SUB-SEMANTIC IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic segmentation tells us what a region is, but many useful visual distinctions depend on how it looks. Bare and moss-covered rock, peeling paint and exposed wood, or grass growing through sand may share semantic labels while forming clearly distinct appearance regions. Traditionally, segmentation partitions scenes into predefined objects and an amorphous "stuff" category. However, as modern segmentation models increasingly integrate with natural language, this rigid dichotomy becomes obsolete, and there is no longer a reason to discard rich visual information into a generic "stuff" bucket. Yet, driven by a strong semantic bias, these advanced models degrade when encountering regions that are too abstract for a fixed dictionary noun, but visually distinct enough to elicit rich human descriptions. To provide a more fitting representation and explicitly target this vulnerability, we introduce *sub-semantic image segmentation*, the task of autonomously discovering such regions, describing each in free-form language and assigning every description an associated mask, all without receiving region prompts or the number of regions in advance. To make this problem measurable, we contribute three complementary task-specific benchmarks spanning annotation-derived natural scenes with variable region counts, independently constructed natural surface transitions and entirely new controlled synthetic compositions, together with validation on an established real-world texture dataset. We propose Detecture, an integrated description-to-partition system that generates structured appearance descriptions, extracts isolated grounding states, predicts description-conditioned masks and resolves them through pixelwise competition. Across all four evaluation routes, Detecture consistently outperforms native GLaMM and current Sa2VA in region covering, and leads the evaluated native language-and-mask model class on both natural benchmarks. Controlled interventions also show that descriptions actively guide region assignment: changing appearance wording alters masks, and mask identities follow appearance edits substantially more often than spatial edits among jointly eligible cases. Together, these results establish language as a practical interface for discovering and addressing visual structure beyond fixed category vocabularies.

1 INTRODUCTION

A rock surface can be bare in one area and moss-covered in another. A weathered board can alternate between peeling paint and exposed wood; a patch of ground can contain both sand and emerging grass. Semantic labels identify familiar classes, but these distinctions concern how a surface looks. The appearance regions of interest need not follow category or object boundaries. We ask a segmentation system to discover those regions and express what distinguishes them, as illustrated by the paired descriptions and masks in Figure 1.

Segmentation tasks organize images around different targets. Semantic segmentation assigns classes; instance segmentation separates particular objects; panoptic segmentation combines the two views (Kirillov et al., 2019b). Texture segmentation instead emphasizes local visual statistics and repeated patterns (Julesz, 1981; Malik & Perona, 1990). A material is a physical substance, such as wood, whose appearance also depends on finish, illumination and wear. *Visible appearance* includes color, texture, surface state and larger structure. We formulate *sub-semantic image segmentation* as discovering regions organized by this appearance criterion and describing the distinctions between

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.

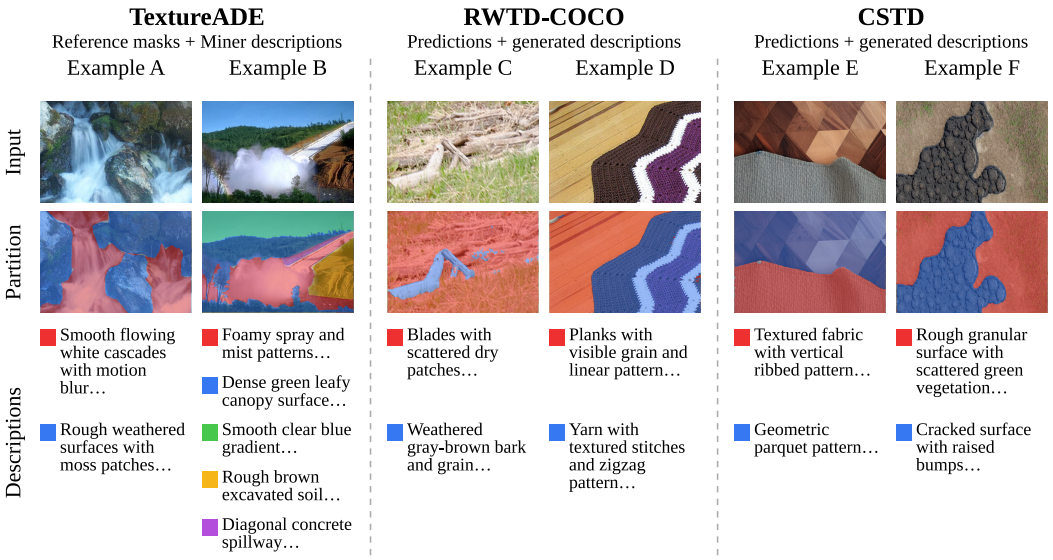


Figure 1: **Discovering and describing appearance regions.** Each column shows an input image, its partition and excerpts of the region descriptions; colors link each phrase to its region. A and B: TextureADE reference masks with Miner descriptions. C and D (RWTD-COCO), E and F (CSTD): Detecture predictions with generated descriptions; C's wood mask is incomplete, and E and F were selected for visual clarity. These illustrative examples are separate from the quantitative evaluation.

them. "Sub-semantic" names the grouping criterion, rather than a fixed hierarchical level below every semantic class.

Descriptions make discovered regions addressable without a fixed vocabulary of region classes. "Weathered granular gray surface" expresses a visible distinction, while its mask identifies the pixels to which it applies. A description and mask together answer what makes a region distinctive and where it is. Referring segmentation locates a region from supplied language (Hu et al., 2016; Wu et al., 2020; Huang et al., 2020; Lüddecke & Ecker, 2022; Wang et al., 2022); our task asks the system to discover the region and generate the description. Entity segmentation (Qi et al., 2021) supplies class-agnostic region discovery, and native grounded-generation systems already produce text and masks (Rasheed et al., 2024; Yuan et al., 2026; Zhou et al., 2026). We bring these capabilities together around appearance-oriented discovery and a coherent partition. Appendix A discusses related work in more detail.

Task contract. Given one image and one fixed global instruction, a system generates a variable list of appearance descriptions and an associated mask slot for each accepted description. No image-specific phrase, category list, point, box or reference count is supplied. A *region* is a pixel set; a *mask* represents that set; a *slot* indexes a candidate output, which may receive no pixels. The described masks and an undescribed residual together form a disjoint label map covering the image. *Grounding* is the association between a description and its image region.

Six design goals. The task asks a system to:

1. **Appearance cohesion:** group pixels sharing visible color, texture or structure relative to neighboring regions.
2. **Category independence:** discover appearance groups across or within categories and objects.
3. **Open descriptiveness:** express a region through free-form appearance wording.
4. **Complete, non-overlapping partition:** aim to cover the image with disjoint described regions, leaving an empty residual.
5. **Granularity dependence:** accommodate useful decompositions at different scales.
6. **Permutation equivalence:** preserve the partition when description-mask pairs are consistently reindexed, keeping residual label zero fixed.

Making this task measurable requires benchmarks that span different forms of appearance structure. We contribute *TextureADE*, mined from ADE20K for natural scenes with variable region counts; *RWTD-COCO*, independently constructed from COCO-Stuff for natural two-surface transitions; and *CSTD*, created through a separate controlled synthetic texture-composition pipeline. Together they provide reproducible reference masks for complex natural scenes, simpler natural transitions and synthetic compositions. The existing Real-world Texture Dataset (RWTD) provides additional external validation.

We propose Detecture, which turns discovered appearance descriptions into a partition. It generates structured descriptions, recomputes isolated grounding states, predicts description-conditioned masks independently and resolves them by pixelwise competition. Across the three contributed benchmarks and RWTD, Detecture exceeds native GLaMM and current Sa2VA in region covering, and leads the evaluated native grounded-generation systems on TextureADE and RWTD-COCO. Controlled interventions connect these results to the language interface: editing descriptions changes masks, and mask identities follow appearance edits more often than spatial edits in eligible clause-swap cases.

Contributions.

- **Task formulation:** sub-semantic image segmentation as autonomous appearance-region discovery with generated descriptions and associated masks.
- **Three complementary benchmarks:** TextureADE, RWTD-COCO and CSTD, built through distinct natural-scene, natural-transition and synthetic construction paths.
- **Detecture:** a novel model that jointly describes and grounds multiple appearance regions in an image; it introduces a new block-diagonal grounding attention mask and a new Shifted-Zero language loss (Appendix C).

2 DETECTURE

Detecture uses descriptions both to express discovered appearances and to compute their masks. It connects language-based discovery to pixel assignment in four steps: generate structured descriptions, extract isolated grounding states, compute each mask independently and let the masks compete. A vision-language model (VLM), Qwen3-VL-8B (Bai et al., 2025), supplies image-conditioned language states; Segment Anything Model 3 (SAM 3; Carion et al., 2025) supplies the segmentation interface and visual features. Grounding markers connect language to masks, building on earlier grounded-generation systems (Lai et al., 2024; Rasheed et al., 2024). Figure 2 shows the architecture.

Structured autonomous generation. One image and the fixed instruction produce descriptions under ordinary causal generation:

`TEXTURE_k: Texture of <appearance>, <location> [SEG].`

The fields express a candidate region in words. The final `[SEG]` marker identifies the hidden state used for its mask. Detecture accepts at most six assistant-region markers as mask slots. The actual count is generated; a candidate slot can later become empty under competition.

Isolated grounding for multi-reasoning. After generation, Detecture recomputes the grounding-marker states. A naive coupling that extracts multiple grounding states from a single causal stream suffers from severe cross-region contamination. Without isolation, grounding states collapse onto one another (e.g., yielding a cosine similarity of 0.74, where a state intended for gravel also retains features of foliage; Appendix C). To enable true multi-reasoning, each description attends to the shared image/instruction prefix and its own earlier tokens, excluding earlier descriptions via a novel block-diagonal attention mask. Original token positions are preserved. This architectural isolation reduces state similarity to 0.16, allowing the system to independently ground multiple distinct appearance regions without interference. Appendix M specifies the exact instruction and inference procedure.

Bridge and independent mask lanes. A *query* is the vector conditioning one mask computation. A learned Bridge maps each recomputed marker state $\mathbf{h}_k \in \mathbb{R}^{4096}$ to the native segmentation width:

$$f_B(\mathbf{h}_k) = \mathrm{Dropout}_{0.4}\big(\mathrm{GELU}\big(\mathrm{LN}(W_B\mathbf{h}_k + b_B)\big)\big), \qquad q_k = R(f_B(\mathbf{h}_k)).$$

LN denotes layer normalization and GELU the Gaussian error linear unit; R is SAM 3's frozen text resizer. The Bridge maps 4096 to 1024 dimensions and R maps 1024 to 256; dropout is disabled at

inference. SAM 3's image backbone provides shared visual features. Each query then independently conditions the transformer encoder and pixel-feature head using those features. A shared learned query-to-mask head produces its logit map $z_k(x)$. A learned residual query follows the same path. The active path uses independent encoding lanes rather than SAM 3's multi-slot mask decoder.

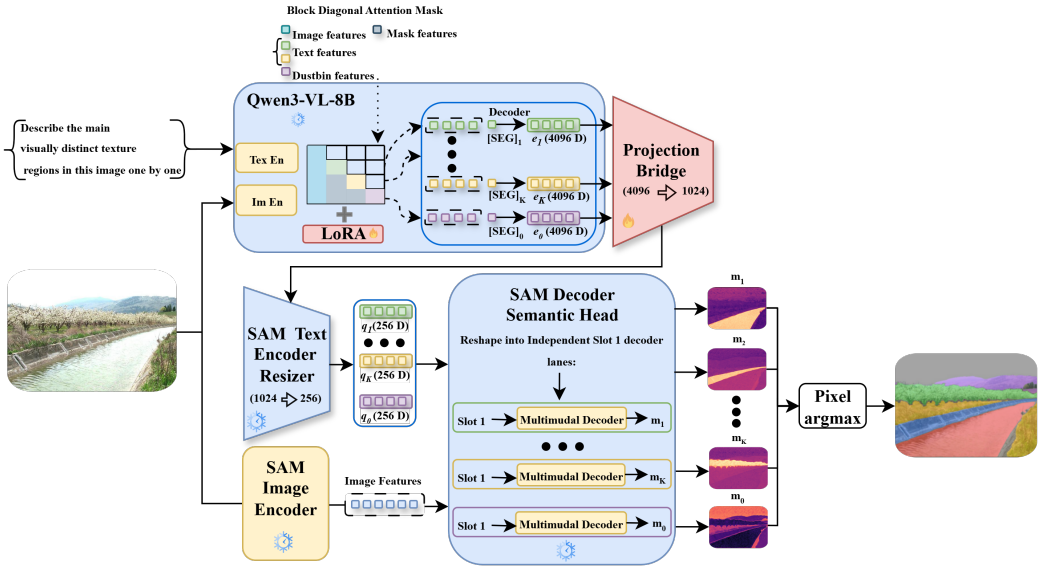


Figure 2: **Detecture's description-to-partition architecture.** Causal generation precedes isolated grounding. Each marker state passes through the Bridge and frozen SAM 3 resizer into an independent mask lane; the external learned residual vector e_0 is prepended after state extraction and follows the same Bridge and resizer. Flame icons identify historically adapted components; evaluation freezes every weight. In the autonomous evaluation, the text prompt is fixed: every image receives the same instruction (Appendix M.1), which contains no image-specific text or per-image region count, so only the image varies.

Competitive pixel assignment. For pixel x, the output is

$$Y(x) = \underset{k \in \{0,\ldots,\hat{K}\}}{\arg\max}\ z_k(x),$$

where $\hat{K}$ is the number of accepted descriptions and zero denotes the residual. Competition produces a disjoint label map. The residual represents pixels outside the described masks, and output diagnostics measure its extent. Matching to reference regions occurs only during evaluation.

Shifted-Zero language loss. Training a VLM head to emit functional segmentation tokens presents a unique optimization challenge. Naively zeroing the direct language loss on the grounding marker—so the state acts purely as a visual query—destroys emission entirely, causing the model to stop generating markers (resulting in a 0.000 evaluation score; Appendix C). To solve this, we introduce the Shifted-Zero language objective. The language loss supervises the emission of grounding markers while setting the direct next-token loss immediately after a marker to zero. Shifted-Zero thus retains marker-emission supervision while removing the marker state's direct loss for predicting the following token. This vital adaptation rescues the autonomous generation capability. The published adaptation recipe first trains the Bridge, mask head and residual, then adds low-rank adaptation (LoRA) of the VLM and updates grounding-marker token parameters, with SAM 3 frozen. Mask cross-entropy, Dice and language supervision train the integrated system. Appendix B specifies the recipe; the experiments below characterize the released checkpoint with all weights frozen.

3 Experiments

We first describe the evaluation data (Section 3.1) and then test whether autonomous systems recover the reference regions (Section 3.2).

3.1 THREE COMPLEMENTARY BENCHMARKS

The three benchmarks make autonomous region discovery measurable under complementary conditions: complex natural scenes with variable region counts, natural two-surface transitions from an independent source and controlled synthetic compositions. Table 1 summarizes their evaluation populations. Each benchmark releases images and reference masks with construction records. TextureADE also supplies descriptions from Detecture Miner, our dataset-curation pipeline (distinct from the Detecture model); RWTD-COCO and CSTD include no reference text.

TextureADE: variable-cardinality natural scenes. TextureADE supplies photographs with multiple candidate appearance regions. Detecture Miner scores annotations from ADE20K (Zhou et al., 2019), selects images, extracts class masks and attaches structured descriptions (Figure 3). The release contains 212 images and 828 described masks, averaging 3.91 regions per image. The image identifiers, masks and metadata preserve the ADE20K source mapping. This natural-scene benchmark complements texture/material resources centered on attributes or category labels (Cimpoi et al., 2014; Bell et al., 2013; 2015).

Miner uses a geometry-first construction: source annotations → candidate scoring → mask extraction → descriptions. Scores combine region structure and boundary geometry; image-language signals can adjust ranking. Selected masks are described from the image, overlay and class hints in a structured format. Masks therefore precede their descriptions. Appendix E gives the formulas and filters, plus the TextureSAM-augmented training-data adapters (Cohen et al., 2025; Tu et al., 2024).

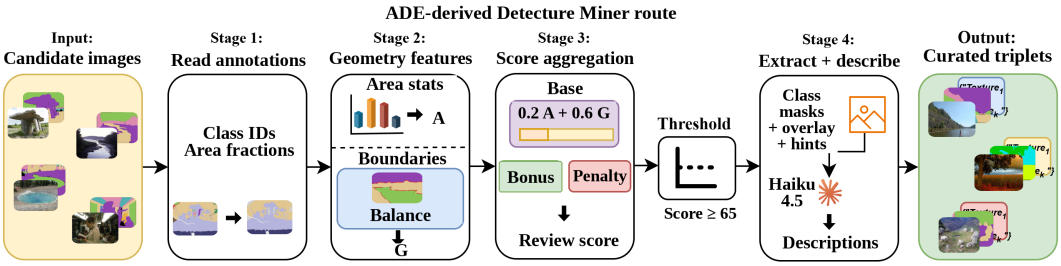


Figure 3: **Detecture Miner constructs TextureADE.** Annotation geometry selects candidate images and masks; image-language signals refine ranking, and structured descriptions make selected regions addressable. RWTD-COCO and CSTD follow their own construction pipelines.

RWTD-COCO: natural two-surface transitions. A separate pipeline finds adjacent surface labels in COCO-Stuff (Caesar et al., 2018), crops around their shared boundary and ranks candidates by coverage, balance and connectedness (Figure 4). Every target pixel is a deterministic reuse of a source annotation. The release contains 2,500 unique-source crops; 212 curator-selected anchors and 44 ranked additions form the 256-image evaluation subset. Assets include crops, binary and source-label masks, source identifiers, coordinates, scores and balancing records. This independently sourced construction isolates simpler natural transitions (Appendix F).

CSTD: controlled synthetic texture compositions. CSTD is our synthetic benchmark, built by a separate texture-composition pipeline (Figure 5). As documented in its dataset card, the construction combines source textures from the Describable Textures Dataset (DTD; Cimpoi et al., 2014) with Stable Diffusion 1.5 and ControlNet conditioning. Targets are the two-region stitching masks supplied to generation. Texture-feature screening and visual selection define the released 256-image evaluation subset. Generated images, binary masks, subset metadata, selected identifiers, screening scores and the screening script document this controlled setting (Appendix G).

External validation. The existing Real-world Texture Dataset (RWTD; Khan et al., 2015; Cohen et al., 2025) contributes 253 real-world texture images to the evaluation. It complements the three new benchmarks as an external source. Section 3.2 evaluates every system on all four datasets.

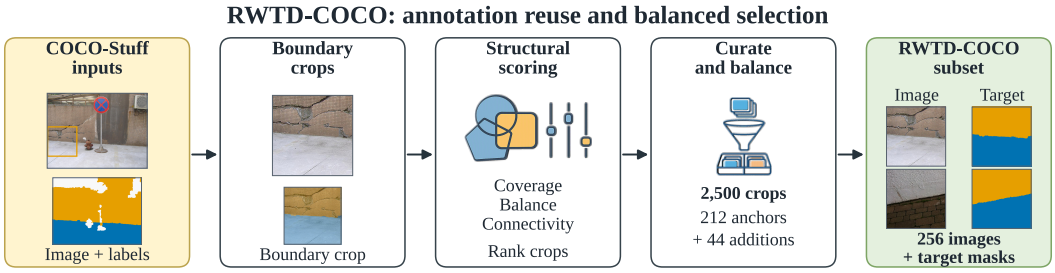


Figure 4: **RWTD-COCO construction.** Adjacent COCO-Stuff labels define boundary crops; structural scoring, curator anchors and balancing select the evaluation subset. Blue and amber trace the inherited pavement and wall masks, respectively. The final column shows two deterministically selected evaluation cases. Target pixels come from source annotations.

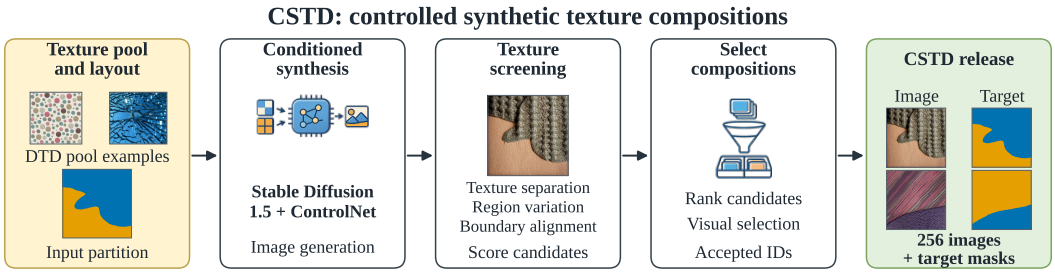


Figure 5: **CSTD construction.** Conditioned synthesis is followed by texture screening and visual selection. DTD thumbnails illustrate the source pool; the input layout and first output belong to released case `005500`. Targets retain the generation-input masks; the output column shows two deterministically selected release cases.

3.2 Autonomous Region Recovery

Can autonomous systems recover the reference regions? We compare systems that receive an image and a general instruction and generate descriptions with masks: GLaMM (Rasheed et al., 2024), Sa2VA-4B, Sa2VA with Qwen3-VL/SAM 3 (Yuan et al., 2026; Sa2VA Team, 2026), SAMTok-3B (Zhou et al., 2026) and Detecture. The baselines use their released text-and-mask interfaces. Eligibility follows the autonomous output interface. Each uses its supported prompt and preprocessing; Detecture requests appearance regions, while native grounded generators request general image descriptions. The comparison concerns complete model-and-instruction systems.

Protocol and metric. All images in Table 1, plus external RWTD, are scored. No reference-derived count, phrase, point or box selects predicted regions. Declared conflict rules convert native masks into disjoint labels (Table 21); failed or empty outputs remain scored. Our primary *bidirectional region covering* averages two area-weighted scores: each reference region's best intersection-over-union (IoU) with a predicted region, and each predicted region's best IoU with a reference region. Both directions use the full annotation-valid area, so missing or residual pixels reduce the score. Appendix K gives the exact definition. Paired significance tests use Holm correction across the

Contributed benchmark	Images	Regions per image	Reference origin
TextureADE	212	variable	ADE20K class masks
RWTD-COCO	256	two	cropped COCO-Stuff masks
CSTD	256	two	synthetic input masks

Table 1: **Three complementary evaluation settings.** TextureADE tests multi-region natural scenes; RWTD-COCO tests simpler natural transitions from another source; CSTD tests controlled synthetic compositions.

System	Contributed benchmarks			External validation
	TextureADE	RWTD-COCO	CSTD	RWTD
GLaMM	61.1 [57.5, 64.6]	55.2 [52.0, 58.2]	73.9 [70.7, 77.0]	65.7 [62.3, 68.9]
Sa2VA-4B	66.1 [62.9, 69.3]	71.9 [68.9, 74.8]	83.6 [80.8, 86.3]	70.8 [67.6, 74.0]
Sa2VA Qwen3-VL/SAM 3	71.8 [68.9, 74.7]	73.6 [70.3, 76.8]	79.7 [76.4, 82.9]	70.4 [66.8, 73.9]
SAMTok-3B	67.4 [64.5, 70.0]	66.5 [63.4, 69.5]	**89.4** [87.1, 91.7]	68.6 [65.3, 71.9]
Detecture	**75.7** [73.8, 77.7]	**85.2** [83.5, 86.8]	87.7 [85.7, 89.6]	**79.4** [77.3, 81.5]

Table 2: **Autonomous appearance-region recovery.** Bidirectional covering in percent (higher is better), with marginal 95% image-bootstrap intervals; bold marks the highest mean per dataset. Three contributed benchmarks are grouped separately from external RWTD. Systems autonomously generate descriptions and masks using their supported instructions. Paired effects and the original 40-contrast Holm family are in Appendix M.

original 40 method-by-dataset comparisons; displayed intervals are marginal 95% image-bootstrap intervals.

Strong task-matched recovery. Detecture exceeds GLaMM and Sa2VA Qwen3-VL/SAM 3 in covering on all four datasets after correction. On the two natural contributed benchmarks, TextureADE and RWTD-COCO, it also exceeds Sa2VA-4B and SAMTok-3B, leading the evaluated native grounded-generation systems. Its CSTD covering of 87.7 trails SAMTok-3B and remains competitive with Sa2VA-4B, whose paired differences are statistically unresolved. The integrated description-to-partition system thus recovers benchmark regions autonomously, generating the descriptions and masks without image-specific prompts or a supplied region count. Complete visual, proposal, panoptic, oracle-assisted and secondary-metric comparisons are in Appendix M. Appendix C traces how Detecture grew out of a two-model cascade experiment and what each design fix contributed.

Variable output in practice. Detecture generates 3.63 nonempty regions per TextureADE image on average, against 3.91 reference regions; exact count agreement is 46.7%. The mean residual occupies 2.0% of annotation-valid pixels. Figure 6 shows a multi-region output selected nearest to the dataset's median covering, with its grounding descriptions and original slot identities. Appendix I reports the full count distribution and Table 33 records residual and empty-output behavior. The generated wording is also varied rather than templated: in archived runs, 98.8% of the 1,857 descriptions are distinct strings (Appendix J).

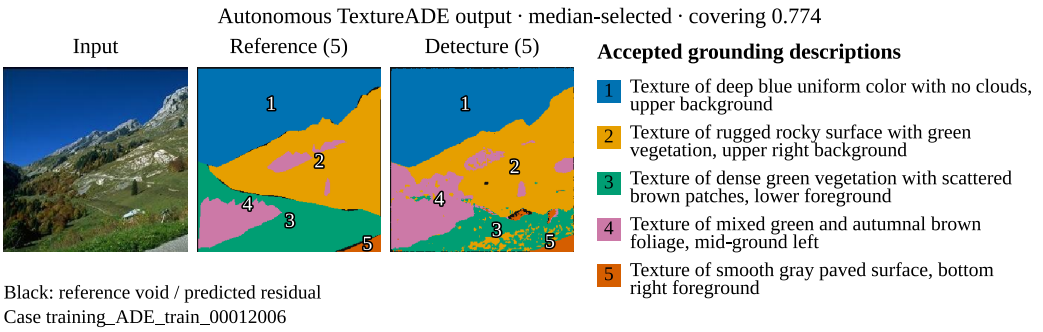


Figure 6: **Autonomous multi-region output on TextureADE.** The example is closest to median covering, with identifier tie-breaks. All five accepted grounding descriptions keep their original mask indices. Sky is recovered; lower vegetation/path regions are fragmented and some boundaries shift. Black denotes reference void or predicted residual.

4 Do Descriptions Functionally Affect Region Assignment?

Descriptions offer a practical interface to the partition only if their content affects region assignment. We test this connection by holding the image and checkpoint fixed, replacing or modifying the descriptions, and recomputing masks. The description-replacement experiment uses 32 images per dataset, selected by hashed identifiers. Every effect compares an edited description with that image's unedited grounding recomputation.

Content changes recovery. Replacing descriptions with nonsense changes covering by −29.7 percentage points on TextureADE (95% interval −35.4 to −23.9) and −21.0 on RWTD-COCO (−29.0 to −12.5). Unrelated descriptions from another image also reduce recovery. These effects survive their original within-dataset correction (Appendix D).

Appearance wording guides identity. A separate clause study exchanges appearance or spatial wording among generated descriptions. Ordinary mask matching can hide a slot exchange, so we also measure identity relative to the pairing obtained before editing. For example, when slot 1 receives slot 2's appearance wording, *following* means that slot 1 selects the region previously paired with slot 2. An image follows only when every slot adopts the association specified by the exchange.

The TextureADE study uses a fixed instruction permitting one to six regions; 98 of 212 images have equal reference, slot and nonempty-region counts in the unedited, appearance-swapped and spatial-swapped conditions. On this joint subset, identities follow appearance edits in 58/98 cases and spatial edits in 1/98. The paired difference is 58.2 percentage points (95% interval 48.0–67.3), significant in the original three-contrast correction. Figure 7 illustrates the distinction between geometric recovery and slot identity.

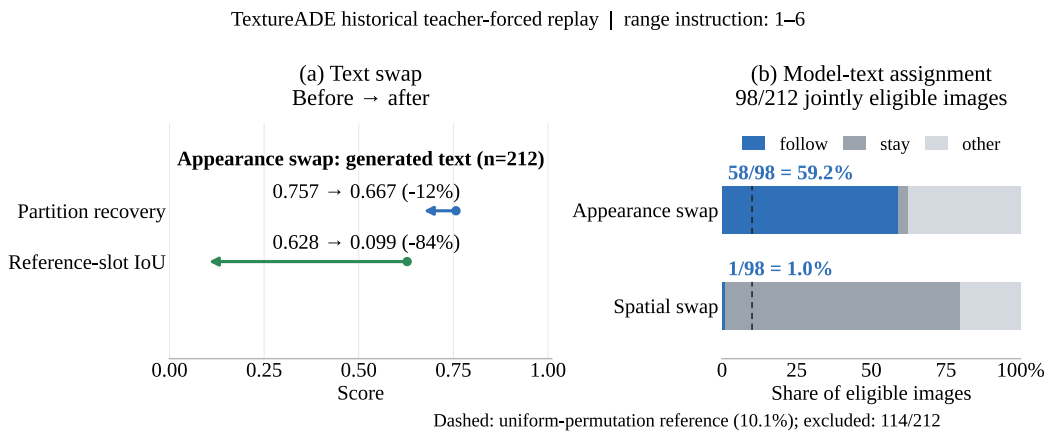


Figure 7: **Appearance edits change mask-slot identity.** The original TextureADE clause experiment replays fixed generated text under a one-to-six-region instruction. (a) Partition recovery uses bidirectional covering; reference-slot IoU keeps the pre-edit pairing fixed. (b) Appearance versus spatial following among 98 jointly eligible images. *Stay* preserves the original associations; *other* denotes another complete assignment. The dashed uniform-permutation reference is separate from the paired appearance-versus-spatial test; full protocols and uncertainty appear in Appendix L.

Together, replacement and clause-edit experiments show that descriptions actively guide region assignment: their content changes masks, and eligible mask-slot identities follow appearance wording more often than spatial wording. This functional dependence connects the language used to describe visual structure to the masks used to address it.

5 Limitations

Reference meaning. TextureADE and RWTD-COCO inherit semantic boundaries; transitions inside a source class can be absent. CSTD uses synthetic input stitching masks, and generated appearance can drift from those masks. Independent appearance-boundary annotations would extend evaluation

beyond these targets (Appendices E–G). Phrase-to-mask correctness also needs independent judgments: the swapped soil and spillway descriptions in the released record behind Figure 1B coexist with accurate masks (Appendix H).

Exposure and experimental scope. Incomplete training, exposure and selection records prevent certifying held-out generalization; RWTD was used during development. Evidence concerns one Detecture checkpoint. Matched retrained controls would be needed to identify individual components' training-time benefits. Historical clause experiments feed fixed text back through the model (teacher-forced recomputation) and require equal counts across conditions; their identity rates describe those subsets. Exact configurations, intervention populations and source-dependence checks are in Appendices B, L and M.

Partition quality and cost. Specialized visual and mask-only systems lead on some routes, and native grounded competitors have higher mean boundary scores across the four datasets (Tables 22, 25, 27, 29 and 31). Region count is imperfect, and every evaluated Detecture output leaves some annotation-valid pixels in the residual (Table 33). Better boundaries, region counts and described coverage are direct improvement targets. Generation and repeated grounding add computation beyond mask-only prediction.

Integration versus cascading. A base-Qwen/SAM 3 cascade, which parses generated descriptions and grounds each through SAM 3's full text encoder, is a legitimate alternative that exceeds Detecture's covering on two-region CSTD and RWTD. Detecture instead joins the two models through the Bridge, a single learned projection followed by SAM 3's frozen resizer; on two-region scenes the cascade is a practical ceiling for this lightweight interface, while Detecture leads it on multi-region TextureADE (Appendix C).

6 Conclusion

Sub-semantic image segmentation asks systems to discover regions defined by how they look, describe what distinguishes them, and associate those descriptions with masks. TextureADE, RWTD-COCO and CSTD make region recovery measurable across complementary natural and synthetic settings, with further validation on external RWTD. Detecture performs this task through a single integrated description-to-partition system, with strong region-covering results among the evaluated native language-and-mask systems. Controlled interventions show that description content functionally guides mask assignment. These contributions make language a practical interface for discovering and addressing visual structure beyond fixed category vocabularies.

AI USE STATEMENT

Generative AI is part of the studied method and data pipeline: Qwen3-VL-8B generates descriptions and grounding states, and the retained Miner annotation configuration uses Claude Haiku 4.5 for image-mask descriptions. AI assistance was also used substantively for research planning, code inspection and implementation, experimental design, frozen-weight evaluation, statistical analysis, literature retrieval, figures, writing, simulated peer review and multilingual readability diagnostics; these diagnostics are not human peer review. GPT Image 1.5 generated decorative process icons for Figures 4 and 5; their dataset images and masks come from existing source files. The authors are responsible for the scientific claims and final submission.

REPRODUCIBILITY STATEMENT

The anonymous supplement supplies compact results, metric checks, scripts, exact model/data revision manifests and claim-to-evidence bindings. Historical data-construction and supervised learnability records are separated from the completed frozen-checkpoint evaluations. No training, weight update, new segmentation-model inference or human study was performed in this editorial revision. The released Detecture configuration can be inspected, but its complete training and selection history is unavailable. Appendix B specifies what is known; Appendix M gives the evaluation contracts. The compact supplement runs metric and statistical checks from derived inputs. Array-level checks require saved predictions; full inference requires the pinned external datasets, pretrained models and per-model environments. Its README separates these entry points. The source package compiles without accessing those external artifacts.

REFERENCES

Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour Detection and Hierarchical Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. doi: 10.1109/tpami.2010.161.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631, 2025.

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A Richly Annotated Catalog of Surface Appearance. *ACM Transactions on Graphics*, 32(4):1–17, 2013. doi: 10.1145/2461912.2462002.

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material Recognition in the Wild with the Materials in Context Database. In *CVPR*, pp. 3479–3487, 2015. doi: 10.1109/cvpr.2015.7298970.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and Stuff Classes in Context. In *CVPR*, pp. 1209–1218, 2018. doi: 10.1109/cvpr.2018.00132.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Dídac Surís, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. arXiv preprint arXiv:2511.16719, 2025.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *CVPR*, pp. 1280–1289, 2022. doi: 10.1109/cvpr52688.2022.00135.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *CVPR*, pp. 3606–3613, 2014. doi: 10.1109/cvpr.2014.461.

Inbal Cohen, Boaz Meivar, Peihan Tu, Shai Avidan, and Gal Oren. TextureSAM: Towards a Texture Aware Foundation Model for Segmentation. arXiv preprint arXiv:2505.16540, 2025.

Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Panoptic Narrative Grounding. In *ICCV*, pp. 1344–1353, 2021. doi: 10.1109/iccv48922.2021.00140.

Eric Hayman, Barbara Caputo, Mario Fritz, and Jan-Olof Eklundh. On the Significance of Real-World Conditions for Material Classification. In *ECCV*, pp. 253–266, 2004. doi: 10.1007/978-3-540-24673-2_21.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from Natural Language Expressions. In *ECCV*, pp. 108–124, 2016. doi: 10.1007/978-3-319-46448-0_7.

Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring Image Segmentation via Cross-Modal Progressive Comprehension. In *CVPR*, pp. 10485–10494, 2020. doi: 10.1109/cvpr42600.2020.01050.

Béla Julesz. Textons, the Elements of Texture Perception, and Their Interactions. *Nature*, 290:91–97, 1981. doi: 10.1038/290091a0.

Naeemullah Khan, Marei Algarni, Anthony Yezzi, and Ganesh Sundaramoorthi. Shape-Tailored Local Descriptors and Their Application to Segmentation and Tracking. In *CVPR*, pp. 3890–3899, 2015. doi: 10.1109/cvpr.2015.7299014.

Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. In *CVPR*, pp. 6392–6401, 2019a. doi: 10.1109/cvpr.2019.00656.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *CVPR*, pp. 9404–9413, 2019b. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kirillov_Panoptic_Segmentation_CVPR_2019_paper.html.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. In *ICCV*, pp. 3992–4003, 2023. doi: 10.1109/iccv51070.2023.00371.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *CVPR*, pp. 9579–9589, 2024. doi: 10.1109/cvpr52733.2024.00915.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint arXiv:2303.05499, 2023.

Timo Lüddecke and Alexander Ecker. Image Segmentation Using Text and Image Prompts. In *CVPR*, pp. 7076–7086, 2022. doi: 10.1109/cvpr52688.2022.00695.

Jitendra Malik and Pietro Perona. Preattentive Texture Discrimination with Early Vision Mechanisms. *Journal of the Optical Society of America A*, 7(5):923–932, 1990. doi: 10.1364/josaa.7.000923.

Jitendra Malik, Serge Belongie, Jianbo Shi, and Thomas Leung. Textons, Contours and Regions: Cue Integration in Image Segmentation. In *ICCV*, pp. 918–925, 1999. doi: 10.1109/iccv.1999.790346.

David R. Martin, Charless C. Fowlkes, and Jitendra Malik. Learning to Detect Natural Image Boundaries Using Local Brightness, Color, and Texture Cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004. doi: 10.1109/tpami.2004.1273918.

Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. doi: 10.1109/tpami.2002.1017623.

Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Zhe Lin, Philip Torr, and Jiaya Jia. Open-World Entity Segmentation. arXiv preprint arXiv:2107.14228, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel Grounding Large Multimodal Model. In *CVPR*, pp. 13009–13018, 2024. doi: 10.1109/cvpr52733.2024.01236.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. arXiv preprint arXiv:2408.00714, 2024.

Sa2VA Team. Sa2VA Qwen3-VL-4B SAM3 Model Release, 2026. URL https://huggingface.co/ByteDance/Sa2VA-Qwen3-VL-4B-SAM3. Released June 2026; revision 3f2bec126b22761a4607eec674781c138f9f9cca.

Gabriel Schwartz and Ko Nishino. Visual Material Traits: Recognizing Per-Pixel Material Context. In *ICCV Workshops*, pp. 883–890, 2013. doi: 10.1109/iccvw.2013.121.

Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H. Adelson. Recognizing Materials Using Perceptually Inspired Features. *International Journal of Computer Vision*, 103(3):348–371, 2013. doi: 10.1007/s11263-013-0609-0.

Peihan Tu, Li-Yi Wei, and Matthias Zwicker. Compositional Neural Textures. In *SIGGRAPH Asia Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3680528.3687561.

Paul Upchurch and Ransen Niu. A Dense Material Segmentation Dataset for Indoor and Outdoor Scene Parsing. In *ECCV*, pp. 450–466, 2022. doi: 10.1007/978-3-031-20074-8_26.

Manik Varma and Andrew Zisserman. A Statistical Approach to Texture Classification from Single Images. *International Journal of Computer Vision*, 62(1–2):61–81, 2005. doi: 10.1007/s11263-005-4635-4.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-Driven Referring Image Segmentation. In *CVPR*, pp. 11676–11685, 2022. doi: 10.1109/cvpr52688.2022.01139.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-Based Image Segmentation in the Wild. In *CVPR*, pp. 10213–10222, 2020. doi: 10.1109/cvpr42600.2020.01023.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *CVPR*, pp. 3858–3869, 2024. doi: 10.1109/cvpr52733.2024.00370.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model. arXiv preprint arXiv:2312.17240, 2023.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with MLLM for Dense Grounded Understanding of Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/tpami.2026.3720379.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A Simple Framework for Open-Vocabulary Segmentation and Detection. In *ICCV*, pp. 1020–1031, 2023. doi: 10.1109/iccv51070.2023.00100.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. doi: 10.1007/s11263-018-1140-0.

Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, Haochen Wang, Haobo Yuan, Jiacong Wang, Lu Qi, Hao Fei, Anran Wang, Zhuochen Wang, Yujing Wang, Cheng Chen, Shunping Ji, and Xiangtai Li. SAMTok: Representing Any Mask with Two Words. In *CVPR*, pp. 37852–37863, 2026.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized Decoding for Pixel, Image, and Language. In *CVPR*, pp. 15116–15127, 2023. doi: 10.1109/cvpr52729.2023.01451.

A Related Work Context

This appendix gives the fuller literature positioning deferred from the main paper. Three neighboring literatures shape the present question. One asks how textures and materials should be represented. A second asks how language can serve as an interface for dense prediction, and a third asks how modern promptable segmentation systems convert text or proposals into masks. Our setting lies at the intersection: we want language to help a model discover texture regions for itself, not merely follow a supplied target.

Texture and material segmentation. Classical work on texture perception emphasizes textons and preattentive texture differences (Julesz, 1981; Malik & Perona, 1990), while descriptor-based and statistical models such as local binary patterns and filter-bank statistics formalized texture as a discriminative appearance signal (Ojala et al., 2002; Varma & Zisserman, 2005). Early cue-integration formulations and later boundary-based segmentation work likewise showed that texture cues are central to separating regions in natural images (Malik et al., 1999; Martin et al., 2004; Arbeláez et al., 2011). Recognition datasets such as DTD (Cimpoi et al., 2014), KTH-TIPS (Hayman et al., 2004) and MINC (Bell et al., 2015), together with material-recognition work in the wild (Sharan et al., 2013), tell us what texture and material categories look like, while OpenSurfaces (Bell et al., 2013), per-pixel material-context work (Schwartz & Nishino, 2013) and newer dense material segmentation datasets (Upchurch & Niu, 2022) bring surface labels into real scenes. These resources emphasize recognition, material labels or supplied targets. Our task instead couples appearance-region discovery with generated descriptions. TextureSAM (Cohen et al., 2025) moves closer to our setting by directly addressing the shape bias of SAM-family models with texture-augmented supervision built from ADE20K (Zhou et al., 2019) via Compositional Neural Textures (Tu et al., 2024). That design choice is already informative: large real-scene texture supervision is scarce, and ADE20K is one of the few dense-mask resources that can be repurposed into texture regions at meaningful scale. TextureSAM showed that SAM-style models can be shifted toward texture boundaries. Its evaluation uses automatic grid proposals and reports ground-truth-aware mask aggregation separately from unaggregated masks. Our comparison uses automatic proposals without reference-based aggregation. Detecture instead studies the upstream problem of discovering and expressing those regions before mask decoding. We reuse TextureSAM's insight that texture supervision improves appearance-sensitive segmentation, and study autonomous appearance descriptions as the interface to isolated grounding and competitive partitioning.

Language-guided dense grounding. A second line of work treats language as a handle for dense prediction. Early referring-expression segmentation asked a model to return the mask named by a supplied expression (Hu et al., 2016), and PhraseCut (Wu et al., 2020) broadened that regime to open-domain phrases in natural images. CMPC (Huang et al., 2020), CLIPSeg (Lüddecke & Ecker, 2022) and CRIS (Wang et al., 2022) strengthened language-conditioned mask prediction before the multimodal large-language-model wave. LISA (Lai et al., 2024) introduced a learnable `[SEG]` token whose hidden state conditions a SAM mask decoder; LISA++ (Yang et al., 2023) extended this line of work. GSVA (Xia et al., 2024) generalized the idea to multiple grounding tokens and explicit rejection. GLaMM (Rasheed et al., 2024) extends this family to grounded conversation and generation of descriptions that include appearance attributes of objects, stuff and parts. Sa2VA (Yuan et al., 2026) and SAMTok (Zhou et al., 2026) also support native generation with masks. These systems show that jointly producing text and masks is not new. Detecture studies a specific structured-appearance interface: it discovers descriptions, recomputes their states under block isolation, and makes the resulting mask scores compete with a learned residual. Native competitors therefore belong in the main comparison; independent thresholding and prompt-format behavior are interface choices to measure, not universal limitations of earlier models.

Entity segmentation (Qi et al., 2021) discovers objects and stuff without assigning semantic classes, with non-overlapping masks. Panoptic Narrative Grounding (González et al., 2021) aligns supplied narrative phrases to panoptic regions. These tasks establish label-free discovery and phrase-mask associations as neighboring goals; sub-semantic segmentation couples discovered appearance descriptions to a competitive partition.

Promptable and open-vocabulary segmentation backbones. A third line of work concerns the masking backbones themselves. SAM (Kirillov et al., 2023), SAM 2 (Ravi et al., 2024) and SAM 3 (Carion et al., 2025) provide strong prompt-conditioned segmentation systems; SAM 3

is especially relevant here because it adds a native text pathway, making it possible to test whether rich appearance descriptions can be grounded directly. Grounding DINO (Liu et al., 2023) plus SAM is a common detector-to-mask pipeline in which open-set detections propose regions before masking. Datasets such as COCO-Stuff (Caesar et al., 2018) make non-object regions explicit at scale, and panoptic or open-vocabulary methods such as Panoptic FPN (Kirillov et al., 2019a), Mask2Former (Cheng et al., 2022), X-Decoder (Zou et al., 2023) and OpenSeeD (Zhang et al., 2023) predict sets of masks across thing and stuff categories. Their semantic and panoptic routes use fixed or extensible class vocabularies. X-Decoder also supports referring segmentation and language outputs, so its capabilities extend beyond those routes. Detecture differs in two ways: its intermediate queries derive from structured appearance descriptions, and its goal is a partition in which multiple descriptions compete for ownership of each pixel. Table 3 summarizes how the evaluation addresses each design goal.

Design goal	Evidence and scope
Appearance cohesion	Reference-mask recovery tests inherited surface boundaries; independent judgments of within-region cohesion remain open.
Category independence	No fixed class list is required at inference. Within-class appearance transitions need independent annotation.
Open descriptiveness	The model generates appearance phrases; clause edits influence mask identities. Independent phrase-to-mask correctness remains unmeasured.
Complete, non-overlapping partition	Pixel competition enforces disjoint labels, including the residual. Measured residual area exposes incomplete described coverage.
Granularity dependence	Variable nonempty counts are evaluated under one fixed instruction. Several valid scales are not separately annotated.
Permutation equivalence	Unordered reference metrics ignore label names. Description reversal tests recomputation and reveals order sensitivity.

Table 3: How the six design goals of Section 1 guide evaluation. Output guarantees, observed behavior and remaining annotation needs are distinct parts of the task formulation.

B Additional Method and Diagnostic Details

B.1 Early VLM-to-SAM Prototype

The original development notes describe an earlier Qwen2.5-VL prototype with explicit start/end markers around texture descriptions and a multi-token projection into a SAM-family decoder. This provides the architectural trajectory toward the released design: structured descriptions, explicit grounding markers, isolated extraction and a native-width Bridge. The present experiments characterize the released Qwen3-VL-8B checkpoint. The early prototype's qualitative observations and unbound training measurements are development context, rather than matched ablation results. Appendix C records the development trail from the first cascade experiment to the released design.

B.2 Block-Diagonal Extraction

A block-diagonal attention mask keeps each description from attending to preceding descriptions during mask-state extraction. Let the assistant begin at token a, and let its grounding markers occur at $s_1 < \cdots < s_m$. With $s_0 = a - 1$, block $\mathcal{G}_j$ contains positions $s_{j-1} + 1, \ldots, s_j$. The shared prefix $\mathcal{P}$ contains positions before a, including the image, system prompt and user instruction. A token $i \in \mathcal{G}_j$ may attend to the prefix and causal positions of its own block:

$$\mathcal{A}(i) = \mathcal{P} \cup \{k \in \mathcal{G}_j : k \leq i\}.$$

This rule is used for recomputation after ordinary causal generation. It preserves the original token positions and shared image/instruction context. Description-order invariance therefore remains an empirical property, not an implication of the mask.

All assistant markers define attention blocks; only the first accepted slots, capped at six, are decoded into masks. Tokens after the last marker retain ordinary causal attention. The implementation uses a finite negative attention floor and has a separate fewer-than-two-marker path. Padding is handled

in the multi-block construction; the compact source specification records these branches. Figure 2 shows recomputation as distinct from initial generation.

B.3 Independent Mask Computation and Batch Multiplexing

This subsection concerns interactions between mask queries. The released model decodes each query independently, pairing it with the same image features. A learned residual query is prepended separately; it is not generated as a text marker. The active SAM path uses its transformer encoder and pixel-feature head, followed by the shared learned query-to-mask head. It does not call SAM's multi-slot transformer decoder. Independent lanes can be batched or processed sequentially for memory management, then their mask scores compete pixelwise.

For each lane k, the frozen SAM encoder conditions image features on query $q_k \in \mathbb{R}^{256}$. Its pixel-feature head produces $E_k \in \mathbb{R}^{256\times h\times w}$. When configured, prompt cross-attention is applied to encoder memory before these pixel features are computed. The learned mask head computes

$$u_k = W_2\,\mathrm{GELU}(W_1 q_k + b_1) + b_2, \qquad F_k(x) = W_p E_k(x) + b_p, \qquad z_k(x) = u_k^\top F_k(x).$$

All three projections have input and output width 256, and W_p is a 1×1 convolution. The head's parameters are shared across lanes, including the residual lane. The per-lane logits then enter pixelwise competition.

The frozen campaign used sequential lanes on the available device. Table 4 records a small numerical comparison with batching. The float32-storage arm permits TensorFloat-32 through the native builder; it is not an end-to-end strict-float32 experiment. This check bounds a numerical implementation difference on the tested images rather than certifying full-dataset equivalence.

Image route	Storage dtype	Max. logit error	Changed pixels (%)	Δ covering (pp)
RWTD	float32	0.004931	0.000	0.000
RWTD	bfloat16	0.066406	0.011	−0.002
TextureADE	float32	0.005498	0.008	0.001
TextureADE	bfloat16	0.093750	0.130	0.017

Table 4: Query-batching check on 2 selected images, using identical image and query activations. Covering differences and changed-pixel fractions compare batched with sequential decoding; weights remain unchanged.

B.4 Shifted-Zero Indexing

The language loss applies to assistant-region next-token targets. The marker itself retains supervision, the target immediately after a grounding marker receives zero direct weight, and ordinary token weights recover exponentially with distance. Literal marker examples in the user prompt are excluded from the assistant-marker set. This indexing matters: zeroing the marker target would weaken emission, whereas zeroing the following target reduces the pressure for its grounding state to predict the next ordinary word. Other loss and context paths still depend on that state.

The released configuration uses the exponential coefficient specified by the retained loss source. With $\alpha = 2.0$ and assistant-marker positions S, the base distance is $d_t = \min_{s\in S} |t - s|$ and the ordinary weight is $w_t = 1 - \exp(-\alpha d_t)$. The implementation then sets $w_s = 1$ for marker targets and $w_{s+1} = 0$ for their following targets, and excludes prefix/padding positions. The distance is measured from marker positions; the shift applies to the explicit zero-weight override. The source-bound objective is

$$\mathcal{L} = 1\mathcal{L}_{\mathrm{CE}}^{\mathrm{mask}} + 3\mathcal{L}_{\mathrm{Dice}} + 0.10\mathcal{L}_{\mathrm{LM}}^{\mathrm{SZ}}.$$

Mask supervision preserves the metadata ordering of reference channels. Shifted-Zero is part of the inspected released recipe; its training-time gain over alternative objectives is not identified by inference interventions.

B.5 THE BRIDGE

The Bridge resolves the width mismatch between the pretrained backbones with a single native-width projection

$$f_B(\mathbf{h}) = \mathrm{Dropout}_{0.4}(\mathrm{GELU}(\mathrm{LN}(W_B\mathbf{h} + b_B))),$$

where $\mathbf{h} \in \mathbb{R}^{4096}$ and $f_B(\mathbf{h}) \in \mathbb{R}^{1024}$. The subsequent $1024 \to 256$ mapping is SAM 3's frozen text resizer. Dropout is disabled for inference. Earlier design notes considered narrower and higher-capacity projections; the available final-checkpoint evidence supports the active mapping, not superiority over those trained alternatives. The native SAM text-encoder comparison changes both the text representation and its downstream interface and is reported as a separate pipeline.

B.6 TRAINING CURRICULUM

The inspected configuration has two stages: first adapt the Bridge, mask head and residual embedding, then add LoRA on language-model attention query/value projections and masked grounding-token rows. The training prompt supplies a fixed range rather than each image's region count. Teacher forcing provides reference descriptions and their termination, with no separate count head. The evaluation instruction instead uses the constant word "multiple." These are source/configuration facts; a complete historical example inventory, optimizer state and checkpoint-selection trajectory are unavailable.

Setting	Retained configuration
Optimizer / weight decay	AdamW / 0.01
Nominal epochs / linear warmup	30 / 5
Microbatch / gradient accumulation	2 / 4
Base rate / cosine minimum	10^{-4} / 10^{-6}
Stage-1 adaptation	8 epochs; Bridge, mask head and residual adapted
Stage-2 Bridge rate multiplier	0.1
Qwen rate multiplier	0.01
Marker-row rate multiplier	1.0
LoRA rank / alpha / dropout	8 / 16 / 0.0
LoRA target projections	`q_proj`, `v_proj`
Gradient norm cap	1.0
SAM square input side	1008

Table 5: Retained Detecture training recipe, extracted from configuration and scheduler source. Rates are configured values, not recovered per-step trajectories. The nominal epoch budget does not establish the selected checkpoint's training duration. No part of this recipe was executed for the frozen-weight evaluation.

Table 6 separates stored tensors from elements eligible for task gradients. The full vocabulary matrices enter the configured optimizer group, although a gradient hook masks all but the grounding rows. Row masking alone does not establish invariance under weight decay. Direct comparison with the base model finds the other rows unchanged at stored precision; missing optimization history prevents reconstructing the reason. The residual embedding's aliased checkpoint entry is counted once. No optimizer, backward operation or training continuation was executed for this revision.

Group	Task-gradient eligible	Stored elements
Qwen LoRA	3,833,856	3,833,856
Bridge	4,197,376	4,197,376
Mask head	197,376	197,376
Residual embedding	4,096	4,096
Vocabulary tensors	8,192	1,242,480,640
Total	8,240,896	1,250,713,344

Table 6: Released-checkpoint parameter accounting. Task-gradient eligibility follows the inspected historical row mask; stored elements count checkpoint tensors. These counts do not imply new training or reconstruct the historical optimizer trajectory.

C Ablation: From a Two-Model Cascade to One Model

The first experiment was a cascade. Our first attempt at autonomous appearance segmentation chained two strong pretrained models without training either of them. A vision-language model wrote a list of appearance descriptions, a parser split that text into separate prompts, and SAM 3 grounded each prompt through its own text encoder. The base-Qwen/SAM 3 cascade in Appendix M instantiates this design with the current backbones (Qwen3-VL-8B-Instruct and SAM 3's semantic head). It is a strong system, reaching 70.1, 85.6, 97.9 and 83.5 covering on TextureADE, RWTD-COCO, CSTD and RWTD (Table 22). We were nevertheless not satisfied with the approach, for three reasons. First, the two models communicate only through text, so a parser sits on the critical path between them: a description it fails to parse yields no mask. Second, SAM 3 re-encodes each parsed description from its words alone, as a separate prompt, discarding the image-conditioned state in which the language model produced it. Third, the result is two separately run models rather than one network whose descriptions and masks are produced together. These limitations led to the architecture of Section 2, which folds the two models into one: the language model's own grounding-marker states, mapped through the Bridge, are SAM 3's queries. Parsing survives only in the optional name-removal step, and an item that cannot be parsed is still grounded from its unedited description.

Folding the models has a cost the cascade does not pay. Detecture joins the two backbones only through the Bridge, a single learned projection followed by SAM 3's frozen text resizer (Appendix B), whereas the cascade reaches SAM 3 through its full native text encoder. On two-region scenes that richer text path is an advantage. Detecture-minus-cascade covering differences are -10.2 points on CSTD and -4.0 on RWTD, and the RWTD-COCO difference (-0.4) is unresolved (Table 23). Routing Detecture's own descriptions through the same native pipeline (condition C9 of the historical intervention study, Table 16) shows the effect from the other side: it changes RWTD covering by +5.3 points. On multi-region TextureADE the order reverses. Detecture leads the cascade by 5.6 points, and C9 changes TextureADE covering by -2.6 points (Table 18). We therefore regard the cascade as a legitimate reference that the lightweight Bridge does not reach on two-region scenes (Section 5), not as a baseline Detecture should be expected to beat everywhere. Because C9 replaces the whole text pathway rather than the projection alone, neither comparison isolates the Bridge.

Wiring the backbones together was not enough. The system of Section 2 is easy to summarize as "a vision-language model wired to a promptable segmenter," and that summary is accurate about the components. It is not accurate about the difficulty. The remainder of this section records the development trail of the integrated model in order, because the intermediate configurations answer a question the final numbers cannot: what happens if the two backbones *are* simply wired together. The short answer is that the naive coupling does not merely work at reduced quality; on the live task it does not work at all.

How to read these numbers. Every number below is a development measurement on RWTD taken under the *exact-K=2* instruction and reported as the original paper's matched-only mIoU (Appendix K); several are teacher-forced. They are therefore not comparable to Section 3.2, whose covering scores come from a protocol in which no method is given the number of regions, and they are not evidence of superiority over any baseline. Like the prototype observations in Appendix B, they are development records rather than matched, source-bound ablation results: internal comparisons under a fixed protocol, useful for attributing gains to mechanisms and for nothing else. Two regimes appear. In *teacher-forced* rows the ground-truth descriptions are supplied, so only the mask pathway is under test. In *live* rows the model must emit its own descriptions and grounding tokens, so the whole system is under test. Keeping the two apart is the point: the trail's central lesson is that they diverge.

The integrated starting point is a naive coupling, and it scores $\mathbf{0.544}$. The first working configuration of the integrated model was exactly the system a reader might reasonably assume: a high-capacity projector from Qwen3-VL-8B's hidden width into the decoder, several `[SEG]` tokens read out of the ordinary causal stream, and a full language-modeling loss over the whole assistant response. It reaches 0.544 RWTD mIoU. Nothing about it is broken in an obvious way; it simply carries three defects at once, each of which the following steps isolate.

Prompt-slot competition costs the most and is the cheapest to fix ($\mathbf{0.544 \rightarrow 0.703}$). Feeding several learned prompts to the segmentation decoder in one call does not produce several masks. A slot diagnostic measured the allocation directly: on RWTD the first slot received 90.5% of the pixels

and the second 0.0%, and the pattern survived swapping the order of the textures, so it is positional rather than semantic. Decoding every texture query in its own independent lane (Appendix B) recovers 0.159 mIoU. This is the single largest gain in the trail.

Cross-region contamination is the second defect ($0.732 \rightarrow 0.810$, teacher-forced). When the grounding states are read from the ordinary causal stream, each one has attended to every preceding texture description, so the state that is supposed to mean "coarse gray gravel" also carries "dense green foliage." The cosine similarity between grounding states was 0.74 under shared context against 0.16 under isolated extraction: the states were nearly collapsed onto one another. Recomputing them under a block-diagonal attention mask, so that each description attends to the shared prefix and to itself but not to its siblings, adds a further 0.078.

Between these two sits the weakest link in the trail ($0.703 \rightarrow 0.732$). A slim bottleneck in place of the high-capacity projector adds 0.029, the smallest gain of any mechanism, and we now regard even that attribution as unsafe. The projection was later replaced by the single native-width Bridge, and the C9 comparison above replaces the whole text pathway rather than the projection alone, so it cannot isolate the projection either. We report the row for completeness and draw no conclusion from it. The projection is plumbing between two hidden widths, not a contribution.

Then the trail breaks, and this is the most informative part of it. At 0.810 the mask pathway was healthy under teacher forcing. Running the same model family live, with no language supervision on the generated descriptions, collapses it to 0.136. The geometry had never been the binding constraint. Re-weighting the language loss with a cosine decay recovers most of it (0.694 live), but the natural next step is a trap: zeroing the loss weight exactly at the `[SEG]` position, so that the grounding state is never asked to predict ordinary continuation text, is clean in principle and catastrophic in practice. It scores $\mathbf{0.000}$ because the output is absent, not degraded: the model learns that emitting `[SEG]` is free of gradient and stops emitting it at all. The Shifted-Zero schedule, which keeps full weight on each `[SEG]` target and zeroes the target immediately after it (Appendix B), together with a masked-row warm start for the `[SEG]` embedding, restores emission and reaches 0.794 live. The final checkpoint reaches 0.8162 under this instruction and 0.7843 under the original paper's count-free shared-format instruction (Appendix M).

What the trail does and does not license. It does not establish that our particular mechanisms beat alternatives that solve the same failures differently; the matched retrained controls that would settle this are not part of this submission (Section 5). It does establish something narrower and more useful: the distance between a naive coupling of these two backbones and a working autonomous partitioner is not a matter of tuning. Two of these defects are visible under teacher forcing and would be caught by any careful implementer; the one that destroys the system outright (0.000) is invisible in any measurement that is not run live, which is precisely why it survived into a configuration that looked principled on paper. A coupling that scores 0.544 in its best case and 0.136 in its realistic one is not a weaker version of the final model. Whether the specific fixes generalize beyond texture is untested. That the coupling needs fixing at all is not.

D Frozen-Weight Interventions

Which parts of the released computation influence its output? The original development trail motivated isolation, independent queries, emission supervision and width adaptation. We test the available checkpoint with frozen-weight interventions, leaving alternative trained objectives and architectures as future work. This separates effects on the current computation from the effects of retraining a different system.

Table 7 uses the fixed diagnostic sample of 32 images per route, chosen by hashed identifiers before viewing intervention outcomes. Text and query controls compare with the same image's unedited recomputation. Removing isolation lowers primary covering on three routes after correction; its RWTD-COCO effect remains unresolved. Reversing complete descriptions changes TextureADE covering by -10.6 percentage points. Thus independent-block extraction does not make the released model insensitive to output order. Nonsense and cross-image descriptions test content dependence; direct attention and zero-image controls test specific visual paths.

Identical positive queries create identical deterministic positive-mask logits up to numerical effects; argmax ties then collapse labels by construction. This control removes query differences but cannot

Fixed-text intervention	RWTD	RWTD-COCO	TextureADE	CSTD
Unedited model (covering, %)	77.8	87.9	72.9	86.5
Remove block isolation	**−17%** −13.4 [−19.7, −7.1]	−8% −7.2 [−13.4, −1.2]	**−15%** −11.0 [−15.2, −7.1]	**−9%** −7.6 [−12.8, −3.0]
Reverse complete descriptions	0% 0.1 [−6.7, 7.1]	−4% −3.3 [−6.5, −0.3]	**−15%** −10.6 [−16.6, −4.9]	−3% −3.0 [−12.0, 6.4]
Appearance clauses only	**−8%** −6.1 [−10.9, −1.9]	0% −0.2 [−1.8, 1.9]	**−5%** −3.5 [−6.3, −1.0]	−5% −4.0 [−8.4, 0.0]
Spatial clauses only	**−15%** −11.7 [−18.0, −5.3]	−6% −5.5 [−11.0, −0.6]	**−11%** −8.4 [−11.7, −5.1]	**−10%** −9.1 [−16.1, −3.2]
Nonsense clauses	**−23%** −17.7 [−24.5, −10.7]	**−24%** −21.0 [−29.0, −12.5]	**−41%** −29.7 [−35.4, −23.9]	**−21%** −18.2 [−26.2, −9.8]
Cross-image descriptions	**−26%** −20.2 [−28.1, −11.9]	**−24%** −20.8 [−27.6, −13.8]	**−21%** −15.5 [−21.7, −9.4]	**−15%** −12.7 [−20.9, −4.1]
Block direct image attention	**−18%** −13.8 [−20.7, −7.4]	**−6%** −5.5 [−9.3, −2.4]	**−8%** −5.8 [−9.3, −2.3]	−10% −8.2 [−15.4, −0.6]
Zero VLM image input	−12% −9.2 [−17.2, −1.4]	**−34%** −29.7 [−35.6, −23.5]	**−31%** −22.9 [−29.0, −17.0]	**−11%** −9.7 [−16.4, −2.9]
Identical region queries	**−24%** −18.7 [−25.5, −11.8]	**−42%** −36.5 [−40.9, −31.5]	**−50%** −36.7 [−41.5, −31.9]	**−32%** −27.4 [−33.1, −21.1]

Table 7: **Frozen-weight interventions.** Each condition changes one part of the released computation on the same fixed sample of 32 images per dataset. The first row gives the unedited model's bidirectional covering (%). Each intervention cell shows the relative change in covering, as a percentage of that unedited score (negative means worse), and below it the absolute change in covering points with its marginal 95% image-bootstrap interval. For example, removing block isolation changes RWTD covering by −17% of its unedited value (−13.4 points from 77.8). Bold: significant after multiple-comparison correction (Table 8). These are inference interventions, not retrained architectural ablations.

Fixed-text intervention	RWTD	RWTD-COCO	TextureADE	CSTD
Remove block isolation	**0.002 / 0.002**	0.141 / 0.141	**<0.001 / <0.001**	**0.024 / 0.024**
Reverse complete descriptions	0.971 / 0.971	0.141 / 0.141	**0.004 / 0.004**	0.532 / 0.532
Appearance clauses only	**0.031 / 0.031**	0.884 / 0.884	**0.014 / 0.014**	0.147 / 0.147
Spatial clauses only	**0.006 / 0.006**	0.141 / 0.141	**<0.001 / <0.001**	**0.028 / 0.028**
Nonsense clauses	**<0.001 / <0.001**	**<0.001 / <0.001**	**<0.001 / <0.001**	**0.002 / 0.002**
Cross-image descriptions	**0.001 / 0.001**	**<0.001 / <0.001**	**<0.001 / <0.001**	**0.037 / 0.037**
Block direct image attention	**0.002 / 0.002**	**<0.001 / <0.001**	**0.007 / 0.007**	0.114 / 0.114
Zero VLM image input	0.065 / 0.065	**<0.001 / <0.001**	**<0.001 / <0.001**	**0.037 / 0.037**
Identical region queries	**<0.001 / <0.001**	**<0.001 / <0.001**	**<0.001 / <0.001**	**<0.001 / <0.001**

Table 8: **Statistical significance of the Table 7 changes.** Each cell gives two p-values from paired sign-flip tests of whether the change differs from zero: the first resamples images, the second resamples reviewed scene clusters. Both are Holm-adjusted across the nine interventions within each dataset. Lower is stronger evidence; bold marks cells where both values are below 0.05, i.e., the change is significant after correction.

by itself establish their linguistic meaning. Zeroing the vision-language model's image tensor still preserves fixed descriptions derived from the image and the separate SAM image pathway. The historical clause-swap study in Appendix L tests the more specific question of whether mask identities follow appearance or spatial wording.

E Constructing TextureADE

This section expands the concise main-text description of TextureADE construction in Section 3.1. We first state the implemented score, then define each component as a deterministic function of the image's dense semantic mask, followed by the optional image-based corrections. Throughout, H and W are the image height and width, and $|I| = HW$ is the total pixel count. The ADE adapter assigns semantic names to `TEXTURE_SURFACE`, `OBJECT` or `AMBIGUOUS` using a fixed keyword taxonomy: object keywords take precedence, followed by texture and ambiguous keywords, with

unmatched names defaulting to OBJECT. Thus a texture class is a positive taxonomy assignment, rather than simply any class outside an object list.

Base review score. For an input image, Detecture Miner computes

$$s_{\text{base}} = \text{clip}_{[0,100]}\big(0.20\, s_A + 0.60\, s_{\text{geom}} + B_{\text{tex}} + B_{\text{reg}} + B_{\text{edge}} - P_{\text{obj}} - P_{\text{amb}} - P_{\text{clutter}}\big). \quad (1)$$

Writing $\text{clip}_{[a,b]}$ for clipping to $[a, b]$, the exact additive terms are

$$B_{\text{tex}} = 20\,\text{clip}_{[0,1]}\big((\phi_{\text{tex}} - 0.65)/0.35\big), \qquad B_{\text{reg}} = \begin{cases} 6, & 2 \le C_{\text{large}} \le 4, \\ 2, & C_{\text{large}} > 4, \\ -8, & C_{\text{large}} < 2, \end{cases}$$

$$B_{\text{edge}} = \begin{cases} 3, & C_{\text{strong}} \ge 1, \\ -4, & C_{\text{strong}} = 0, \end{cases} \qquad P_{\text{obj}} = 180\,\text{clip}_{[0,1]}(\phi_{\text{obj}} - 0.002),$$

$$P_{\text{amb}} = 120\,\text{clip}_{[0,1]}(\phi_{\text{amb}}), \qquad P_{\text{clutter}} = 4\,\text{clip}_{[0,99]}(C_{\text{large}} - 5).$$

Here ϕ_{tex}, ϕ_{obj} and ϕ_{amb} are the corresponding pixel fractions; the component counts are defined below. The extraction configuration thresholds the final review score at 65.

Mask-structure score s_A. For the ADE adapter, let N be the number of nonzero semantic classes present in the dense mask, let $\{a_i\}_{i=1}^{N}$ be their total pixel areas, and let $r_i = a_i/|I|$. All pixels of a class contribute to the same a_i, including disconnected instances. Define $r_{\max} = \max_i r_i$, $r_{\text{med}} = \text{median}_i\, r_i$ and $\phi_{\text{small}} = \frac{1}{N}\sum_i \mathbf{1}[r_i < 10^{-3}]$. The Shannon entropy of the area distribution (base 2) is

$$H(\{r_i\}) = -\sum_{i=1}^{N} \tilde{r}_i \log_2 \tilde{r}_i, \qquad \tilde{r}_i = \frac{r_i}{\sum_j r_j}. \quad (2)$$

Five normalized component scores are then defined, each clipped to $[0, 1]$:

$$\begin{aligned} n_{\text{score}} &= \min\big(1,\, N/120\big), & \ell_{\text{score}} &= 1 - \min\big(1,\, r_{\max}/0.65\big), \\ m_{\text{score}} &= 1 - \min\big(1,\, r_{\text{med}}/0.03\big), & p_{\text{score}} &= \min\big(1,\, \phi_{\text{small}}/0.55\big), \\ e_{\text{score}} &= \min\big(1,\, H(\{r_i\})/4.0\big). \end{aligned} \quad (3)$$

The mask-structure score is their convex combination, expressed on a 0–100 scale:

$$s_A = 100 \cdot \big(0.28\, n_{\text{score}} + 0.24\, \ell_{\text{score}} + 0.18\, m_{\text{score}} + 0.18\, p_{\text{score}} + 0.12\, e_{\text{score}}\big). \quad (4)$$

Intuitively, n_{score} rewards a rich class-mask inventory, ℓ_{score} penalizes a single dominating class, m_{score} penalizes a coarse median, p_{score} rewards small class masks and e_{score} rewards an entropically balanced area distribution. These are semantic-class statistics in this adapter, distinct from the connected components used next.

Texture-boundary geometry score s_{geom}. Let $\mathcal{R} = \{R_1, \ldots, R_M\}$ be the components obtained by eight-connectivity labeling separately within each texture class, and let $\mathcal{L} = \{R_j : a(R_j) \ge 0.08|I|\}$ be the subset of *large* components. Define:

- $C_{\text{large}} = |\mathcal{L}|$, the count of large textured regions.
- $C_{\text{strong}} = \big|\{(R_i, R_j) \in \mathcal{L} \times \mathcal{L},\, i < j : |\partial(R_i, R_j)| \ge 40\}\big|$, where shared boundaries count horizontal and vertical pixel adjacencies.
- $\beta = \frac{1}{2(H+W)} \sum_{R_i, R_j \in \mathcal{L},\, i<j} |\partial(R_i, R_j)|$, total shared boundary length between large components, normalized by image perimeter.
- $b = 1 - \frac{\max_{R_j \in \mathcal{L}} a(R_j)}{\sum_{R_k \in \mathcal{L}} a(R_k)}$ when at least two large components exist, and $b = 0$ otherwise.
- $\phi_{\text{obj}} \in [0, 1]$, the object-taxonomy pixel fraction, used as a penalty multiplier.

The four positive terms are clipped to $[0, 1]$ and combined:

$$\begin{aligned} \rho_{\text{score}} &= \min\big(1,\, C_{\text{large}}/3\big), & \sigma_{\text{score}} &= \min\big(1,\, C_{\text{strong}}/2\big), \\ \beta_{\text{score}} &= \min\big(1,\, \beta/0.35\big), & b_{\text{score}} &= b, \end{aligned} \quad (5)$$

$$s_{\text{geom}}^{\text{raw}} = 100 \cdot \big(0.28\,\rho_{\text{score}} + 0.34\,\sigma_{\text{score}} + 0.26\,\beta_{\text{score}} + 0.12\,b_{\text{score}}\big), \tag{6}$$

$$s_{\text{geom}} = s_{\text{geom}}^{\text{raw}} \cdot \big(1 - \min(0.75,\ 1.15\,\phi_{\text{obj}})\big). \tag{7}$$

Equation (6) rewards several large textured regions, strong inter-region boundaries, substantial total boundary length relative to image perimeter, and balanced large-region areas. Equation (7) damps the score on object-dominated scenes without making that multiplier zero.

Acceptance rule and post-acceptance extraction. The pipeline is geometry-first, with optional Contrastive Language-Image Pretraining (CLIP) and VLM corrections before thresholding. With CLIP similarity available, the score becomes $0.92\,s_{\text{base}} + 0.08\,s_{\text{CLIP}}$; with a VLM score available it becomes $0.90\,s_{\text{base}} + 0.08\,s_{\text{CLIP}} + 0.02\,s_{\text{VLM}}$. CLIP scores are scaled to $[0, 100]$ using their empirical lower and upper quantiles, with range/constant-score fallbacks in the implementation. VLM flags subtract 2.0 for object-centric content, 1.4 for too many objects, 1.6 each for low texture content or no clear texture boundary, and 2.4 each for mosaic/collage or synthetic/graphic content. A VLM match decision adds 2.0 and a non-match subtracts 2.0; the final score is clipped to $[0, 100]$. The base training-data configuration enables both auxiliary stages. Thus a VLM may influence selection as well as supply later language labels.

The extraction configuration accepts scores of at least 65 and then pools all pixels of each nonzero semantic class. It retains the largest classes that cover at least 1% of the image, up to 5 masks, requiring at least 2 masks for an item. This extraction stage does not apply the texture taxonomy again or absorb small fragments into a nearest neighbor; excluded pixels remain in the residual/background label. Scoring therefore precedes this class-mask extraction.

Dual-source mining: training vs. evaluation. The construction has two complementary sources. The training route uses TextureSAM's CNT-augmented ADE20K imagery (Tu et al., 2024; Cohen et al., 2025), while TextureADE evaluates natural ADE20K scenes. The released source includes two augmented adapters: one scores original ADE images and maps each source score to its textured variants; the full textured-images adapter caches shared annotation geometry but can compute CLIP/VLM scores separately for each variant. A further degree-expansion step can include higher-degree variants above the earliest passing degree while reusing their source masks. Consequently, these routes should not be described as one identical, strict per-variant filtering operation. The Detecture training configuration points to the full textured-images metadata; the executed training and validation manifests were not recovered, so their exact sizes, expansion history and source-image exposure remain unverified.

The pinned natural-image TextureADE release contains 212 images and 828 region masks with nonempty descriptions, a mean of 3.91 regions per image. Its source IDs identify 195 ADE training images and 17 validation images. This is an ADE-derived evaluation set; those source IDs do not establish disjointness from the missing Detecture training manifest.

Claude Haiku 4.5 (model ID `claude-haiku-4-5-20251001`) is the configured language annotator. Its fixed pattern is "Texture of <semantic name>, <distinct visual features>, <spatial context>", targeting 10–15 words after "Texture of". The full image, region overlay, region count and available ADE class-name hints condition a joint description response. Claude is used here as a language-labeling tool: it describes the supplied masks rather than inventing their pixels. The resulting supervision therefore contains a semantic-name field; any removal of that field for inference is a separate model operation.

Throughout the paper, Detecture refers to the model, Detecture Miner to the curation pipeline and TextureADE to the ADE20K-derived dataset contribution produced by that pipeline.

TextureADE motivation and source dependence. Texture segmentation has a data bottleneck. RWTD provides real-world appearance partitions but is too small to support broad training studies on its own. Other resources (DTD, KTH-TIPS, MINC, OpenSurfaces, dense material segmentation datasets) are organized around categories or surface labels rather than blind multi-region texture partitioning (Cimpoi et al., 2014; Sharan et al., 2013; Bell et al., 2015; 2013; Upchurch & Niu, 2022). TextureADE addresses this scarcity by mining an appearance-oriented route from ADE20K, a canonical dense-mask scene-parsing corpus. The augmented route provides region/description supervision for learning a language-to-mask interface.

The training corpus and TextureADE evaluation share an ADE source family. Their roles differ: TextureADE probes operation on natural scenes from this source family, while RWTD-COCO

and CSTD provide complementary construction domains, with RWTD adding external validation. Source separation alone does not certify absence of pretraining exposure or development use, and the unrecovered Detecture training inventory prevents calling TextureADE a verified held-out set. As TextureSAM also used ADE-derived texture supervision (Cohen et al., 2025), mining useful supervision from an existing dense-mask corpus remains a practical part of the contribution. The scientific limit is that inherited semantic masks cannot independently establish within-class appearance boundaries.

The division of labor is deliberate: Detecture Miner constructs candidate regions and descriptions; Detecture learns to discover and ground regions at inference time. This supplies an implementable supervision route for the proposed task, while independent appearance annotation remains necessary to validate its full category-independent scope.

F THE RWTD-COCO BENCHMARK

This appendix documents RWTD-COCO construction, release structure, curated-anchor balancing and the historical supervised learnability study. The construction and trained ranker/U-Net records below predate this revision, which preserves and checks their existing metadata and scalar results. They are distinct from the fixed-checkpoint autonomous comparisons in the main paper.

Why the benchmark is needed. The Real-world Texture Dataset (Khan et al., 2015; Cohen et al., 2025) exemplifies the regime we care about: real photographs whose target is an image-specific partition into a few regions coherent in surface appearance, where region A need not carry the same name across images. Its scale reflects the cost of finding suitable images and tracing the intended appearance boundary. Larger resources solve adjacent problems instead: DTD (Cimpoi et al., 2014) describes texture attributes, MINC (Bell et al., 2015) recognizes materials, OpenSurfaces (Bell et al., 2013) catalogs surface appearance, and general contour benchmarks (Arbeláez et al., 2011) mix boundaries arising from shape, occlusion and illumination. These resources do not directly supply the intended image-specific two-region appearance partitions at scale.

The key observation. Dense semantic corpora contain reusable human boundary annotations. If one restricts attention to surface labels, rejects things and clutter, and crops around a shared boundary using annotation geometry, complex scenes become candidates for simple appearance-partition examples. This reuses existing annotations to construct suitable image-specific targets without fitting a segmentation model to generate target pixels. It remains a proxy: a semantic boundary need not be a strong appearance transition, and one semantic region may contain multiple appearances.

F.1 CONSTRUCTION

Let $Y \in \{0, \ldots, 181, 255\}^{H \times W}$ be a COCO-Stuff (Caesar et al., 2018) thing+stuff map, where 0–90 are things and 255 is void, and let $\mathcal{S}$ be the hand-curated whitelist of 28 surface-like stuff labels (blanket, carpet, cloth, dirt, floor-marble, floor-other, floor-stone, floor-tile, floor-wood, grass, gravel, ground-other, mat, moss, mud, pavement, rock, rug, sand, stone, textile-other, wall-brick, wall-concrete, wall-other, wall-stone, wall-tile, wall-wood, wood). These are the source PNG IDs; official label IDs stored in release metadata are one greater. All thing labels are excluded from candidate pairs and penalized wherever they occur; sky, water and every other non-pair label count as nuisance rather than clean background.

Candidate pairs and boundary-centered crops. A pair $a, b \in \mathcal{S}$ is retained when each covers at least 2.5% of the map and the pair shares at least 12 four-neighbor edges. Shared-edge coordinates are projected onto their first principal axis; candidate centers are the mean boundary coordinate plus 5 fixed quantiles along that axis. Around each center we enumerate boxes with short side in $\{192, 256, 320, 384, 448, \min(H, W)\}$ and aspect ratio in $\{3/4, 1, 4/3\}$, clipped to the image, with minimum short side 176. Duplicate boxes are removed and the first 75 lexicographically sorted boxes per pair are considered. Cheap eligibility requires at least 6% support for each label and 8 shared edges. This crop search is deterministic and annotation-driven; RGB is read subsequently for materialization checks and for the separate balancing step below.

Structural score. Within a crop C, let p_a, p_b be the two label fractions, $c = p_a + p_b$ the pair coverage and r the balance,

$$r = \frac{2\min(p_a, p_b)}{p_a + p_b}. \tag{8}$$

Let t, n, v be thing, other-stuff nuisance and void fractions; let β be the number of four-neighbor a/b shared edges divided by $\sqrt{|C|}$; let q be the smaller of the two regions' connectedness values (largest eight-connected-component area over total label area); and let the fragmentation term be $f = \min(1, \max(0, c_a + c_b - 2)/20)$, where c_a, c_b count their components. The score is

$$s(C) = 3c + 1.4r + 0.8\,\frac{\min(\beta, 1.5)}{1.5} + 0.8q - 5t - 3n - 2v - 0.4f. \tag{9}$$

The coefficients express the dataset definition rather than a fit to observed RGB or to benchmark performance. Cheap area and boundary statistics shortlist boxes using $q = 0.75$ and $f = 0$; full connected-component analysis is applied to the best 14. Tier-1 candidates are preferred, then the score breaks the ranking; one crop is kept per source image.

Tiers and masks. Table 9 gives the two threshold presets. Candidates are ranked, capped per pair and limited to one crop per source image. Materialization computes an exact SHA-256 hash of the uncompressed RGB and a perceptual hash, and rejects duplicates or nearly degenerate RGB crops; the archived materialization summary records no such rejections. The recommended ground-truth format maps $A \to 1$, $B \to 2$ and nuisance to 0 (ignore); a binary view ($A = 255$) is provided for compatibility, and the untouched cropped source map is preserved so a release row can be reconstructed independently. No model-predicted pixel enters the ground truth.

Criterion	Tier-1	Tier-2	Criterion	Tier-1	Tier-2
$a + b$ coverage, min	0.920	0.840	Void, max	0.020	0.050
Each region, min	0.170	0.130	Boundary norm, min	0.070	0.045
Thing area, max	0.015	0.030	Connectedness, min	0.600	0.500
Other stuff, max	0.080	0.140	Components/region, max	12	24

Table 9: **Structural thresholds.** Tier-2 contains additional crops that fail Tier-1 but pass the looser constraints. Values are generated from the pinned construction configuration.

F.2 Release and Structural Character

The archived scan covers $123{,}287$ training and validation maps and finds $3{,}703$ Tier-1 and $2{,}098$ additional Tier-2 source images before size and diversity caps; the release materializes $1{,}000$ and $1{,}500$, respectively. All $2{,}500$ source IDs and recorded uncompressed RGB hashes are unique. An extra-strict preset still returns $2{,}580$ eligible crops (Figure 8, left), supporting useful yield under stricter constraints. The right panel shows the residual semantic imbalance that motivates the balancing step below: pair caps prevent domination but do not make the release class-balanced.

Table 10 reports the annotation-derived structure of each tier. These statistics operationalize the intended "two dominant fields" character more directly than semantic category counts. The release contains 163 distinct pairs in Tier-1, 191 in Tier-2 and 220 in their union.

Metric	Tier-1 (1,000)	Tier-2 (1,500)
Pair coverage	0.991 / 1.000 / 1.000	0.870 / 0.918 / 0.980
Smaller-region fraction	0.468 / 0.490 / 0.499	0.221 / 0.387 / 0.460
Thing fraction	0.000 / 0.000 / 0.003	0.000 / 0.009 / 0.027
Other-stuff fraction	0.000 / 0.000 / 0.003	0.000 / 0.058 / 0.120
Boundary norm	1.430 / 1.702 / 2.433	1.125 / 1.649 / 3.199
Connectedness	1.000 / 1.000 / 1.000	0.775 / 1.000 / 1.000

Table 10: **Annotation-derived structure of the released tiers.** Values are 10th / 50th / 90th percentiles recomputed from the archived manifest measurements; headers give the number of released crops per tier.

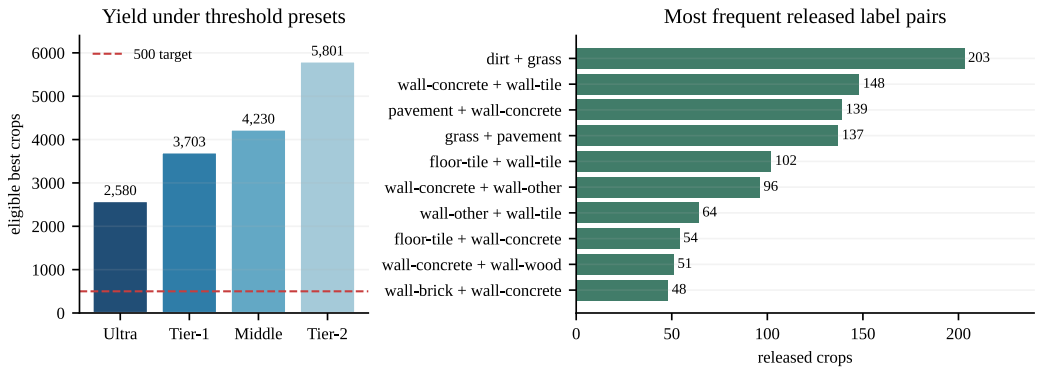


Figure 8: **Scale and composition.** Left: archived annotation-only yield across threshold presets. Right: most frequent label pairs after release caps, combining both tiers.

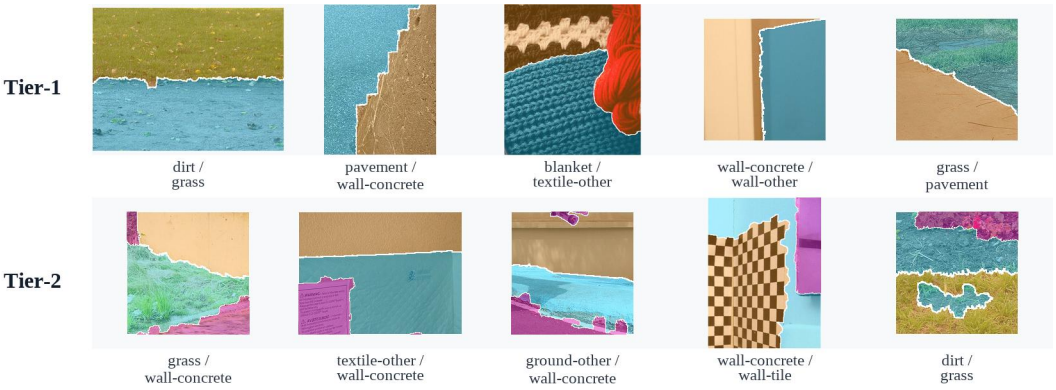


Figure 9: **The two release tiers.** Tier-1 examples are almost pure two-region partitions. Tier-2 permits a small third/nuisance region (magenta) while preserving a dominant human-annotated A/B boundary. Structurally clean textured silhouettes are retained; the construction does not require a straight or background-only boundary.

F.3 Curated-Anchor Balancing

The mined reservoir is deliberately broad, but a broad reservoir is not a benchmark distribution: frequent COCO transitions such as grass/dirt dominate a naively ranked subset. The historical construction therefore uses 212 curator-selected examples as positive anchors and treats the other 2,288 released crops as *unlabeled* rather than negative, since absence from the curated set does not imply rejection.

Each crop is described by 14 structural features and 9 lightweight RGB diagnostics (including inter-region Lab separation, texture-energy and roughness differences, shared-boundary contrast and gradient, gradient-histogram divergence, entropy and color variation). A 25-bag class-balanced logistic model estimates positive-unlabeled preference and a shrinkage Mahalanobis distance measures proximity to the multivariate anchor profile. Each bag uses all anchors and a threefold sample of unlabeled crops, with standardized features and logistic regularization $C = 0.35$. The combined score is 0.75 times the mean positive-unlabeled probability plus 0.25 times prototype similarity. Semantic label, exact-pair and family identities are excluded from this preference model and enter only through the subsequent balancing objective. A separate 5-fold single-logistic proxy gives 0.779 ROC-AUC and 0.239 average precision. These values describe separability of the curation view, not texture-boundary accuracy or validation of the entire selection pipeline, because the unlabeled pool was never exhaustively rejected.

An admissible extension must lie inside anchor-calibrated envelopes for coverage, smaller-region support, thing/nuisance area, connectedness, fragmentation, boundary clarity and profile distance, and within the top 30% of unlabeled combined preference scores. The archived threshold is 0.5134 (rounded here), leaving 535 eligible crops. A deterministic greedy objective combines 68% anchor-aligned quality with 32% diversity reward, subject to caps on exact pairs, family pairs and individual labels, over 6 coarse provenance families (earth/vegetation, masonry/tile, mineral, other surface, textile, wood) spanning 21 unordered family-pair types. The diversity weights are 0.14 for family balance, 0.09 for exact-pair balance, 0.06 for label balance and 0.03 for a previously absent pair; count-based rewards decrease as a group fills. Caps preserve any concentration already present in the anchor set. Neither the ranker nor the balancing step alters a target pixel.

View	Images	Anchors	Additions	Exact pairs	Effective pairs	Family entropy
Curated core	212	212	0	103	83.4	0.876
Ultra-conservative	128	128	0	90	82.1	0.908
Balanced-256	256	212	44	136	110.7	0.913
Expanded-454	454	212	242	162	129.1	0.913

Table 11: **Benchmark views.** Effective pairs are $\exp(H)$ for exact-pair Shannon entropy H; family entropy is normalized within the 21 family-pair types. **Balanced-256 is the RWTD-COCO view evaluated in the main paper.** The historical U-Net study below instead holds out the curated core.

Balanced-256 retains all 212 anchors and adds 44 examples, raising exact-pair support from 103 to 136 and effective exact-pair diversity from 83.4 to 110.7, while the largest family-pair share falls from 17.5% to 14.5% (Table 11, Figure 10). Median coverage is 1.000 versus 0.999 for the anchors, median connectedness 1.000 versus 1.000, and median boundary clarity 0.540 versus 0.538. The additions' appearance-contrast standardized mean difference from the anchors is 0.017. These are operational matching results, not a formal claim of distributional equivalence; other measured feature shifts remain in the archived comparison table.

F.4 Leakage-Audited Learnability

Clean structure does not by itself establish that the RGB images support a learnable appearance partition, so the historical release includes a supervised protocol that is harder than a random image split. The 212 curated crops are removed before splitting and held out as the curated test. From the remaining $2{,}288$ sources, complete unordered label pairs are held out until reaching 303 images, producing 35 pair types with zero training overlap while every one of the 28 individual labels remains represented in training. The remainder is stratified by tier and family pair into $1{,}485$ training, 250 validation and 250 random-test images. The preserved split files are source-image disjoint and jointly cover the reservoir. Here "leakage-audited" refers to those source and pair checks; it does not certify absence of foundation-model pretraining exposure or every perceptual near-duplicate.

Because the target is an *unordered* bipartition, a prediction and its complement are equivalent. Let y be the binary target, z the predicted logits and σ the sigmoid function. With binary cross-entropy (BCE) on logits, training chooses the smaller per-image masked objective under label exchange,

$$\min_{\pi \in \{y, 1-y\}} \big[\mathrm{BCE}(z, \pi; m) + 1 - \mathrm{Dice}(\sigma(z), \pi; m)\big],$$

and averages over the batch, using valid pixels m and unit smoothing in soft Dice. The historical ImageNet-initialized ResNet-18 U-Net run configuration records 256-pixel square inputs, batch size 32, learning rate 0.0002, at most 24 epochs, validation patience 6 and an NVIDIA GeForce RTX 3090. Only the encoder is initialized from ImageNet; no COCO segmentation weights are used. Evaluation aligns the two labels globally per image; boundary F1 uses tolerance 0.0075 times the image diagonal, rounded to pixels with a minimum of one.

The archived first-seed run reaches 84.8 mIoU and 46.0 boundary F1 on the random holdout, $81.0/41.9$ on unseen pairs and $85.3/47.2$ on the curated test, on a 0–100 scale. Across three seeds, mIoU mean and sample standard deviation are 85.1 ± 1.4, 80.9 ± 0.6 and 86.8 ± 1.7, respectively. The first-seed paired mIoU gain over Lab+XY two-means is 11.1 points (saved 95% bootstrap CI 8.4–13.9) on the random holdout and 9.1 (6.2–12.0) on the curated test. Existing per-image scalar records reproduce these means, counts and seed statistics; an independent scalar check also reproduces

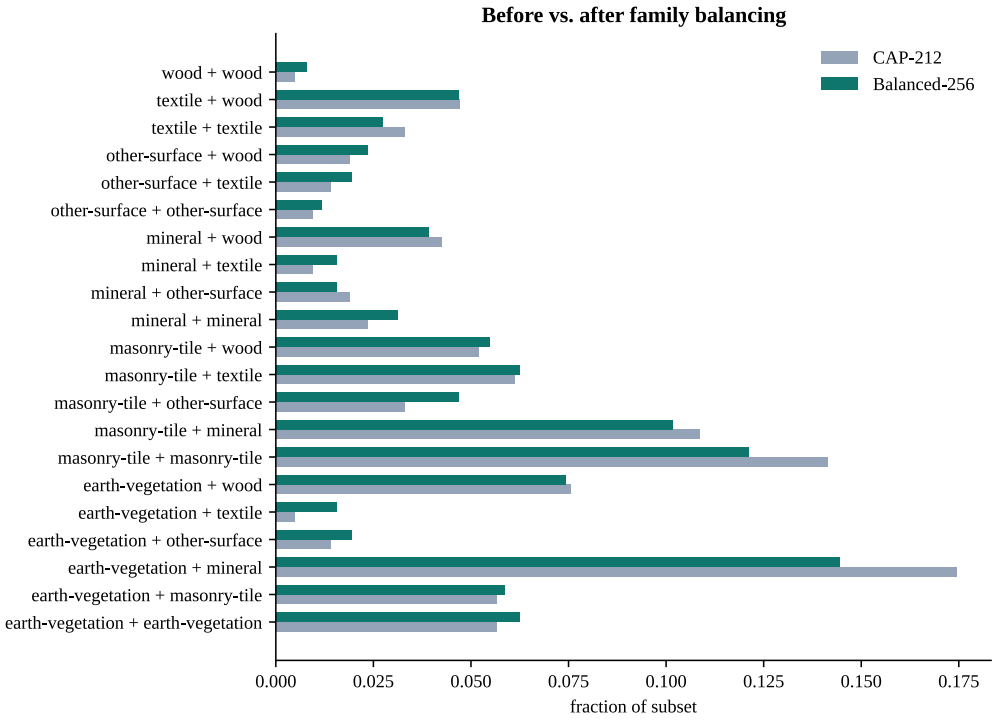


Figure 10: **Before and after balancing.** Balanced-256 preserves every curated anchor while limiting concentration and filling underrepresented family-pair cells with 44 matched extensions. Bars are each family pair's share of its subset, so the comparison is independent of set size.

the saved confidence intervals using the archived seed and resampling recipe, without model replay. The historical checkpoints are locally available and hashed, but are not included in the compact scalar/source annex.

The gain over this color/position baseline supports learned RGB partitioning on the constructed labels, while the boundary scores leave substantial room for improvement. This study supplies the two-region target structure and trains a supervised model; its metrics do not measure autonomous cardinality discovery and should not be pooled with the main Detecture comparison. Its three-seed result also does not establish training-seed stability for Detecture. Current source copies document the recovered implementation but do not, by themselves, certify the exact code bytes at every historical training execution.

F.5 Governance, Limitations and Intended Use

The construction manifests preserve source dataset, split, image ID and URL, crop coordinates, source label IDs and names, score components, tier, hashes and paths to the original-label crop. A read-only correspondence check binds all 256 Balanced-view images and both masks per image byte-for-byte to the pinned RWTD-COCO evaluation release; remapping each local source-label crop also reproduces its stored trinary target. Intended uses are appearance partitioning, class-agnostic two-region segmentation, boundary evaluation and data-centric selection research. Out-of-scope uses include identity recognition, surveillance and claims of exhaustive semantic-class coverage; thing area is explicitly suppressed, but incidental people may remain, and redistribution must preserve original provenance and per-image licensing (Caesar et al., 2018).

Three limitations bound what the release establishes. First, COCO-Stuff annotators were asked for semantic stuff, not texture homogeneity: a single class may contain multiple appearances and some semantic boundaries are weak in texture, so pair frequency reflects COCO photography and our whitelist rather than a sample of the visual world. Second, the curated anchors reflect one

curation process rather than measured inter-reviewer agreement, and positive-unlabeled ranking can reproduce that process's blind spots. The 44 admitted additions remain proposals for independent adjudication despite satisfying explicit matching envelopes; the stored review template is not a completed consensus study. The released RWTD-COCO region-description fields are empty and cannot validate generated phrase semantics. Third, comparison to the original Real-world Texture Dataset is operational rather than distributional, as no shared annotation protocol links them. Multi-annotator agreement, adjudication and cross-dataset transfer remain natural next steps. RWTD-COCO is therefore an independently sourced, annotation-reuse route with human-derived target pixels, not a claim of measured human consensus about appearance boundaries.

G CSTD: Controlled Synthetic Texture Compositions

CSTD is the third contributed benchmark, complementing the natural TextureADE and RWTD-COCO datasets. Its synthetic pipeline is separate from Detecture Miner and COCO-Stuff crop construction. The published dataset card describes Stable Diffusion 1.5 with ControlNet conditioning over source textures from the Describable Textures Dataset (DTD; Cimpoi et al., 2014). The two-region stitching masks supplied to generation define the targets.

Texture-feature screening and visual selection produce the released 256-image evaluation subset. The release supplies generated images, two binary masks per image, subset metadata, accepted identifiers, screening scores/ranks and the screening script. The frozen evaluation uses this entire selected view. Its region-description fields, like RWTD-COCO's, are empty; intervention evidence therefore comes from the separately specified model-text studies.

The source-card construction attribution and saved screening and subset files establish the reported provenance. The full historical synthesis and visual-selection process has not been independently replayed or adjudicated. Generated appearance can depart from its input stitching mask, so target agreement measures recovery of the imposed composition and is not human perceptual boundary validation. The selected population also differs from the unfiltered synthetic corpus. The supplement preserves the source-card and annotation-audit bindings. No new benchmark images or annotations were generated in this revision.

H Description-Mask Association Failure

Figure 1B shows a TextureADE example whose released record contains an association error: the soil and spillway descriptions are linked to each other's reference regions. The lower-right soil region carries the spillway description, and the diagonal strip beside the spillway carries the soil description. The figure displays the corrected pairing; the released annotation itself is unchanged. The masks can still agree geometrically with the inherited annotation because reference recovery and phrase association ask different questions. Independent judgments of the intended phrase-region pairs would be needed to measure phrase correctness across the dataset. The clause interventions in Appendix L instead test changes in mask identity relative to each image's unedited assignment.

I Autonomous Region-Count Accuracy

Because Detecture chooses its descriptions under a fixed global instruction, a natural diagnostic is how closely its nonempty-region count matches the reference count. We measure this on all 212 TextureADE images from the frozen autonomous run used in the revised evaluation. Let K_{GT} be the number of positive reference regions and $\hat{K}_{\mathrm{region}}$ the number of positive output labels occupying at least one annotation-valid pixel. The learned residual and reference void are excluded. The reference counts range from 2 to 5.

Markers, slots and regions. Three counts must be kept separate: visible grounding-marker strings in the saved caption, accepted decoder slots and nonempty regions after pixel competition. The decoder accepts at most 6 slots. In 3 TextureADE captions, the saved text contains 7 visible markers although only 6 slots are accepted. This is a count of decoded strings; the original generation token IDs were not retained. The mean visible-marker count is 3.64, while the mean accepted-slot count is 3.63. On this TextureADE run, every accepted slot has a nonempty region on the valid domain, so slot and

nonempty-region counts agree; this equality is an observed result, not a general guarantee of the architecture.

Aggregate accuracy. The nonempty-region count is exact on 99 of 212 images (46.7%) and lies within one reference region on 90.6%. Mean absolute count error is 0.63 regions and root mean squared error is 0.90. The mean predicted count is 3.63 against a mean reference count of 3.91, giving a signed bias of -0.28 regions. The model under-counts on 39.2% of images and over-counts on 14.2%. These quantities measure count agreement with the supplied annotations; they do not identify the perceptual cause of a missed region or establish a uniquely correct granularity.

Behavior by count. Figure 11 gives the complete joint counts. Exact agreement is 68.1% for $K_{\mathrm{GT}} = 3$, 39.8% for $K_{\mathrm{GT}} = 4$ and 30.2% for $K_{\mathrm{GT}} = 5$. The distribution shows an overall tendency to under-count, with both over-counted and under-counted cases retained. The figure and summaries use the same nonempty-region definition throughout.

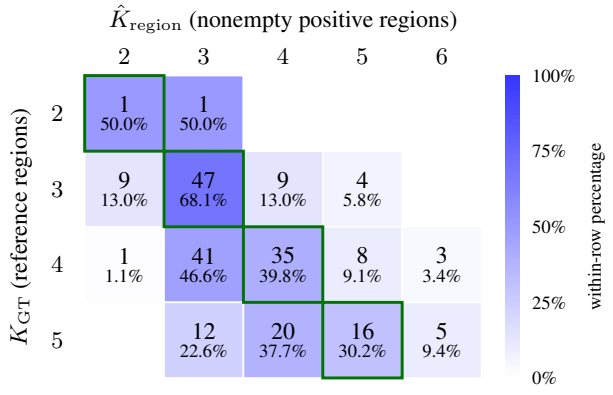


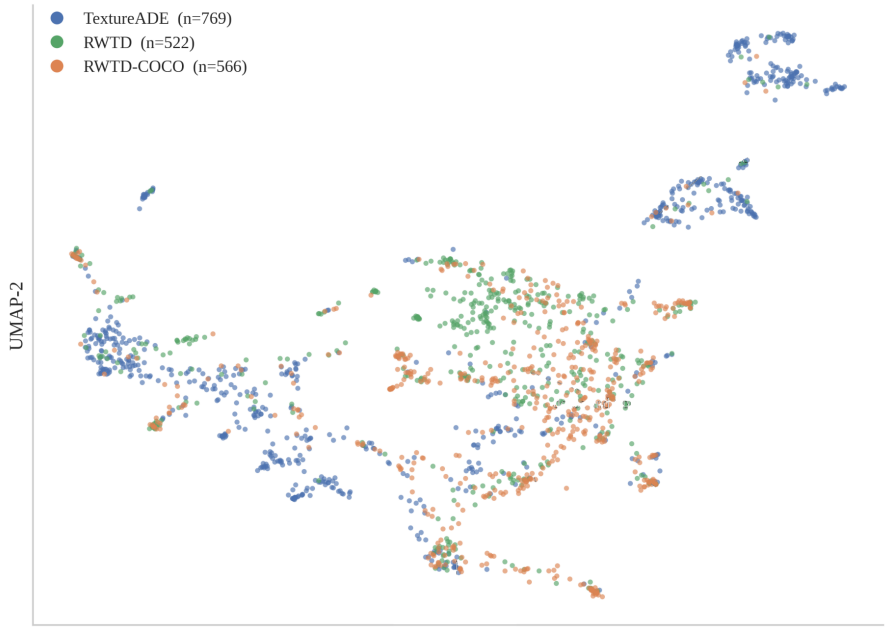
Figure 11: **Autonomous region-count confusion matrix on TextureADE.** Rows are positive reference counts K_{GT}; columns are nonempty positive output counts $\hat{K}_{\mathrm{region}}$ on annotation-valid pixels. All 212 images from the frozen autonomous run are included. Cells show image counts and within-row percentages; shading is row-normalized and green outlines mark exact-count cells. The learned residual is excluded; blank cells are zero. Counts come from the same frozen run as the autonomous comparison.

J Linguistic Richness of Descriptions

Population-level view of Detecture's appearance descriptions. A natural concern about any VLM-driven segmentation system is that the language head collapses onto a small set of rigid templates and the "descriptions" carry little real textural content. To examine this for Detecture, we take every `TEXTURE_i` description it emitted on RWTD, RWTD-COCO and TextureADE in the archived shared-instruction runs, whose instruction asks for "multiple" regions and supplies no count (Table 36): 522, 566 and 769 descriptions respectively, 1,857 in total, in the canonical "Texture of ⟨body⟩" form. These archived outputs are separate from the evaluation in Section 3.2.

A direct count answers the template question first: 1,835 of the 1,857 descriptions (98.8%) are distinct strings, so the language head does not repeat a few fixed phrasings. To see how these descriptions spread, we embed each one with the sentence-transformer `all-MiniLM-L6-v2` and project the 384-dimensional embeddings to two dimensions with UMAP under a cosine metric ($n_{\mathrm{neighbors}}$=25, min_dist=0.15). Figure 12 shows the resulting 2D landscape, colored by source dataset. Three properties stand out. First, the descriptions spread widely across the projection rather than collapsing into a handful of point-like clusters. Second, the datasets overlap substantially: both two-region routes, RWTD and RWTD-COCO, are largely interleaved, indicating that Detecture uses a *shared appearance vocabulary* rather than dataset-specific phrasing, while TextureADE's multi-region scenes also form a few groups of their own. Third, six representative samples from across the projection are listed under the plot: every one describes material or structure (*fine granular surface*, *rough concrete surface*, *dense and fibrous with…*) rather than naming an object, illustrating the breadth of the texture

vocabulary Detecture produces. The projection is a qualitative view; proximity in it does not measure whether a phrase correctly describes its mask.



Original annotation phrases (callouts omitted to prevent overlap):

smooth reflective
vibrant green uniform
rough concrete surface
dense and fibrous with
fine granular surface
dark and rough with

Figure 12: **Semantic landscape of Detecture's appearance descriptions.** 2D UMAP projection of `all-MiniLM-L6-v2` sentence embeddings of every `TEXTURE_i` description Detecture emitted on RWTD, RWTD-COCO and TextureADE in the archived shared-instruction runs (n=1,857, canonical "Texture of ⟨body⟩" form; cosine metric, $n_{\text{neighbors}}$=25, min_dist=0.15). Points are colored by source dataset. The descriptions spread widely with substantial cross-dataset overlap. Six representative descriptions, listed below the plot, illustrate the breadth of the texture vocabulary.

K Evaluation Metrics

This appendix collects the definitions used in Section 3. Let $\mathcal{P} = \{P_1, \ldots, P_K\}$ be nonempty positive predicted regions and $\mathcal{G} = \{G_1, \ldots, G_M\}$ the reference regions, all restricted to reference-valid pixels. Define $N = \sum_j |G_j|$ and $\mathrm{IoU}(A, B) = |A \cap B|/|A \cup B|$. Prediction label zero is the undescribed residual; reference label zero is annotation void and is excluded. Scores are averaged over images, including failed or empty outputs.

Shared scoring path. Each method's native output is converted into a disjoint label map using its declared conflict rule, then restored to the reference geometry. Primary scores use the raw label map. No residual filling, reference-count truncation or reference-based proposal selection enters the autonomous comparison. An output with no positive region receives zero covering, IoU, ARI and boundary F; its count and residual fraction remain measured.

Hungarian-matched IoU. For one-to-one pairings Π of size $\min(K, M)$,

$$\mathrm{IoU}_{\text{penalized}}(\mathcal{P}, \mathcal{G}) = \frac{1}{\max(K, M)} \max_{\pi \in \Pi} \sum_{(i,j) \in \pi} \mathrm{IoU}(P_i, G_j). \tag{10}$$

Unmatched regions contribute zero. The original paper's matched-only mIoU divides by $\min(K, M)$ instead and can reward one accurate mask despite missing regions. Historical reproduction tables

retain that convention explicitly; it is not the primary IoU reported here. Both matchings are invariant to a relabeling of predictions.

Region covering. The reference-to-prediction term follows region covering (Arbeláez et al., 2011); the primary metric symmetrizes it with a prediction-to-reference term:

$$C_G = \frac{1}{N}\sum_{j=1}^{M} |G_j| \max_i \mathrm{IoU}(G_j, P_i), \qquad C_P = \frac{1}{N}\sum_{i=1}^{K} |P_i| \max_j \mathrm{IoU}(P_i, G_j),$$
$$C_{\mathrm{bi}} = \frac{C_G + C_P}{2}. \tag{11}$$

Both directional terms use the same reference-valid area N. Predicted residual pixels therefore reduce coverage rather than disappearing through normalization. The original one-way statistic is C_G. The regenerated historical plots use C_{bi}; the original one-way numbers remain in the archived reconciliation record.

Boundary F. Boundary pixels lie on either side of a horizontal or vertical label transition whose two pixels are annotation-valid; the image frame is excluded. Precision and recall use Euclidean distance-transform proximity within 0.75% of the native image diagonal. Boundary F is $2PR/(P+R)$, with zero when only one map has boundaries and one when both are boundary-free. The empty-positive-output override above still assigns zero. This proximity rule differs from the bipartite boundary assignment of Martin et al. (2004). Primary scoring uses the raw residual-containing map; the original nearest-region-filled variant is historical only.

ARI. The adjusted Rand index compares pixel-pair co-membership, including the residual cluster on valid pixels. With contingency counts n_{ij}, row sums a_i, column sums b_j and N pixels,

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}. \tag{12}$$

It is chance-corrected, can be negative and is insensitive to label names. The implementation handles degenerate partitions and uses floating-point-safe arithmetic; metric tests cover empty, split, merged, extra and permuted regions.

Reference-slot IoU. Let π_{ref} be the Hungarian pairing obtained once for the unedited descriptions, and let K' count nonempty positive regions after an intervention. The fixed-assignment score is

$$\mathrm{IoU}_{\mathrm{fixed}}(\mathcal{P}, \mathcal{G}) = \frac{1}{\max(K', M)} \sum_{(i,j)\in\pi_{\mathrm{ref}}} \mathrm{IoU}(P_i, G_j). \tag{13}$$

Here P_i retains decoder-slot identifier i after editing; missing or empty slots are empty sets and are never renumbered. Missing referenced slots contribute zero. The score measures assignment stability with a cardinality penalty. Both generated-prompt arms and the separate supplied-reference track freeze the unedited Hungarian pairing. Supplying reference wording does not supply the pairing used by this regenerated diagnostic. Phrase correctness requires a separate annotation of the intended phrase-to-region association.

Follow, stay, other. A known swap moves content from slot $\sigma(i)$ into slot i. The expected association is $\pi_{\mathrm{ref}}(\sigma(i))$. A complete image-level *follow* event requires every slot's new association to equal this expectation; *stay* retains the original association; other complete associations form the third category. Eligibility requires equal reference, accepted-slot and nonempty-region counts in the reference and edited conditions. There is no IoU-accuracy threshold. Appearance-versus-spatial comparisons use the intersection of eligible images. This conditional event measure can be reliable about identity changes without establishing accurate masks or correct descriptions.

Uncertainty and comparisons. Paired image bootstrap intervals and sign-flip tests use 10,000 draws. Intervals are marginal. Holm correction covers 40 primary method-by-route contrasts together, 9 mechanism contrasts separately within each route and 13 historical C0-versus-condition contrasts separately within each dataset/arm. Image and reviewed-scene tests use separate corrections. The exact two-sided McNemar test for appearance-versus-spatial follow events uses discordant pairs and corrects the three dataset/arm contrasts together. Its null differs from the separate uniform-permutation $1/K!$ reference used in historical chance analyses. Sign-flip inference assumes exchangeability or an

appropriate sign-symmetric difference distribution; the observational method comparisons do not randomize model assignment. A single released checkpoint supplies no training-seed uncertainty.

Reviewed-scene cluster resampling and deterministic deduplication assess identified source dependence. They do not certify training independence or discover every duplicate. The analyses are exploratory, with dated protocol records rather than an external preregistration. Metric validation and the complete original testing families are included in the compact supplement.

L Language-Intervention Study: Protocol and Full Results

Section 4 asks whether description content influences region assignment. This appendix preserves the original clause-intervention study while distinguishing its historical prompt settings from the autonomous frozen diagnostics of Appendix D. All reported replays use the existing weights without further learning.

Protocol. Historical variants A and B both use archived model-generated descriptions. Variant A requests name, appearance and spatial fields, then removes the name; RWTD variant B generates under a prompt that omits the name request. The regenerated reference-slot diagnostic fixes the unedited Hungarian association in both arms. The RWTD prompt explicitly requests two regions, while TextureADE uses the historical range instruction. Text is supplied by teacher forcing during mask-state recomputation. The unedited C0 sequence is therefore the reference for its own intervention family, not the autonomous score in Table 2. The released checkpoint reproduces the archived mask pathway closely at the aggregate level, with remaining per-image numerical differences recorded in the supplement.

The replay covers the archived populations: 252 RWTD variant-A cases and 233 variant-B cases, with 231 available for variant-B C3, versus 253 images in the autonomous release. The construction code excludes generations with mismatched marker counts, fewer than two descriptions or undecomposable fields. Variant-B C3 additionally requires a matching set of name fields from the original three-field generation. Per-image exclusion manifests were not retained, so these rules do not certify every historical omission. These availability counts precede the equal-count assignment eligibility test; Tables 16–18 retain each condition's own denominator.

The separate supplied-reference track reads the stored RWTD description pairs: O0 keeps their order, O1 exchanges whole descriptions and O2 substitutes nonsense. Its full-text swap includes spatial wording and therefore is not an appearance-only intervention. The regenerated fixed-assignment metric uses the unedited matching here as well. These reference-description records do not independently establish human authorship or phrase correctness.

Concrete appearance swap. The smallest archived RWTD image ID (1) gives the following pair. The same image and spatial clauses are retained; the appearance clauses exchange slots.

> **C0.**
> TEXTURE_1: Texture of intricate brown and white mosaic pattern, resting on sandy seabed <|seg|>
> TEXTURE_2: Texture of fine granular beige surface with scattered small particles, surrounding shell <|seg|>
>
> **C5.**
> TEXTURE_1: Texture of fine granular beige surface with scattered small particles, resting on sandy seabed <|seg|>
> TEXTURE_2: Texture of intricate brown and white mosaic pattern, surrounding shell <|seg|>

The retained paraphrases were produced by the historical annotator and may overlap the reference wording. Nonsense is word-count matched, not necessarily matched in tokenizer length. Blocking direct marker-to-image attention leaves indirect image influence through preceding description tokens; the broader assistant block still retains other available context and the separate segmentation image path. These distinctions determine what each control tests.

Partition quality versus identity. The regenerated plots use primary bidirectional covering for geometric recovery and reference-slot IoU for assignment stability. An appearance swap can preserve

Condition	Change relative to the same historical image and reference description
C0 / C0v	Templated generated text / verbatim generated text.
C1 / C2	Retain only appearance / spatial clauses.
C3	Retain only the leading category/name clause.
C4a–c	Three archived paraphrases intended to retain appearance and location.
C5 / C6	Cyclically exchange appearance / spatial clauses among slots.
C7	Replace content with archived word-count-matched nonsense.
C8a / C8b	Block direct image-token attention from grounding markers / the full assistant region.
C9	Route text through the native SAM 3 semantic-head pipeline.

Table 12: The original language interventions. Clause edits leave the image and model weights fixed. C9 changes a pipeline rather than isolating the Bridge alone. The two generation prompts and their count assistance are retained explicitly.

much of the geometric division while moving the slot that owns each region. This is the signature tested by the original two-panel figures, rather than an assumption that every swap succeeds.

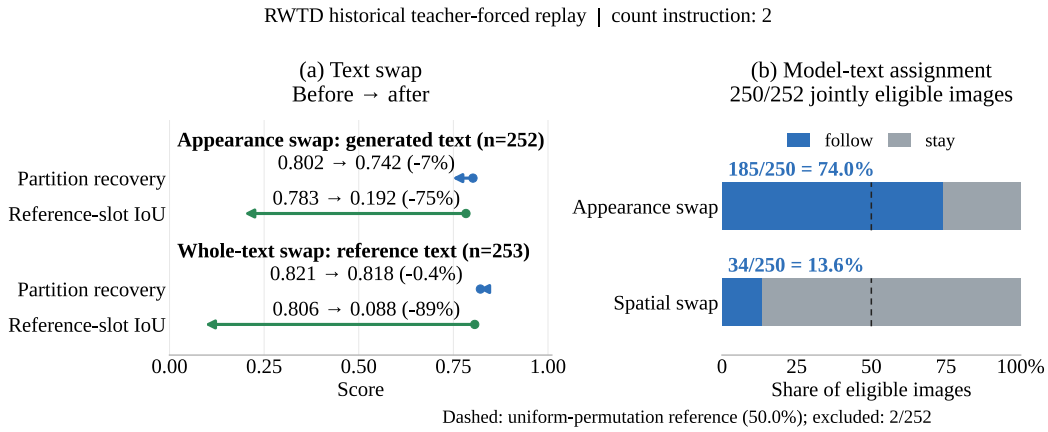


Figure 13: **Text swaps and region identities on RWTD.** Left: the generated-text appearance swap C5 (variant A) and the separate supplied-reference whole-description swap O1, each relative to its own unchanged text. O1 uses 253 images and swaps spatial wording along with appearance; it is not variant B. Right: appearance- versus spatial-follow events in jointly eligible variant-A images. All replays use the two-region instruction.

Dataset	Historical arm	Joint n	Appearance	Spatial	Difference [95% CI]	Holm p
RWTD	variant A	250	185	34	60.4 [53.6, 66.8]	<0.001
RWTD	variant B	233	159	34	53.6 [45.9, 60.9]	<0.001
TextureADE	variant A	98	58	1	58.2 [48.0, 67.3]	<0.001

Table 13: Jointly eligible historical identity contrasts. Appearance and spatial columns are event counts, not percentages. Differences are percentage points with marginal paired intervals. The original exact paired tests use Holm correction across these three contrasts.

Eligibility is restrictive on TextureADE: 98 of 212 images have complete assignments across the relevant conditions. This denominator defines the identity result. It does not justify extrapolating its event rate to images with incomplete or unequal counts. Hungarian ties are assessed by a separate unique-optimum sensitivity; no mask-accuracy cutoff selects these images.

Full corrected effects. Tables 16–18 retain every historical condition's covering effect and marginal interval with both image and scene corrections. Their correction family differs from the three identity contrasts above. Some archived corrected identifiers recur in other summaries; the claim ledger points to the causal-replay family explicitly. These are exploratory historical controls, not preregistered

Dataset	Variant	Appearance only	Spatial only	Exact Holm p	Cluster Holm p
RWTD	A	157	6	<0.001	<0.001
RWTD	B	136	11	<0.001	<0.001
TextureADE	A	57	0	<0.001	<0.001

Table 14: Discordant follow events and exact McNemar versus reviewed-scene sign-flip sensitivity, each with its original three-contrast correction. “Appearance only” and “spatial only” here mean following just one edit, not the C1/C2 clause-isolation conditions.

evidence of conceptual understanding. The functional conclusion is that structured wording affects the released model’s masks and eligible identities.

Historical arm	Joint n	Unique n	Follow difference [95% CI]
RWTD / variant A	250	250	60.4 [53.6, 66.8]
RWTD / variant B	233	233	53.6 [45.9, 60.9]
TextureADE / variant A	98	98	58.2 [48.0, 67.3]

Table 15: Identity sensitivity restricted to unique optimal assignments. Joint and unique denominators are shown separately; effects are appearance-minus-spatial differences in follow probability (percentage points).

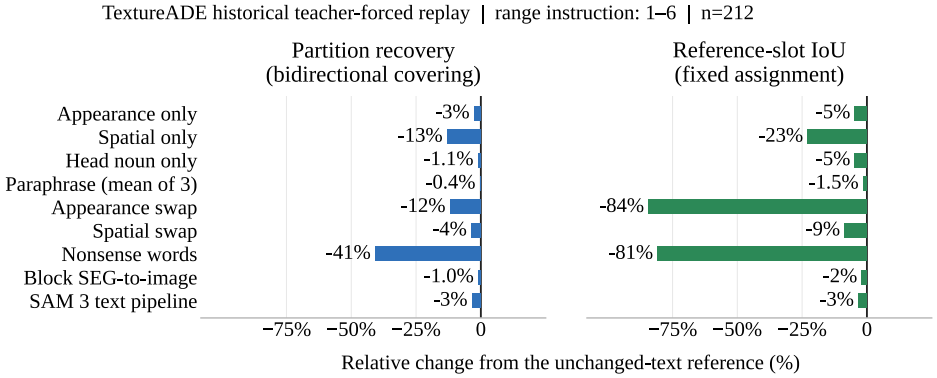


Figure 14: **Description interventions on TextureADE.** Historical fixed-text, range-prompt replay. Percent changes use the same unedited reference for bidirectional covering and reference-slot IoU. Dataset-dependent effects of partial clauses and native text routing are retained.

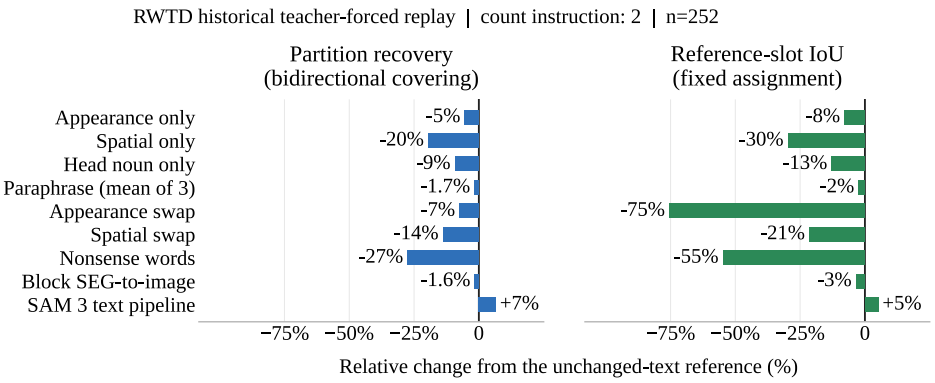


Figure 15: **Description interventions on RWTD.** The original generated-text, two-region historical replay compares covering and reference-slot IoU with unchanged text. Relative percent changes are distinct from the percentage-point effects in the full statistical tables. The SAM 3 condition changes the text-to-mask pipeline.

Intervention	n	Difference [95% CI]	Image p	Scene p
C0v: Verbatim generated text	252	2.6 [1.4, 3.8]	0.001	0.001
C1: Appearance only	252	−4.4 [−6.0, −2.7]	0.001	0.001
C2: Spatial only	252	−15.6 [−17.9, −13.5]	0.001	0.001
C3: Head noun only	252	−7.4 [−9.8, −5.1]	0.001	0.001
C4a: Archived paraphrase 1	252	−1.0 [−1.9, 0.0]	0.055	0.069
C4b: Archived paraphrase 2	252	−1.4 [−2.5, −0.4]	0.010	0.014
C4c: Archived paraphrase 3	252	−1.5 [−2.5, −0.7]	0.003	0.004
C5: Appearance swap	252	−6.0 [−8.9, −3.0]	0.001	0.001
C6: Spatial swap	252	−11.1 [−13.2, −9.0]	0.001	0.001
C7: Word-count nonsense	252	−22.0 [−24.6, −19.4]	0.001	0.001
C8a: Block SEG-to-image attention	252	−1.3 [−2.4, −0.1]	0.055	0.069
C8b: Block assistant-to-image attention	252	−9.7 [−11.7, −7.8]	0.001	0.001
C9: SAM 3 semantic-head pipeline	252	5.3 [2.6, 7.9]	0.001	0.001

Table 16: RWTD variant A (model-generated text): primary covering changes in percentage points, paired intervals and original image/scene corrected tests.

Intervention	n	Difference [95% CI]	Image p	Scene p
C0v: Verbatim generated text	233	0.0 [0.0, 0.0]	1.000	1.000
C1: Appearance only	233	−6.4 [−8.3, −4.6]	0.001	0.001
C2: Spatial only	233	−10.0 [−12.4, −7.7]	0.001	0.001
C3: Head noun only	231	−5.3 [−7.8, −2.9]	0.001	0.001
C4a: Archived paraphrase 1	233	−0.8 [−1.5, −0.2]	0.076	0.060
C4b: Archived paraphrase 2	233	−1.0 [−1.9, 0.0]	0.195	0.203
C4c: Archived paraphrase 3	233	−0.2 [−0.9, 0.6]	1.000	1.000
C5: Appearance swap	233	−6.8 [−9.6, −3.9]	0.001	0.001
C6: Spatial swap	233	−10.5 [−12.7, −8.3]	0.001	0.001
C7: Word-count nonsense	233	−22.5 [−25.2, −19.8]	0.001	0.001
C8a: Block SEG-to-image attention	233	0.4 [−0.7, 1.5]	1.000	1.000
C8b: Block assistant-to-image attention	233	−6.2 [−7.8, −4.6]	0.001	0.001
C9: SAM 3 semantic-head pipeline	233	7.0 [4.5, 9.5]	0.001	0.001

Table 17: RWTD variant B: model-generated text from the noun-free prompt. C3 borrows matching name fields from the three-field generation. Both generated-prompt arms use the unedited Hungarian association for the identity diagnostic.

Intervention	n	Difference [95% CI]	Image p	Scene p
C0v: Verbatim generated text	212	1.4 [0.6, 2.2]	0.005	0.005
C1: Appearance only	212	−2.0 [−2.9, −1.1]	0.001	0.001
C2: Spatial only	212	−9.9 [−11.6, −8.2]	0.001	0.001
C3: Head noun only	212	−0.9 [−1.9, 0.2]	0.495	0.495
C4a: Archived paraphrase 1	212	−0.3 [−0.8, 0.3]	0.939	0.939
C4b: Archived paraphrase 2	212	−0.3 [−0.9, 0.3]	0.939	0.939
C4c: Archived paraphrase 3	212	−0.3 [−0.9, 0.3]	0.939	0.939
C5: Appearance swap	212	−9.0 [−11.2, −7.0]	0.001	0.001
C6: Spatial swap	212	−2.8 [−4.0, −1.6]	0.001	0.001
C7: Word-count nonsense	212	−30.9 [−33.1, −28.6]	0.001	0.001
C8a: Block SEG-to-image attention	212	−0.8 [−1.4, −0.1]	0.084	0.084
C8b: Block assistant-to-image attention	212	−4.2 [−5.5, −3.0]	0.001	0.001
C9: SAM 3 semantic-head pipeline	212	−2.6 [−4.1, −1.1]	0.006	0.006

Table 18: TextureADE variant A (model-generated text) under its retained range prompt: primary covering changes in percentage points. Every condition is compared with its own historical reference.

M Evaluation Protocol and Full Results

This appendix expands the autonomous comparison in Section 3.2: what each method receives, how outputs are converted to labels, and how uncertainty and dependence affect the conclusions. All campaign evaluations use frozen pretrained checkpoints and retain every released image. The compact supplement contains the remaining diagnostic and reproduction workstreams without turning each into a separate main-paper claim.

M.1 Exact Task Instructions

System message.

> You analyze surface textures in images. Always respond in the exact format requested, with no extra text.

Fixed autonomous user instruction (verbatim).

> This image contains multiple main visually distinct regions separated by a boundary (for example, a prominent foreground object and its background, or contrasting materials).
>
> Write a single, highly descriptive phrase (approximately 10-15 words) for each region. Include the following precise information:
> 1. Semantic Name: A natural, common name for the material or object.
> 2. Distinct Visual Features: The core visual attributes like color, pattern, or texture that strongly contrast with other regions.
> 3. Spatial Context: A brief note on its general position (e.g., 'foreground', 'background', 'top-left').
>
> IMPORTANT: Describe the ENTIRE region as a collective group, NOT individual objects within it. Think of each region as a surface/area, not as a single object.
>
> Format your response exactly like this for the number of regions present:
> `TEXTURE_1: Texture of <description> <|seg|>`
> `TEXTURE_2: Texture of <description> <|seg|>`
> ...

The literal token `<|seg|>` is denoted `[SEG]` elsewhere. Variant-A RWTD replays substitute `2` for `multiple`; TextureADE replays substitute `between 1 and 6`. These substitutions retain the rest of the instruction.

RWTD variant-B generation instruction (verbatim).

> This image contains 2 main visually distinct regions separated by a boundary (for example, a prominent foreground object and its background, or contrasting materials).
>
> Write a single, highly descriptive phrase (approximately 10-15 words) for each region. Include the following precise information:
> 1. Distinct Visual Features: The core visual attributes like color, pattern, or texture that strongly contrast with other regions.
> 2. Spatial Context: A brief note on its general position (e.g., 'foreground', 'background', 'top-left').
>
> IMPORTANT: Do NOT name the object or the material. Describe only how the region looks. Describe the ENTIRE region as a collective group, NOT individual objects within it. Think of each region as a surface/area, not as a single object.
>
> Format your response exactly like this for the number of regions present:
> `TEXTURE_1: Texture of <description> <|seg|>`
> `TEXTURE_2: Texture of <description> <|seg|>`
> ...

GLaMM native instruction.

> Could you please give me a detailed description of the image? Please respond with interleaved segmentation masks for the corresponding parts of the answer.

Sa2VA native instruction.

> Could you please give me a brief description of the image? Please respond with interleaved segmentation masks for the corresponding parts of the answer.

SAMTok native instruction.

> Could you please give me a brief description of the image? Please respond with interleaved segmentation masks for the corresponding parts of the answer and highlight the phrases. A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>

The Sa2VA image marker is followed by a newline before the question. The older adapter supplies that marker; the current Qwen implementation removes the marker before chat processing while retaining the newline. Exact whitespace and tokenizer identifiers are retained in the run manifests. The Qwen/SAM 3 cascades use the Detecture instruction with its grounding-marker literal removed. The shared-format controls reuse that template with each model's supported marker spelling.

Phrase source	Dataset	Phrases	Truncated	Max. tokens	Collision images
Supplied annotations	RWTD	506	0	10	0
Supplied annotations	TextureADE	828	0	26	0
C9, variant A	RWTD	507	1	38	2
C9, variant A	TextureADE	775	0	26	1
C9, variant B	RWTD	466	5	47	0
Base Qwen generation	CSTD	511	0	25	0
Base Qwen generation	RWTD	498	0	30	0
Base Qwen generation	RWTD-COCO	547	0	28	0
Base Qwen generation	TextureADE	722	0	25	0

Table 19: SAM 3 text-encoder input diagnostics. Autonomous, historical C9 and supplied-description inputs are separate rows. Truncated phrases exceed the encoder capacity; collision images contain two phrases mapped to the same retained tokens. These are interface-capacity checks, not semantic-faithfulness scores.

M.2 Inference Interfaces

Detecture. Both image paths receive the published square-warped input. Greedy generation uses seed 42 and a 300-token cap. Assistant-region markers are extracted, grounding states are recomputed with isolation, and each query produces mask logits through the Bridge, frozen SAM features and learned mask head. If parsed slot count equals accepted slot count and at least one slot parses, leading name clauses are removed, unparsed slots are retained, and grounding is recomputed again. Otherwise, the original sequence is used. Unused channels are suppressed before argmax; nearest-neighbor label resizing restores native reference geometry. Descriptions keep their original slot order.

The fixed instruction asks for multiple appearance regions. It supplies no per-image reference count. The release's six-slot cap is a model capacity constraint, not a supplied target count. Count, residual and empty-region diagnostics expose its consequences rather than silently filling or dropping outputs.

The parser accepts anchored `TEXTURE` lines containing a grounding marker and removes the literal prefix and grounding marker before splitting on commas. It discards empty pieces; the first remaining piece is the name, the last is spatial context, and the interior pieces form appearance. At least three pieces are required. If parsed-line count equals accepted-slot count and any line decomposes, parsed name fields are removed while other lines retain their wording. Otherwise, the original caption supplies grounding. The raw generated caption is saved separately from this grounding text; figures state any display shortening. Removing a name field does not remove category information from the remaining words or image-conditioned states.

Visual and proposal controls. DINOv2 grouping applies Ward clustering to normalized patch features, selecting cluster count with a declared silhouette rule and a single-region fallback. TokenCut uses its pretrained feature and normalized-cut saliency pipeline; the evaluated binary partition adds the saliency complement and bilateral refinement, explicitly imposing a fixed binary prior. Mask2Former uses its COCO panoptic checkpoint and official processor. TextureSAM uses an automatic point grid with declared proposal thresholds and deterministic conflict resolution. None of these interfaces emits descriptions; they are strong tests of reference-region recovery, not complete text-and-mask systems.

Grounded generation. GLaMM runs its native grounded-conversation generation (GCG) route. The older Sa2VA-4B and current Qwen3-VL/SAM 3 variants use their respective released native interfaces. SAMTok uses its grounded-generation checkpoint and decoded mask-token pathway. Native output formats are retained instead of forcing a common artificial syntax. Model and tokenizer revisions, decoding settings and repairs are recorded in the manifests; only completed, validated runs enter the tables.

Base-Qwen/SAM 3 cascades. A pretrained base Qwen3-VL-8B-Instruct model generates descriptions under a fixed instruction, and SAM 3 then grounds them. The semantic-head interface and official detector interface are separate rows because they expose different mask paths. Both interfaces use autonomous descriptions; neither uses reference phrases. Timing of the detector-only route excludes cached text generation and is labeled accordingly in the resource records. The semantic-head input audit found no truncated phrases and no identical encoded phrase pairs within an image on any autonomous route (Table 19); these interface checks do not measure phrase correctness. Supplied reference-text and point-oracle comparisons are separate diagnostics.

Protocol quantity	Value
Detecture new-token cap	300
LISA new-token cap	512
GSVA extension new-token cap	512
GLaMM new-token cap	512
Sa2VA-4B native new-token cap	2,048
Current Sa2VA native new-token cap	2,048
SAMTok native new-token cap	2,048
SAM 3 native detection confidence	0.50
SAM 3 text context, including boundary tokens	32
TextureSAM points per grid side	64
TextureSAM points per batch	64
TextureSAM predicted-IoU threshold	0.80
TextureSAM stability threshold	0.20
TextureSAM mask-logit threshold	0
TextureSAM minimum region area	0
Historical mask CE coefficient	1
Historical Dice coefficient	3
Historical weighted-LM coefficient	0.10
Fixed diagnostic images per route	32
Fixed diagnostic images in total	128
Boundary tolerance (% of image diagonal)	0.75
Bootstrap / sign-flip Monte Carlo draws	10,000

Table 20: Source-extracted protocol constants. Historical objective weights describe the inspected original configuration. Inference settings and diagnostic sample counts refer to the frozen campaign; they do not imply that historical runs used identical settings.

M.3 Complete Autonomous Comparison

Reading the comparison. Detecture's covering advantage over native GLaMM and current Sa2VA is supported on all routes. Older Sa2VA and SAMTok comparisons support an advantage on RWTD, RWTD-COCO and TextureADE; CSTD remains unresolved after correction (Table 24). Strong visual grouping, panoptic and semantic-head cascade outputs lead on some datasets. The native grounded models' higher mean boundary scores further distinguish recovery of whole regions from boundary precision. This pattern supports the combined appearance-description interface while identifying concrete segmentation improvements.

Interface	Conversion to a disjoint label map
Detecture	Bilinearly enlarge competing logits to the evaluation square; argmax includes the learned residual, with lower index winning exact ties. Resize labels to reference geometry by nearest neighbor.
GLaMM / Sa2VA	GLaMM thresholds native mask logits above zero; Sa2VA returns binary masks. For GLaMM and current Sa2VA, remove instruction-origin masks before empty-mask filtering. Older Sa2VA selects generated-token states directly. Later generated masks win overlap; restore labels by nearest neighbor.
SAMTok	Bilinearly enlarge native mask logits to image size, then threshold above 0.5. Later generated masks win overlap.
TextureSAM automatic	Retain all proposals passing the declared automatic-generator thresholds. Overlap goes to highest predicted IoU, then stability score, then earlier proposal index.
Qwen + SAM 3 semantic	Bilinearly enlarge per-phrase logits; compete with a zero-logit residual, then restore labels by nearest neighbor. Lower index wins exact ties.
Qwen + SAM 3 detector	Use native binary instance masks above the detection confidence threshold. Retained instances form each phrase's union. Greatest detection score wins overlap; earlier phrase/instance wins ties.
Visual label maps	DINOv2 and TokenCut already define disjoint groups. Mask2Former uses its official panoptic output and retains unlabeled pixels.

Table 21: Native output conversion used by the reported autonomous runs. Zero positive masks remain an empty output; uncovered pixels retain label zero. The numeric automatic-generator and detection thresholds appear in Table 20. No rule uses reference masks to choose outputs.

Method	Text	RWTD	RWTD-COCO	TextureADE	CSTD
DINOv2 grouping	No	71.7	68.5	60.4	76.4
TokenCut + bilateral refinement	No	87.2	79.8	51.3	98.1
Mask2Former (COCO panoptic)	No	62.0	86.8	81.1	56.6
TextureSAM automatic	No	70.2	74.0	64.1	90.7
Qwen3-VL + SAM 3 semantic head	Yes	83.5	85.6	70.1	97.9
Qwen3-VL + SAM 3 detector	Yes	59.4	61.6	48.2	79.9
GLaMM (native GCG)	Yes	65.7	55.2	61.1	73.9
Sa2VA-4B (native GCG)	Yes	70.8	71.9	66.1	83.6
Sa2VA Qwen3-VL/SAM 3	Yes	70.4	73.6	71.8	79.7
SAMTok-3B (GCG checkpoint)	Yes	68.6	66.5	67.4	89.4
Detecture	Yes	79.4	85.2	75.7	87.7

Table 22: All eleven autonomous systems under the primary raw bidirectional-covering contract, on a percentage scale. "Text" denotes generated descriptions, not independently validated phrase faithfulness. RWTD-COCO is the balanced view called CAP in the release and artifact keys.

M.4 Secondary Metrics and Source Dependence

Tables 25–32 report the same complete predictions under penalized IoU, raw boundary F and ARI, with scene-grouped and deduplicated sensitivities. Duplicate candidates came from exact hashes, perceptual hashes, frozen visual features and geometric follow-up; scene groups were then visually inspected. They expose identified dependence without certifying complete independence. Source training/validation identifiers likewise do not establish absence from all pretraining or Detecture development.

Comparator	RWTD	RWTD-COCO	TextureADE	CSTD
DINOv2	7.7 [4.9, 10.6]	16.7 [14.4, 19.1]	15.3 [13.2, 17.5]	11.3 [8.5, 14.1]
TokenCut + refinement	−7.7 [−10.3, −5.2]	5.4 [2.8, 8.0]	24.4 [22.3, 26.5]	−10.4 [−12.4, −8.4]
Mask2Former	17.5 [13.5, 21.4]	−1.5 [−3.8, 0.7]	−5.3 [−6.9, −3.6]	31.0 [26.7, 35.4]
TextureSAM	9.2 [5.8, 12.5]	11.2 [8.8, 13.7]	11.6 [9.2, 14.2]	−3.1 [−5.5, −0.6]
Qwen + SAM 3 semantic	−4.0 [−6.5, −1.6]	−0.4 [−2.1, 1.3]	5.6 [3.6, 7.6]	−10.2 [−12.3, −8.2]
Qwen + SAM 3 detector	20.0 [16.0, 24.1]	23.6 [20.1, 27.2]	27.6 [24.3, 30.8]	7.7 [4.0, 11.6]
GLaMM	13.7 [10.1, 17.5]	30.1 [26.7, 33.4]	14.7 [11.4, 18.1]	13.8 [10.0, 17.5]
Sa2VA-4B	8.6 [4.9, 12.3]	13.3 [10.2, 16.4]	9.6 [6.7, 12.7]	4.0 [0.7, 7.4]
Sa2VA (Qwen3)	9.1 [5.2, 12.9]	11.6 [8.4, 15.0]	3.9 [1.4, 6.7]	8.0 [4.3, 11.7]
SAMTok	10.8 [7.4, 14.3]	18.8 [15.6, 21.9]	8.4 [6.2, 10.8]	−1.8 [−4.6, 1.1]

Table 23: Detecture-minus-comparator covering differences in percentage points, with marginal paired image-bootstrap 95% intervals. Signs and intervals retain strong alternatives and unresolved comparisons.

Comparator	RWTD	RWTD-COCO	TextureADE	CSTD
DINOv2	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004
TokenCut + refinement	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004
Mask2Former	0.004 / 0.004	0.545 / 0.557	0.004 / 0.004	0.004 / 0.004
TextureSAM	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004	0.091 / 0.091
Qwen + SAM 3 semantic	0.009 / 0.008	0.614 / 0.621	0.004 / 0.004	0.004 / 0.004
Qwen + SAM 3 detector	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004
GLaMM	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004
Sa2VA-4B	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004	0.091 / 0.091
Sa2VA (Qwen3)	0.004 / 0.004	0.004 / 0.004	0.017 / 0.017	0.004 / 0.004
SAMTok	0.004 / 0.004	0.004 / 0.004	0.004 / 0.004	0.545 / 0.557

Table 24: Original primary-family Holm-adjusted sign-flip tests; each cell gives image / reviewed-scene-cluster p-values. The complete method-by-route family is retained; a marginal interval does not replace multiplicity correction.

Method	GT covering	Pred. covering	Penalized IoU	ARI	Boundary F
DINOv2 grouping	73.2	70.2	62.6	53.3	17.7
TokenCut + bilateral refinement	86.5	87.8	82.8	71.6	58.6
Mask2Former (COCO panoptic)	60.5	63.5	46.3	31.0	24.9
TextureSAM automatic	73.2	67.2	56.7	71.4	44.1
Qwen3-VL + SAM 3 semantic head	84.9	82.0	81.4	82.2	69.9
Qwen3-VL + SAM 3 detector	61.2	57.6	58.9	71.4	57.8
GLaMM (native GCG)	67.3	64.1	58.4	67.8	55.0
Sa2VA-4B (native GCG)	72.2	69.5	65.2	67.6	54.6
Sa2VA Qwen3-VL/SAM 3	71.2	69.5	66.0	69.0	58.6
SAMTok-3B (GCG checkpoint)	70.6	66.6	67.3	74.2	57.9
Detecture	80.0	78.9	72.2	60.8	32.1

Table 25: RWTD: secondary metrics under the primary contract, on a percentage scale. IoU charges unmatched regions, and boundary F uses raw predictions.

Comparator	Image estimate [95% CI]	Scene-cluster CI	Deduplicated [95% CI]
DINOv2	7.7 [4.9, 10.6]	[4.8, 10.6]	7.6 [4.7, 10.6]
TokenCut + refinement	−7.7 [−10.3, −5.2]	[−10.3, −5.2]	−8.0 [−10.6, −5.4]
Mask2Former	17.5 [13.5, 21.4]	[13.6, 21.5]	17.2 [13.3, 21.3]
TextureSAM	9.2 [5.8, 12.5]	[5.9, 12.4]	9.6 [6.2, 12.9]
Qwen + SAM 3 semantic	−4.0 [−6.5, −1.6]	[−6.4, −1.6]	−3.9 [−6.3, −1.4]
Qwen + SAM 3 detector	20.0 [16.0, 24.1]	[15.9, 24.3]	19.9 [16.0, 23.9]
GLaMM	13.7 [10.1, 17.5]	[10.1, 17.4]	14.0 [10.3, 17.7]
Sa2VA-4B	8.6 [4.9, 12.3]	[4.9, 12.3]	8.5 [4.8, 12.3]
Sa2VA (Qwen3)	9.1 [5.2, 12.9]	[5.3, 12.9]	9.1 [5.3, 13.1]
SAMTok	10.8 [7.4, 14.3]	[7.4, 14.2]	10.6 [7.2, 14.1]

Table 26: RWTD: Detecture-minus-comparator covering differences, marginal image/scene intervals and deterministic deduplication sensitivity. Identified scene groups do not establish training independence.

Method	GT covering	Pred. covering	Penalized IoU	ARI	Boundary F
DINOv2 grouping	71.0	66.1	54.7	53.4	17.7
TokenCut + bilateral refinement	79.3	80.3	72.4	58.2	38.4
Mask2Former (COCO panoptic)	86.9	86.6	75.2	78.1	50.2
TextureSAM automatic	77.4	70.6	51.2	73.0	30.0
Qwen3-VL + SAM 3 semantic head	87.5	83.8	83.7	79.6	47.2
Qwen3-VL + SAM 3 detector	63.6	59.6	62.0	71.7	39.6
GLaMM (native GCG)	56.6	53.8	44.4	65.2	35.8
Sa2VA-4B (native GCG)	73.6	70.3	66.8	73.0	38.2
Sa2VA Qwen3-VL/SAM 3	74.6	72.5	68.1	76.1	45.8
SAMTok-3B (GCG checkpoint)	68.8	64.1	62.5	70.1	39.8
Detecture	86.2	84.3	80.8	74.3	27.6

Table 27: RWTD-COCO: secondary metrics under the primary contract, on a percentage scale. IoU charges unmatched regions, and boundary F uses raw predictions.

Comparator	Image estimate [95% CI]	Scene-cluster CI	Deduplicated [95% CI]
DINOv2	16.7 [14.4, 19.1]	[14.3, 19.1]	16.6 [14.2, 19.0]
TokenCut + refinement	5.4 [2.8, 8.0]	[2.8, 8.0]	5.5 [2.9, 8.1]
Mask2Former	−1.5 [−3.8, 0.7]	[−3.7, 0.8]	−1.5 [−3.7, 0.8]
TextureSAM	11.2 [8.8, 13.7]	[8.8, 13.7]	11.3 [8.9, 13.8]
Qwen + SAM 3 semantic	−0.4 [−2.1, 1.3]	[−2.1, 1.3]	−0.4 [−2.1, 1.3]
Qwen + SAM 3 detector	23.6 [20.1, 27.2]	[20.0, 27.3]	23.5 [19.9, 27.2]
GLaMM	30.1 [26.7, 33.4]	[26.6, 33.4]	29.9 [26.5, 33.3]
Sa2VA-4B	13.3 [10.2, 16.4]	[10.2, 16.3]	13.2 [10.2, 16.3]
Sa2VA (Qwen3)	11.6 [8.4, 15.0]	[8.5, 14.9]	11.6 [8.5, 15.0]
SAMTok	18.8 [15.6, 21.9]	[15.7, 21.9]	18.7 [15.6, 21.8]

Table 28: RWTD-COCO: Detecture-minus-comparator covering differences, marginal image/scene intervals and deterministic deduplication sensitivity. Identified scene groups do not establish training independence.

Method	GT covering	Pred. covering	Penalized IoU	ARI	Boundary F
DINOv2 grouping	62.0	58.9	46.5	49.7	19.3
TokenCut + bilateral refinement	48.0	54.7	29.7	26.6	30.7
Mask2Former (COCO panoptic)	80.9	81.2	69.2	77.0	59.8
TextureSAM automatic	67.5	60.7	49.5	69.7	53.6
Qwen3-VL + SAM 3 semantic head	71.6	68.7	59.9	71.2	57.0
Qwen3-VL + SAM 3 detector	50.5	45.9	42.3	57.5	47.4
GLaMM (native GCG)	62.4	59.7	50.7	65.1	53.4
Sa2VA-4B (native GCG)	67.9	64.3	58.9	69.2	54.5
Sa2VA Qwen3-VL/SAM 3	72.6	71.0	61.4	72.9	55.7
SAMTok-3B (GCG checkpoint)	70.1	64.6	60.6	69.9	57.1
Detecture	75.7	75.7	63.0	69.8	43.7

Table 29: TextureADE: secondary metrics under the primary contract, on a percentage scale. IoU charges unmatched regions, and boundary F uses raw predictions.

Comparator	Image estimate [95% CI]	Scene-cluster CI	Deduplicated [95% CI]
DINOv2	15.3 [13.2, 17.5]	[13.2, 17.5]	15.3 [13.2, 17.5]
TokenCut + refinement	24.4 [22.3, 26.5]	[22.3, 26.5]	24.4 [22.3, 26.5]
Mask2Former	−5.3 [−6.9, −3.6]	[−6.9, −3.6]	−5.3 [−6.9, −3.6]
TextureSAM	11.6 [9.2, 14.2]	[9.2, 14.2]	11.6 [9.2, 14.2]
Qwen + SAM 3 semantic	5.6 [3.6, 7.6]	[3.6, 7.6]	5.6 [3.6, 7.6]
Qwen + SAM 3 detector	27.6 [24.3, 30.8]	[24.3, 30.8]	27.6 [24.3, 30.8]
GLaMM	14.7 [11.4, 18.1]	[11.4, 18.1]	14.7 [11.4, 18.1]
Sa2VA-4B	9.6 [6.7, 12.7]	[6.7, 12.7]	9.6 [6.7, 12.7]
Sa2VA (Qwen3)	3.9 [1.4, 6.7]	[1.4, 6.7]	3.9 [1.4, 6.7]
SAMTok	8.4 [6.2, 10.8]	[6.2, 10.8]	8.4 [6.2, 10.8]

Table 30: TextureADE: Detecture-minus-comparator covering differences, marginal image/scene intervals and deterministic deduplication sensitivity. Identified scene groups do not establish training independence.

Method	GT covering	Pred. covering	Penalized IoU	ARI	Boundary F
DINOv2 grouping	78.0	74.8	67.6	62.4	14.4
TokenCut + bilateral refinement	98.0	98.1	97.5	95.8	94.9
Mask2Former (COCO panoptic)	54.4	58.8	43.0	29.6	28.1
TextureSAM automatic	92.0	89.4	85.9	91.8	60.7
Qwen3-VL + SAM 3 semantic head	98.2	97.6	98.1	97.1	88.6
Qwen3-VL + SAM 3 detector	80.5	79.4	80.5	91.9	74.9
GLaMM (native GCG)	74.4	73.4	66.5	78.9	62.7
Sa2VA-4B (native GCG)	83.8	83.4	78.3	86.4	70.2
Sa2VA Qwen3-VL/SAM 3	79.6	79.8	77.2	83.2	69.8
SAMTok-3B (GCG checkpoint)	90.0	88.9	88.0	90.8	79.1
Detecture	88.0	87.3	85.8	76.7	31.3

Table 31: CSTD: secondary metrics under the primary contract, on a percentage scale. IoU charges unmatched regions, and boundary F uses raw predictions.

Comparator	Image estimate [95% CI]	Scene-cluster CI	Deduplicated [95% CI]
DINOv2	11.3 [8.5, 14.1]	[8.5, 14.1]	11.3 [8.5, 14.1]
TokenCut + refinement	−10.4 [−12.4, −8.4]	[−12.4, −8.4]	−10.4 [−12.4, −8.4]
Mask2Former	31.0 [26.7, 35.4]	[26.7, 35.4]	31.0 [26.7, 35.4]
TextureSAM	−3.1 [−5.5, −0.6]	[−5.5, −0.6]	−3.1 [−5.5, −0.6]
Qwen + SAM 3 semantic	−10.2 [−12.3, −8.2]	[−12.3, −8.2]	−10.2 [−12.3, −8.2]
Qwen + SAM 3 detector	7.7 [4.0, 11.6]	[4.0, 11.6]	7.7 [4.0, 11.6]
GLaMM	13.8 [10.0, 17.5]	[10.0, 17.5]	13.8 [10.0, 17.5]
Sa2VA-4B	4.0 [0.7, 7.4]	[0.7, 7.4]	4.0 [0.7, 7.4]
Sa2VA (Qwen3)	8.0 [4.3, 11.7]	[4.3, 11.7]	8.0 [4.3, 11.7]
SAMTok	−1.8 [−4.6, 1.1]	[−4.6, 1.1]	−1.8 [−4.6, 1.1]

Table 32: CSTD: Detecture-minus-comparator covering differences, marginal image/scene intervals and deterministic deduplication sensitivity. Identified scene groups do not establish training independence.

M.5 Output Behavior and Alternative References

Dataset	Mean nonempty K	Correct K	Empty	Fully described	Residual
RWTD	2.07	76.3	0.0	0.0	2.5
RWTD-COCO	2.20	82.8	0.0	0.0	2.0
TextureADE	3.63	46.7	0.0	0.0	2.0
CSTD	2.00	100.0	0.0	0.0	1.9

Table 33: Detecture output behavior on all images. K counts nonempty positive regions on reference-valid pixels. Correct-count, empty-output and fully-described columns are percentages; residual is mean valid-pixel percentage. A disjoint map can still leave pixels undescribed.

The available Kaust256 copy of RWTD has different masks from the released evaluation references on common image identifiers. Its distribution is pinned separately; no original-release digest identifies which copy changed the source labels. Table 34 therefore tests sensitivity to this available reference set rather than replacing the primary labels. Supplied-text controls (Table 35) and predicted-category coarsening (Table 40) are likewise separate from the autonomous task.

Method	Released masks [95% CI]	Available-copy masks [95% CI]
DINOv2	71.7 [69.5, 74.0]	71.8 [69.6, 74.1]
TokenCut + refinement	87.2 [85.4, 88.9]	87.1 [85.2, 88.8]
Mask2Former	62.0 [58.4, 65.6]	61.8 [58.2, 65.5]
TextureSAM	70.2 [67.3, 73.2]	70.3 [67.3, 73.2]
Qwen + SAM 3 semantic	83.5 [81.1, 85.8]	83.8 [81.5, 86.2]
Qwen + SAM 3 detector	59.4 [55.4, 63.4]	59.6 [55.5, 63.5]
GLaMM	65.7 [62.3, 68.9]	65.9 [62.4, 69.0]
Sa2VA-4B	70.8 [67.6, 74.0]	71.1 [67.9, 74.4]
Sa2VA (Qwen3)	70.4 [66.8, 73.9]	70.7 [67.1, 74.2]
SAMTok	68.6 [65.3, 71.9]	68.8 [65.5, 72.1]
Detecture	79.4 [77.3, 81.5]	79.9 [77.7, 81.9]

Table 34: Available-copy RWTD mask sensitivity on shared image identifiers. These are alternate-reference comparisons, not a claim to have recovered an authenticated original annotation release. The original correction family is preserved.

Dataset	Detecture autonomous	SAM 3 semantic head	SAM 3 detector
RWTD	79.4 [77.3, 81.5]	83.6 [81.4, 85.8]	65.7 [62.1, 69.4]
TextureADE	75.7 [73.8, 77.7]	79.2 [77.3, 81.1]	58.8 [55.4, 62.1]

Table 35: Supplied-description SAM 3 controls. Reference phrases are oracle-assisted inputs, distinct from autonomous generation; the masks remain scored under the declared contract.

M.6 Original Shared-Format Experiment

The original paper applied one output-template instruction to LISA, a GSVA adaptation, GLaMM, older Sa2VA and Detecture. Each prompt's count field was the constant `multiple`, but native chat templates and required image tokens differed. Assistant-only grounding-marker extraction excluded marker examples in the prompt. GSVA required a generation adaptation because its released entry point was teacher-forced; that row is an adaptation, not native GSVA. Empty outputs remained scored.

This historical comparison is useful for prompt-transfer behavior, but weak shared-format emissions cannot establish that a model is structurally incapable of producing multiple regions. The native results above are the fairer headline comparison. Tables 36 and 37 retain the original setting and its reproduction differences. Legacy matched-only IoU and filled-boundary scores are explicitly separated from primary raw metrics.

Method	RWTD	RWTD-COCO	TextureADE	CSTD
LISA	54.4	49.3	34.4	52.2
GSVA extension	37.6	38.4	18.4	69.0
GLaMM	45.0	38.3	17.0	55.4
Sa2VA-4B	38.3	20.8	16.6	59.7
Sa2VA Qwen3-VL/SAM 3	1.7	0.4	2.4	0.2
Detecture	79.4	85.2	75.7	87.7

Table 36: Original shared-format experiment under the retained metric conventions. Method behavior is specific to this prompt/interface adaptation; it is not a general capacity claim.

Method	Dataset	Archived	Live	Difference [95% CI]
Detecture	RWTD	79.459	79.959	0.501 [−0.468, 1.444]
Detecture	RWTD-COCO	85.363	85.623	0.261 [−0.049, 0.604]
Detecture	TextureADE	74.793	75.090	0.296 [−0.274, 0.895]
Detecture	CSTD	88.355	88.046	−0.309 [−1.079, 0.453]
LISA	RWTD	54.192	54.263	0.071 [−0.046, 0.187]
LISA	RWTD-COCO	48.171	48.187	0.016 [−0.029, 0.072]
LISA	TextureADE	31.547	31.540	−0.007 [−0.028, 0.012]
LISA	CSTD	50.936	50.820	−0.115 [−0.430, 0.072]
GLaMM	RWTD	42.013	42.456	0.444 [−0.220, 1.209]
GLaMM	RWTD-COCO	35.582	34.828	−0.754 [−1.510, −0.105]
GLaMM	TextureADE	16.739	16.691	−0.048 [−0.463, 0.401]
GLaMM	CSTD	48.439	48.225	−0.214 [−1.361, 0.901]
GSVA extension	RWTD	42.298	42.165	−0.133 [−0.387, 0.115]
GSVA extension	RWTD-COCO	42.471	42.457	−0.014 [−0.489, 0.465]
GSVA extension	TextureADE	20.915	21.192	0.277 [−0.119, 0.755]
GSVA extension	CSTD	71.449	71.574	0.125 [−0.082, 0.337]
Sa2VA-4B	RWTD	39.871	40.063	0.192 [−2.495, 2.762]
Sa2VA-4B	RWTD-COCO	23.979	21.735	−2.243 [−5.016, 0.304]
Sa2VA-4B	TextureADE	17.934	18.097	0.163 [−1.990, 2.301]
Sa2VA-4B	CSTD	59.578	60.626	1.048 [−1.730, 3.886]

Table 37: Fresh historical-format reproduction versus archived results. The declared tolerance is a diagnostic of aggregate drift, not a statistical-equivalence margin. Exact original orchestration and numerical paths are not fully recoverable.

M.7 Additional Visual and Reference-Assisted Comparisons

The full comparison also includes a separate classical two-region texture method, reference-informed point prompts and predicted-category coarsening. These settings test different interfaces and are kept separate from the autonomous description-and-mask ranking.

Classical texture partitioning. The release port of STLD (Khan et al., 2015) uses its fixed binary initialization and descriptor/PDE settings on all 253 RWTD images. A typed-store repair makes the released mask write valid on the evaluated platform; the result is a port, not a reconstruction of the historical paper's scores. Detecture minus the port is -1.8 covering points (marginal 95% interval -4.7 to 1.1; Table 38). This late exploratory comparison is outside the main 40-contrast family.

Method	Covering [95% CI]	Matched IoU
Detecture	79.4 [77.3, 81.5]	72.2
STLD release port	81.2 [79.3, 83.1]	78.5

Table 38: Classical two-region comparison on RWTD. Covering uses the primary raw contract and marginal intervals. Matched IoU is the historical matched-only measure and differs from the penalized IoU in the secondary tables. The port emits no descriptions.

Reference-informed points. The TextureSAM point control (Table 39) provides one positive point at a reference region's distance-transform maximum, revealing both location and count. Automatic grid proposals receive neither. The oracle-assisted input denotes extra reference information, not an accuracy upper bound; the same checkpoint produces both outputs. Full per-image and source-group sensitivities remain in the compact evidence.

Dataset	Automatic grid	Reference-point oracle	Oracle minus automatic [95% CI]
RWTD	70.2 [67.3, 73.2]	68.9 [65.9, 71.9]	-1.4 [-3.6, 0.9]
RWTD-COCO	74.0 [71.3, 76.5]	70.0 [67.4, 72.5]	-4.0 [-5.8, -2.4]
TextureADE	64.1 [61.1, 67.0]	66.2 [63.7, 68.6]	2.1 [0.5, 3.7]
CSTD	90.7 [89.0, 92.3]	90.7 [88.9, 92.3]	0.0 [-1.8, 1.7]

Table 39: Automatic TextureSAM versus reference-informed points. Primary bidirectional covering and oracle-minus-automatic effects with marginal 95% intervals. The point arm is distinct from the autonomous task.

Predicted-category coarsening. Merging Mask2Former regions with the same predicted class tests a coarser, still reference-blind visual partition. It changes neither model parameters nor target masks. The source-bound comparisons in Table 40 preserve all four datasets and their marginal uncertainty.

Dataset	Panoptic	Class merge	Panoptic minus merge [95% CI]
RWTD	62.0	62.2	-0.3 [-0.6, 0.0]
RWTD-COCO	86.8	86.8	0.0 [0.0, 0.0]
TextureADE	81.1	81.2	-0.1 [-0.3, 0.0]
CSTD	56.6	56.6	0.0 [0.0, 0.0]

Table 40: Panoptic outputs and predicted-class merging. Primary covering and paired differences in percentage points. Category merging uses predicted labels; no reference region count selects the groups.

M.8 Reproducibility and Remaining Scope

The supplement preserves full prompt-transfer, input/decoding, output, metric, scene-dependence and numerical-repair analyses. Data and model hashes, code snapshots and finite-output checks accompany completed runs. Two early scoring populations were independently recomputed with retained code and agree byte-for-byte; missing historical orchestration copies are recorded rather than retrospectively inferred. The initial Detecture run manifest also lacks its own orchestration-script

digest. These limitations concern exact historical execution identity, while saved predictions and the final metric calculations remain auditable.

All frozen executions disable gradients and prohibit optimizer/backward operations; model hashes and parameter-version checks accompany inference. The historical RWTD-COCO supervised study in Appendix F predates this revision and is not pooled with the autonomous results reported here. No new learned checkpoint was created. The scientific gap that remains is independent appearance annotation inside source classes, paired with judgments of description faithfulness; existing mask overlap and slot-following evidence measure different aspects of the proposed task.