense reward encourages retention of annotated evidence across the observation factors. Inference follows *preview* $\rightarrow$ *action* $\rightarrow$ *render* $\rightarrow$ *task model*. Registry, reference, training, and cost details are in Appendices A.6 and A.7.

6 Experiments

6.1 Evaluation protocol and fair budgets

One observation-randomized task checkpoint per domain is frozen across strategies and policy seeds. Metrics are CT target hit rate, Charades-STA Recall@1 at tIoU 0.5/mean tIoU, and ScanRefer Acc@.25/Acc@.5; the system selects one action per query, and splits are separated by patient/case, video, or scene. OneThinker-8B and BUTD-DETR are the video and point-cloud backbones, with Group-Free proposals rebuilt from each retained cloud (Appendices A.1 and A.7).

Table 3: Hard-budget calibration. B_T includes the charged preview; point groups and unique points are architecture-specific accounting units. Caps are rounded upward.

Domain	Anchor	B_T	T_p	Encoder-input anchor	Calibration
CT	8 slices	3,000 tok.	256 tok.	8 slices	$324\times8+256$
CT	32 slices	11,000 tok.	256 tok.	32 slices	$324\times32+256$
Video	4 frames	2,700 tok.	256 tok.	4 frames	$576\times4+256$
Video	8 frames	5,000 tok.	256 tok.	8 frames	$576\times8+256$
ScanRefer	1,024 points	192 groups	64 groups	1,024 unique points	$128+64$ groups
ScanRefer	40,000 points	192 groups	64 groups	40,000 unique points	$128+64$ groups

Table 4: Cross-domain task outcomes at low and standard anchors. BS is development-selected best-static; NP is best-static without a charged preview. Metrics are CT hit rate, video R@1 at tIoU .5, and ScanRefer Acc@.25; policy entries are three-seed means.

Domain	Anchor	Metric	Fixed	BS	Policy	Reference	NP
CT	8 slices	Hit	30.1	39.4	56.8	75.9	43.2
CT	32 slices	Hit	55.2	57.1	58.4	75.9	58.7
Video	4 frames	R@1	31.7	38.2	46.9	61.8	41.0
Video	8 frames	R@1	48.7	49.5	50.1	52.5	50.4
ScanRefer	1,024 pts	Acc@.25	27.8	33.7	42.6	55.4	36.4
ScanRefer	40,000 pts	Acc@.25	50.6	51.8	53.2	57.1	53.5

Default fixed, development-selected best-static, random, policy, and reference strategies share model and decoding. Controlled comparisons charge the same preview; a no-preview best-static deployment baseline may spend the full cap on its final observation. Low/standard anchors are eight/32 CT slices, four/eight video frames, and 1,024/40,000 unique points; realised processor units and encoder-input anchors determine feasibility. Policy entries average three independently trained selector seeds.

6.2 Intervention results across domains and budgets

Table 4 summarises the matched-budget task outcomes; complete sampling, adequacy, random-baseline, and secondary-metric results are in Appendix Tables 14– 16. Sampling and multi-axis adequacy remain distinct, and the reference is a diagnostic rather than an accuracy oracle.

Figure 4 shows larger gains over default fixed observations at low budgets (26.7/15.2/14.8 points in CT/video/ScanRefer) than at standard budgets (3.2/1.4/2.6). Low-budget gains over uniform random selection are 23.2/12.0/11.7 points, showing that registry access alone does not explain the improvement.

6.3 Controls and paired attribution

Support-only, task-reward-only, adaptation, and feasibility-matched tuple-permutation controls separate cropping, adequacy reward, task-model adaptation, and query-dependent selection; restricted policies are retrained (Appendix Table 18). The primary diagnostic-value test is policy-minus-fixed improvement within strata defined by the fixed observation's G and m. Action frequencies alone cannot identify causal factors.

Table 17 reports six exhaustive outcomes for adequacy-assessable targets, separating unassessable annotations from no-candidate cases. Within reference-eligible targets, the paired gap is miss/reference-hit minus reverse discordance on one denominator; full-population, assessable, and eligible rates remain separate.

Threshold, proposal, replay, and blank-observation controls are reported in Appendix B.

Table 5 turns the aggregate gains into failure attribution. At low budgets, the policy reduces eligible miss/reference-hit outcomes from 47.3 to 22.0 points in CT, 31.9 to 18.3 in video, and 29.3 to 15.7 in ScanRefer. Reference-recoverable misses decrease substantially, while joint system–reference misses change comparatively little. Remaining policy A_{MH} outcomes measure selection headroom;

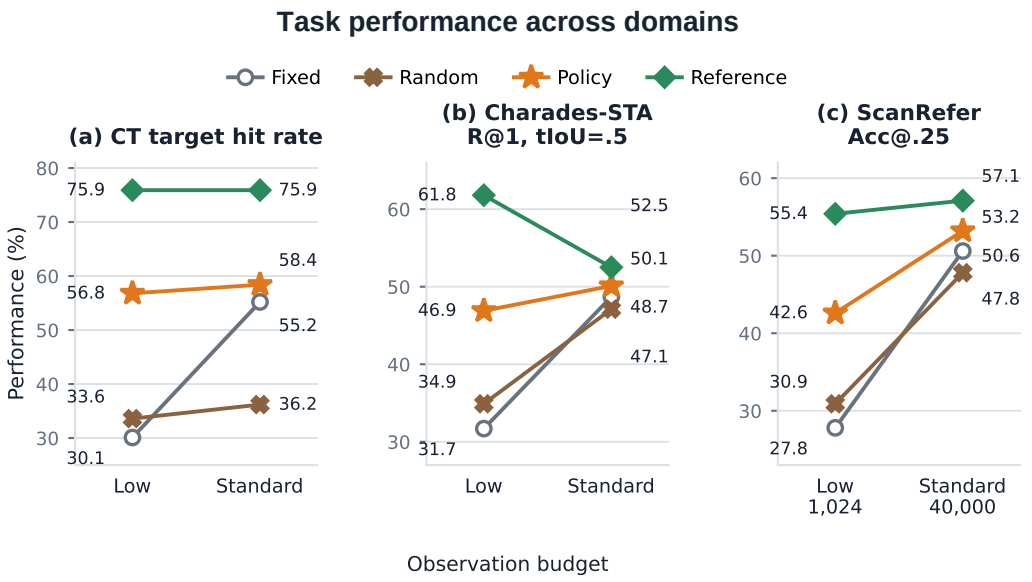

Figure 4: Task performance across domains and observation budgets for fixed, random, policy, and geometry-guided reference observations. Reference–policy gaps are diagnostic and are not L3 incidence; that attribution requires the paired outcomes in Table 5. Tuned static baselines, including the no-preview deployment comparison, are reported in Table 4.

Table 5: Low-budget paired outcomes over targets with assessable evidence. N_H/N_M are no-candidate hits/misses. $A_{MH}, A_{MM}, A_{HM}, A_{HH}$ denote eligible system/reference miss–hit, both-miss, hit–miss, and both-hit outcomes. Thus N_M is L1; A_{MH} is fixed L1 or policy L3; A_{MM} is L2; and A_{HM} is reverse discordance. Every row sums to 100.

Domain	System	N_H	N_M	A_{MH}	A_{MM}	A_{HM}	A_{HH}
CT, 8 slices	fixed	0.6	5.4	47.3	17.2	1.5	28.0
	policy	1.0	5.0	22.0	16.2	2.5	53.3
Video, 4 frames	fixed	0.7	7.3	31.9	29.1	1.8	29.2
	policy	0.9	7.1	18.3	27.7	3.2	42.8
ScanRefer, 1,024 pts	fixed	0.8	9.2	29.3	33.7	1.7	25.3
	policy	1.1	8.9	15.7	32.8	2.6	38.9

A_{MM} cases are not recovered by the current adequate reference, without ruling out recovery by another observation. Standard-budget outcomes are reported in Appendix Table 17.

Table 6: Core low-budget controls using the same adapted task checkpoint. Values are CT hit rate, video R@1 at tIoU .5, and ScanRefer Acc@.25. Tuple permutation preserves feasible complete-action frequencies.

Variant	CT	Video	ScanRefer
Full policy	56.8	46.9	42.6
Support only	49.8	41.3	37.5
Task reward only	50.9	42.0	38.1
Query-only selector	42.4	38.7	34.6
Feasibility-matched tuple permutation	41.7	38.6	35.1
Fixed observation, adapted model	30.1	31.7	27.8

The controls in Table 6 locate the source of the gain. Support selection explains a large share but not all of it: joint selection adds 7.0, 5.6, and 5.1 points over support-only selection in CT, video, and ScanRefer. Removing the dense adequacy reward also degrades every domain. Query-only selection and feasibility-matched tuple permutation remain well below the full policy, so language priors and

the marginal action distribution do not fully explain the result. The adapted checkpoint is shared throughout. Further restrictions and leakage checks appear in Appendix Table 18.

Table 7: Low-budget outcomes in strata defined once by the fixed observation. Gain is policy minus fixed in percentage points. $G < 1$ identifies geometry vulnerability, while $m = 0$ records realised omission.

Domain	Outcome	$G < 1$	$G \geq 1$	$m = 0$	$m \geq 1$
CT	fixed	21.8	46.4	10.8	39.6
	policy	50.4	69.3	44.0	63.1
	gain	28.6	22.9	33.2	23.5
Video	fixed	20.0	41.2	8.0	33.7
	policy	38.6	53.7	16.0	49.5
	gain	18.6	12.5	8.0	15.8
ScanRefer	fixed	16.5	35.1	8.0	29.1
	policy	35.1	47.5	20.0	44.0
	gain	18.6	12.4	12.0	14.9

The fixed observation defines every stratum before policy evaluation, preventing the policy from changing the comparison population. CT gains are largest for $G < 1$ and $m = 0$, consistent with recoverable sampling failures. Video and ScanRefer also gain more in the geometry-vulnerable group, although their realised- coverage patterns differ, showing why G, m, and task outcomes must be reported separately. Mixture weights and threshold sensitivities are specified in Appendix B.3.

Stricter intersection thresholds preserve the policy's sampling advantage; complete sensitivity results are reported in Appendix Table 19.

7 Discussion

Failure composition depends on target population, action, and budget. The framework identifies observation bottlenecks, and the joint policy tests whether they can be reduced without changing model capability. Matched-budget gains within fixed G/m strata support recoverable headroom; support, reward, adaptation, permutation, and no-preview controls assess the contribution of query-dependent joint selection.

G diagnoses density only; L1/L2/L3 require realised multi-axis evidence and paired outcomes from one frozen model. Attribution remains relative to the registry, budget, checkpoint, and evidence definition, and matched compute includes previews and proposal generation. Video pseudo-tubes and within-domain proxies bound adequacy; distributed or unannotated reasoning requires another definition. In particular, L2 means failure under the selected adequate reference, not impossibility under every feasible observation.

8 Conclusion

We introduced a matched-counterfactual framework that separates input-stage, model, and selection failures in observation-limited vision-language systems. Cross-domain analysis revealed strong dependence on target geometry and budget, motivating a Joint Adaptive Observation Policy that coordinates support, density, resolution, and mapping. With the task model frozen, the policy improved low-budget CT, video, and ScanRefer performance by 26.7, 15.2, and 14.8 points over default fixed observations. Paired evaluation linked these gains to fewer reference-recoverable misses, while joint system–reference misses changed comparatively little. The framework therefore guides an effective observation intervention while retaining an explicit reference-dependent attribution boundary. Future work should extend adequacy assessment to distributed evidence and test richer observation registries under the same resource controls.

AI USE STATEMENT

Generative AI tools assisted with portions of code drafting, language editing, and visual polishing of figures and tables. The authors reviewed and verified the resulting code, analyses, numerical results, and manuscript, and take full responsibility for the submission.

REFERENCES

Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3D: Advancing 3d medical image analysis with multi-modal large language models. In *International Conference on Learning Representations*, 2025. arXiv:2404.00578.

Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision*, 2020. doi: 10.1007/978-3-030-58580-8_33.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3d object localization in RGB-D scans using natural language. In *European Conference on Computer Vision*, pp. 202–221, 2020. doi: 10.1007/978-3-030-58565-5_13.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-73004-7_2.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. OneThinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5443, 2026. arXiv:2512.03043.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, 2026. doi: 10.1038/s41551-025-01599-y.

Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, Daguang Xu, and Wenqi Li. VISTA3D: A unified segmentation foundation model for 3D medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20863–20873, 2025.

Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *European Conference on Computer Vision*, 2012. doi: 10.1007/978-3-642-33712-3_25.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. ActiveVLA: Injecting active perception into vision-language-action models for precise 3D robotic manipulation. *arXiv preprint arXiv:2601.08325*, 2026.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 2017. arXiv:1706.02413.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.01357.

Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: What can 8 learned tokens do for images and videos? In *Advances in Neural Information Processing Systems*, 2021. arXiv:2106.11297.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *arXiv preprint arXiv:2507.20198*, 2025.

Gunnar A. Sigurdsson, Guel Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016. doi: 10.1007/978-3-319-46448-0_31.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. arXiv:2401.06209.

Yuli Wang, Jian Peng, Yuwei Dai, Craig Jones, Haris Sair, Jinglai Shen, Nicolas Loizou, Jing Wu, Wen-Chi Hsu, Maliha Imami, Zhicheng Jiao, Paul Zhang, and Harrison Bai. Enhancing vision-language models for medical imaging: Bridging the 3D gap with innovative slice selection. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. doi: 10.52202/079017-3171.

Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. arXiv:2105.03245.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D and 3D medical data. *Nature Communications*, 16, 2025. doi: 10.1038/s41467-025-62385-7.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. *arXiv preprint arXiv:2312.14135*, 2023.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. PointLLM: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72698-9_8.

Ce Zhang, Jinxi He, Yaqi Xie, and Katia Sycara. LENS: Adaptive spatio-temporal zooming for keyframe sampling in long-form videos. In *European Conference on Computer Vision*, 2026. arXiv:2607.25125.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025. arXiv:2410.02713.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, 2026. arXiv:2505.14362.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Mingyu Liu, Zheng Huang, Anzhou Li, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, and Chunhua Shen. ACTIVE-o3: Empowering MLLMs with active perception via pure reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2505.21457.

Table 8: Quantities used in failure attribution. G and m diagnose observation formation; paired outcomes from the same frozen model assign model and observation-selection failures.

Quantity	Role	Limitation
$G = e/\Delta$	screen vulnerable targets, parameters, and budgets	not a realised miss; proxy in point clouds and video
$m(a, y; \xi)$	test adequacy of a realised observation and select references	does not by itself assign a failure class
Paired outcomes	assign L1/L2/L3 under matched conditions	relative to budget, feasible set, tools, and checkpoint

Table 9: Closest adaptive-observation families. These methods motivate domain-specific experimental baselines; our contributions are matched failure attribution and a joint observation policy evaluated as a controlled intervention.

Setting	Representative methods	Selected observation	Difference from our framework
CT volumes	Vote-MI	representative slices	optimises compact volume input; no paired failure attribution
Long video	Adaptive Focus; LENS	frames, regions, temporal ranges	optimises recognition or keyframe efficiency
General VLMs	V*; DeepEyes; ACTIVE-o3	crops and visual tool calls	learns active search without matched adequacy classes
Embodied 3D	ActiveVLA	informative views	couples perception to control rather than diagnosing a frozen task model
Ours	attribution framework and joint policy	support, density, resolution, mapping	separates L1/L2/L3 under one registry and budget

A Implementation Details

A.1 Tasks, datasets, and evaluation units

CT organ grounding. The organ task localises a named anatomical structure with one or more normalised 3D boxes in a full CT volume. We derive boxes from masks in the public segmentation sources used by the VISTA3D data registry (He et al., 2025): AbdomenCT-1K, AMOS22, BTCV, FLARE22, CT-ORG, AeroPath, TotalSegmentator V2, VerSe20, and Medical Segmentation Decathlon Tasks 03, 06, 07, 09, and 10. VISTA3D is a segmentation model and data-integration protocol, not the name of a dataset. Splits are made by 3D case. The training set contains 89,316 positive queries; the held-out set contains 10,978 queries from 371 volumes. The reported geometry screen removes tumor and cyst labels, leaving 10,904 anatomical queries and 11,092 target boxes. A query is counted as vulnerable if any of its target boxes has $G < 1$. The task checkpoint reads 64 uniformly spaced slices with a minimum rendered side of 512 pixels and predicts coordinates in $[0, 1000]^6$.

CT lesion grounding. The lesion task localises one or more findings named by an organ–disease query. Its evaluation set contains 3,135 query records and 3,677 lesion boxes: 75.7% of records express coordinates in an organ crop, 20.6% in a region crop, and 3.6% in a full CT. Organ- and region-conditioned records remain lesion targets; they are not counted as organ grounding. The primary task checkpoint uses 32 slices and in-context examples. The geometry analyses operate per lesion box. We call a lesion small when its longest side is at most 15 voxels. Predictions and targets are paired by Hungarian matching within each size group. A small target is a hit when the dimensionless centre displacement, normalised by the mean target–prediction box diagonal, is below 1.5; a large target requires 3D IoU above 0.1. Because these criteria differ, absolute small- and large-target hit rates are not compared as measures of relative difficulty.

Video and point-cloud grounding. The video evaluation targets the Charades-STA test split: 3,720 queries from 1,334 videos, with Recall@1 at temporal IoU 0.5 and mean temporal IoU. We retained separate versioned identities for OneThinker evaluation intervals and earlier geometry anno-

Table 10: CT vulnerability under a 32-slice observation. Rates for organs and all lesions use all targets; rates by lesion scale condition on model misses. $G < 1$ identifies geometry-based candidates, not attributed L1 failures.

Target population	Conditioning	$G < 1$ candidate rate
Organs	all queries	2.8
Lesions	all targets	9.7
Large lesions	model misses	5.7
Small lesions	model misses	61.5

tations where they differed. Actual decoded frame timestamps, not maximum annotated endpoints, define the inference observation. The point-cloud geometry screen uses a converted file with 36,665 expressions from 562 scenes; its missing official annotation IDs and unverified split exclude it from held-out performance evaluation. The ScanRefer evaluation uses ID-complete official validation annotations and licensed ScanNetV2 assets, with Acc@.25 and Acc@.5 under the official evaluator's threshold convention.

Splits and units for the intervention study. Adaptation, policy training, and development used patient/case, video, and scene partitions disjoint from final evaluation. ScanRefer development scenes were held out from its training split, and no settings were tuned on final validation scenes. The split manifest recorded stable IDs and passed an overlap audit. CT system actions are query-level and scores are target-weighted, with a query-weighted sensitivity analysis; ScanRefer additionally reports unique-object weighting. Original organ/region-conditioned CT crops are task inputs, not evidence that a policy located the region from a full volume; their provenance was recorded. Predictions are mapped back to the original task coordinate frame and scored against unchanged targets, including targets excluded by the chosen crop.

A.2 CT OBSERVATION CONSTRUCTION

CT boxes are converted from normalised or crop-relative coordinates into the represented query stack before computing geometry. At 32 slices, the target-weighted median axial depth of the lesion stacks is 275 native indices, giving median nominal spacing 8.84. Separately, the query-weighted median in-plane short side is 209 voxels; rendering uses a minimum side of 256 pixels, and 16.6% of query volumes are reduced in-plane. These statistics use different declared weighting units and are reported only to characterise the observation pipeline. They do not establish that every failure occurs on the axial axis.

The organ geometry comparison applies the same endpoint-inclusive 32-slice idealisation to each full-volume anatomical box. For a query containing multiple instances, the query is marked vulnerable if any instance has $G < 1$. This yields the 2.8% organ-query rate in Table 10; the lesion all-target rate is 9.7%. The organ task-model result is not used to compute either geometry rate.

A.3 GEOMETRY SCREEN, REALISED ADEQUACY, AND CANDIDATE SEARCH

For a represented stack of depth Z and endpoint-inclusive budget n_z, the nominal ideal-pattern spacing is

$$\Delta = (Z - 1)/(n_z - 1). \tag{5}$$

If $Z \leq n_z$, every index is retained and we set $\Delta = 1$. The realised indices are obtained by rounding an endpoint-inclusive linear grid. For a continuous target interval $[z_1, z_2]$, m is the number of realised indices inside the interval. Consequently, $G < 1$ and $m = 0$ answer different questions: the former identifies phase vulnerability, whereas the latter establishes realised omission for $k = 1$.

The geometry-only candidate search is performed after realised inadequacy. For each density in $\{8, 16, 32\}$, target-centred windows use widths from the target-tight interval, n_z, $2n_z$, $4n_z$, $8n_z$, and the complete represented stack, clipped to the valid range. A candidate is adequate when its realised count reaches k. The primary analysis uses $k = 1$; Table 11 reports stricter requirements. All rows condition on 1,617 model-unmatched lesion targets from the measured 32-slice run. This annotation-centred search is not the policy-accessible registry of Section 6 and does not establish

matched-budget L3 incidence. Separately, over all 3,677 targets, the fixed eight/32-slice grids omit 1,214/110 targets, giving the 67.0%/97.0% sampling-coverage rates in Table 14.

Table 11: Sensitivity of the CT lesion screen among model-unmatched targets. Panel A recomputes G at counterfactual density budgets. Panel B holds the realised 32-slice observation fixed and searches the declared reference candidates at each adequacy threshold. Values are percentages, not failure attribution.

A. Sampling-budget screen

Budget n_z	$G \geq 1$	$G < 1$
32	87.9	12.1
16	58.2	41.8
8	20.5	79.5

B. Realised-adequacy threshold

k	$m \geq k$	$m < k$, candidate	no candidate
1	96.6	2.8	0.6
2	73.5	25.2	1.4
4	30.7	65.3	4.0

A.4 Size-controlled and intensity-window analyses

For the size-controlled analysis, lesion targets are separated by the 15-voxel longest-side threshold. Within each group, axial extent is divided into quartiles; each quartile is then split at its median Δ. The reported fine-minus-coarse difference is the quartile-size-weighted mean. We additionally fit

$$\Pr(\text{hit}) = \sigma(c + a \log \max(e, 0.5) + b \log \Delta) \tag{6}$$

with $L_2 = 10^{-4}$. Standard errors use the inverse observed Hessian. This target-level analysis establishes a predictive association; it does not identify a causal acquisition effect or account for all within-volume correlation.

Table 12: Complete size-controlled lesion-grounding analysis. Significance marks use the target-level Hessian approximation: *** denotes $p < 10^{-3}$.

	Large ($n = 3{,}057$)	Small ($n = 620$)
Fine-minus-coarse hit-rate difference	$+28.1$	$+21.6$
a, coefficient on $\log e$	$+1.07^{***}$	$+0.84^{***}$
b, coefficient on $\log \Delta$	-1.90^{***}	-1.76^{***}
z-statistic on b	-19.1	-8.0
$-b/a$	1.78	2.08
Pseudo-R^2, $\log G$	.097	.063
Pseudo-R^2, extent alone	.022	.025
Pseudo-R^2, spacing alone	.036	.073

The intensity analysis uses 3,666 valid lesion targets; eleven targets are excluded because either the volume, target interior, or surrounding shell cannot be evaluated. Target and shell Hounsfield units are summarised in 10-HU bins over $[-1100, 1100]$. The target shell expands each box by half its side length in each dimension and removes the target interior. The default window has level 40 and width 400. A window is visibility-adequate when no more than 20% of target voxels are saturated and the target–shell median difference spans at least two 8-bit grey levels. The dense search evaluates levels from -900 to 900 in 25-HU steps and widths in $\{50, 80, 120, 160, 200, 250, 300, 400, 500, 650, 800, 1000, 1300, 1600, 2000, 2500\}$, for 1,168 level–width pairs. It optimises this visibility proxy, not task-model performance under every window.

A.5 Cross-domain geometry proxies

For video, event extent is the annotated temporal interval and the available-duration proxy is the maximum annotated endpoint D for that video. Endpoint-inclusive spacing is $D/(N-1)$. Because this can underestimate the true video duration, the resulting G is a screening proxy. For point clouds, event extent is replaced by the minimum target-box side, and the primary effective-spacing proxy

is $\Delta_{\text{eff}} = (V_{\text{scene}}/N)^{1/3}$. An alternative surface proxy $(A_{\text{scene}}/N)^{1/2}$ produces lower absolute candidate rates, confirming that these values are not calibrated physical guarantees.

Table 13: Complete proxy budget sweeps over all annotated targets. Values are $G < 1$ candidate rates in percent.

Video frames	Duration proxy	Points	Volume proxy	Surface proxy
4	44.9	1,024	39.3	9.2
8	3.0	4,096	20.0	1.9
16	0.1	16,384	9.0	0.4
32	0.0	40,000	4.6	0.1

A.6 Action registries and adequacy

We fixed a finite base registry before policy training, independently of evaluation annotations. It recorded crop-coordinate grids, density, rendered resolution, mapping, sampling seeds, invalid combinations, and a stable ID for every complete action. Feasible sets filtered this registry by realised processor and encoder caps, and the default fixed action remained feasible. Random selection was uniform over feasible complete actions. Best-static was selected on development data from configurations feasible for every eligible input; its deployment variant used the same registry without a preview charge. Neither used evaluation labels.

Coverage and adequacy. Support is the fraction of annotated evidence contained in the selected support. Sampling uses exact realised slice/frame intersections and, for point clouds, unique retained instance points and their fraction of source instance points. Representation measures target-bearing processor patches or point groups; mapping uses target–shell contrast or retained voxel coverage. We computed these quantities in the actual rendered coordinate frame. The sampling screen used $k = 1$ with $k = 2, 4$ sensitivities in CT/video; all-axis adequacy additionally required every axis-specific threshold. Normalisation scales and thresholds were frozen on development data. The measured CT window proxy in Appendix A.4 is not a validated universal adequacy rule.

Video temporal annotations do not supply spatial evidence tubes. We generated quality-checked detector–tracker pseudo-tubes as offline training/audit annotations, marked missing or unreliable tubes as unassessable rather than adequate, and reported temporal coverage independently.

Reference selection. For full-action attribution, the evaluator enumerates the system's feasible action IDs using the declared sampling realizations. Among adequate actions it ranks, in order, minimum normalised adequacy margin, realised target coverage, target-bearing representation, widest retained context, lower encoder cost, and stable action ID. Axis scales and within-domain context measures were locked with the thresholds; task outcomes never enter this ranking. Target-specific references are offline diagnostics, not additional system observations: in multi-target CT they do not imply one action simultaneously achieves every target's reference outcome. Their aggregates are not a jointly achievable query-level upper bound. The actual system still uses one query-level action. Target-tight crops not present in the registry are labelled privileged controls and cannot contribute to L3.

Across budgets the base registry, thresholds, and action realizations remained fixed. Increasing both caps with the same preview retained every previously feasible action. This guarantees nondecreasing adequate candidate availability, not nondecreasing accuracy of the geometry-ranked reference. The replay control retains the complete lower-budget observation and decoding settings; any new reference selection is logged.

A.7 Training, inference, and budget accounting

Backbones. The measured CT analyses use Qwen3.5-9B lesion checkpoints with and without demonstrations. The video backbone is OneThinker-8B; the point-cloud grounder is BUTD-DETR with Group-Free proposals recomputed from each observation's retained cloud. We used point-group/neighbourhood variation only for architecture-compatible configurations. If a point encoder

requires 2,048 input rows, a 1,024-point observation may repeat only its retained points, while logging unique and presented counts separately. Repetition adds neither target coverage nor new evidence. Full-source geometry defines GT boxes; no full-cloud proposal, feature, or annotation may enter the budget-limited prediction path.

Adaptation and policy controls. Observation-randomized adaptation samples feasible training observations while preserving the original task target under coordinate conversion, including targets outside the chosen crop. The resulting task model is frozen across fixed, random, policy, and reference comparisons. Policy training uses only training labels; three independent policy seeds share the same task checkpoint. All variants received the same training-data and optimisation budget. The adaptation control also runs the unadapted checkpoint on identical fixed observations; it is not mixed into the policy-only gain.

Preview and costs. The previews are a low-resolution CT mosaic, a sparse full-duration video strip, and a coarse point-cloud projection. The preview is available to the selector but not passed to the final task model; fixed and random strategies ignore its content in the controlled comparison. It is charged once per query, without cross-query amortisation. Thus final-observation coverage does not claim that the selector itself never saw a target in the preview. The deployment static baseline omits this stage entirely.

Image/video accounting uses actual processor grids; point-cloud accounting records architecture-specific group/features and raw-point encoder work, not an assumed one-point/one-token conversion. Every online visual encoder, including preview encoding and proposal generation, contributes to the encoder budget. Offline annotation-guided reference enumeration is excluded from deployable latency; its rendered observation and task-model inference still obey the online cap.

Statistical analysis and audit. Stable query and target IDs joined original annotations, actual frame timestamps or source-point indices, action IDs, checkpoint hashes, predictions, parse status, costs, and outcomes. Evaluator thresholds and sampling realizations were fixed across paired runs. Completed but malformed predictions counted as misses; missing assets, decoder errors, and inference failures did not. The final population was fixed before inference and all exclusions were disclosed.

Baseline G bins were < 0.5, $[0.5, 1)$, $[1, 2)$, $[2, 4)$, and ≥ 4; video m bins were 0, 1, 2, 3, and ≥ 4, and point bins were 0, 1–4, 5–15, 16–63, and ≥ 64. Sampling coverage alone is not evidence of a policy benefit; the diagnostic test is paired improvement in fixed-baseline strata.

B Supplementary Results

The tables below report supplementary comparisons; geometry analyses are in Appendix A.

B.1 COMPLETE TASK COMPARISONS

Table 14: CT comparison at low and standard anchors. Sampling is the percentage with $m \geq 1$; the two fixed sampling rates reuse measured geometry, not new inference. Adequacy requires all four axes.

Anchor	Strategy	Sampling (%)	Adequacy (%)	Hit rate
8 slices	fixed	67.0	58.6	30.1
	best-static	78.5	68.4	39.4
	random	73.8	59.7	33.6
	policy	91.6	86.2	56.8
	reference	96.0	94.0	75.9
	best-static, no preview	83.4	73.1	43.2
32 slices	fixed	97.0	90.5	55.2
	best-static	97.4	92.1	57.1
	random	78.2	62.8	36.2
	policy	97.8	94.1	58.4
	reference	99.0	98.0	75.9
	best-static, no preview	98.0	93.6	58.7

Table 15: Charades-STA comparison. Sampling is temporal $m \geq 1$ coverage, whereas adequacy requires all axes. Spatial adequacy additionally depends on a validated pseudo-tube audit.

Anchor	Strategy	Sampling	Adequacy	R@1, 0.5	Mean tIoU
4 frames	fixed	92.2	65.4	31.7	25.8
	best-static	94.0	73.6	38.2	31.4
	random	88.1	62.7	34.9	28.3
	policy	97.0	85.6	46.9	39.1
	reference	98.4	92.0	61.8	51.0
	best-static, no preview	96.1	78.9	41.0	34.0
8 frames	fixed	99.0	89.4	48.7	40.4
	best-static	99.1	91.0	49.5	41.2
	random	95.6	80.3	47.1	39.0
	policy	99.2	93.5	50.1	41.7
	reference	99.6	97.0	52.5	43.6
	best-static, no preview	99.4	92.7	50.4	42.0

Table 16: BUTD-DETR comparison. Sampling is the percentage of targets with at least one retained instance point, not the fraction of each object's points retained. Adequacy requires all declared thresholds. Values are percentages; a 4,096-point sweep is secondary.

Anchor	Strategy	Sampling	Adequacy	Acc@.25	Acc@.5
1,024 points	fixed	94.0	61.8	27.8	16.4
	best-static	96.0	70.5	33.7	20.8
	random	92.1	59.3	30.9	18.6
	policy	97.4	82.6	42.6	28.1
	reference	98.6	90.0	55.4	38.7
	best-static, no preview	96.8	75.9	36.4	23.0
40,000 points	fixed	99.8	92.4	50.6	35.8
	best-static	99.8	93.4	51.8	37.0
	random	98.1	84.6	47.8	33.4
	policy	99.8	95.3	53.2	38.4
	reference	99.9	98.0	57.1	41.7
	best-static, no preview	99.9	94.6	53.5	38.6

Table 17: Standard-budget six-way percentages for fixed and policy systems; every row sums to 100. N_H/N_M are no-candidate hits/misses. $A_{MH}, A_{MM}, A_{HM}, A_{HH}$ are candidate-eligible system/reference miss–hit, both-miss, hit–miss, and both-hit outcomes. N_M is L1; A_{MH} is fixed L1 or policy L3; A_{MM} is L2; A_{HM} is reverse discordance. Percentages are computed over targets with assessable evidence.

Domain	Anchor	System	N_H	N_M	A_{MH}	A_{MM}	A_{HM}	A_{HH}
CT	32	fixed	0.2	1.8	22.3	20.7	1.6	53.4
	32	policy	0.3	1.7	19.6	20.3	2.0	56.1
Video	8	fixed	0.3	2.7	6.8	41.8	3.0	45.4
	8	policy	0.4	2.6	6.0	41.3	3.5	46.2
ScanRefer	40,000	fixed	0.2	1.8	8.7	38.9	2.2	48.2
	40,000	policy	0.3	1.7	6.8	38.3	2.8	50.1

B.2 Additional action ablations

The additional controls in Table 18 complement the primary controls in Table 6. Single-factor restrictions are retrained with that factor fixed to its default and the others free within the same caps; they are not test-time overrides. CT axial support and xy crop, and video temporal support and xy crop, are distinct controls. Unsupported configurations are omitted.

The permutation control shuffles complete realised action tuples among queries with the same feasibility signature at a fixed budget and sampler seed. It therefore preserves tuple frequencies within these groups without creating invalid combinations. It reports the fraction of queries whose action actually changes and does not generalise a null result from singleton groups. Its changed-action shares are 82.0%, 79.0%, and 84.0% for CT, video, and ScanRefer. Query-only selection and a blank-observation reference control test language and crop-coordinate priors. The blank control retains the query, crop transform, and valid interface shape but removes visual values; its all-target score includes the default fallback where no adequate reference exists.

Table 18: Additional low-budget controls. n/a denotes a domain-inapplicable factor. The blank-reference row is a leakage diagnostic rather than a deployable strategy.

Variant	CT hit	Video R@1	ScanRefer Acc@.25
Fixed observation, unadapted model	27.6	29.4	25.1
Fixed axial/temporal support	37.1	39.8	n/a
Fixed xy crop	55.9	46.0	n/a
Fixed 3D crop	n/a	n/a	31.6
Fixed density	53.7	44.8	40.1
Fixed representation	55.3	45.9	41.4
Fixed modality mapping	56.5	46.7	40.8
Feasibility-matched tuple permutation	41.7	38.6	35.1
Blank observation, reference transform	13.2	10.6	8.4

B.3 Stratification, adequacy sensitivity, and reference controls

The low-budget stratification in Table 7 uses the fixed observation to define both groups, then compares fixed and policy outcomes on those same targets. The $G < 1$ mixture weights are 66.2%, 44.9%, and 39.3% for CT, video, and ScanRefer. The $m = 0$ shares are 33.0% for the reused CT geometry and 7.8%/6.0% for video/ScanRefer. Weighted fixed and policy rates recover the main-table results within rounding. Stratum differences alone do not establish an interaction or a size-controlled association; these require per-target statistical analysis.

Table 19: Low-budget sensitivity to required intersection count k. Coverage columns report $m \geq k$; reference availability additionally requires all other adequacy axes. Observations and predictions are fixed while the offline audit threshold varies.

Domain	k	Fixed coverage	Policy coverage	Reference available
CT	1	67.0	91.6	94.0
CT	2	35.8	79.3	86.0
CT	4	11.6	58.7	68.0
Video	1	92.2	97.0	92.0
Video	2	53.0	81.4	84.0
Video	4	12.0	38.5	55.0

At $k = 4$, policy coverage remains 58.7% versus 11.6% for fixed CT observation and 38.5% versus 12.0% for video. Reference availability declines, so absolute failure-class rates remain definition-dependent.

The ScanRefer proposal diagnostic in Table 20 uses target recall among the same budget-specific top-128 proposals before language-conditioned box refinement. It is not an upper bound on final grounding accuracy when refinement can move a proposal. The intermediate anchor is an additional sweep point.

Table 20: ScanRefer proposal audit and intermediate sweep. Proposals are rebuilt from each strategy's retained cloud, including the reference's default fallback.

Unique points	Strategy	Proposal R@.25	Proposal R@.5	Acc@.25	Acc@.5
1,024	fixed	43.5	29.6	27.8	16.4
	policy	63.8	46.5	42.6	28.1
	reference	75.2	59.0	55.4	38.7
4,096	fixed	59.8	43.1	39.5	26.3
	policy	70.2	53.8	47.7	32.9
	reference	78.4	62.7	56.3	40.1
40,000	fixed	73.6	56.2	50.6	35.8
	policy	77.4	60.1	53.2	38.4
	reference	81.0	66.0	57.1	41.7

A replay control carries the exact low-budget reference observations, including fallbacks, into the standard cap with the same model and deterministic decoding. Its scores remain 75.9, 61.8, and 55.4 for CT, video, and ScanRefer, versus 75.9, 52.5, and 57.1 when reselecting the standard-budget reference. This illustrates why action reselection can change accuracy without violating budget nesting. The blank-reference results in Table 18 likewise do not establish that crop-coordinate leakage is absent.