CALIBRATION-GUIDED STATE SELECTION AND CLASS-CONDITIONED FUSION FOR FROZEN DIFFUSION SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion representations vary substantially across network layers and denoising timesteps. Prior work exploits this variation through fixed states, task-level selection, shared aggregation, or pixel-adaptive rules, yet these approaches do not explicitly model class-dependent reliability after downstream decoding. We find that a state with strong average performance is not uniformly reliable across semantic classes. Based on this finding, we propose **C**lasswise **S**tate **R**outing (**CSR**), a label-calibrated, training-free framework for diffusion segmentation. Using a small labeled calibration set, CSR estimates classwise reliability over the joint layer–timestep space to guide compact state-region selection and class-conditioned response fusion before the frozen decoder. No model parameter is updated, and all routing decisions remain fixed during evaluation. Controlled ablations with the selected state pools held fixed show that class-conditioned weighting improves over a class-agnostic shared-weighting baseline on all four evaluated datasets. Across six backend-task settings and four benchmarks, the complete pipeline yields improvements in 10 of 12 open-vocabulary semantic segmentation (OVSS) comparisons and all 12 category-agnostic comparisons, with maximum gains of 12.79 and 6.88 mIoU percentage points. The procedure can also be reapplied to alternative diffusion backbones after backbone-specific recalibration.

1 INTRODUCTION

Training-free diffusion segmentors turn the internal responses of a frozen text-to-image model into semantic masks Wang et al. (2025); Sun et al. (2026); Tian et al. (2024); Couairon et al. (2024). This is possible because diffusion features and attention maps retain rich spatial and semantic information Tang et al. (2023a); Baranchuk et al. (2021); Ke et al. (2024). However, these responses are not obtained from a single representation. Every denoising timestep and network layer defines a different diffusion state. Changing this state can substantially alter the final mask even when the backbone and segmentation decoder remain fixed. State choice is therefore a central part of the readout procedure: which diffusion state should a segmentation method trust?

Existing methods answer this question in several ways. Some use a fixed state or select a state or feature subset for the complete task Kim et al. (2025a); Liu et al. (2026); Meng et al. (2024a). Others aggregate multiple layers and timesteps using a task-level rule Luo et al. (2023); Zuo et al. (2025); Meng et al. (2026). Recent work also selects timesteps for individual pixels from temporal stability Nakhli et al. (2026). Although these approaches establish state heterogeneity, their policies are generally shared across class channels or derived without classwise segmentation quality measured after the complete frozen pipeline. Consequently, a task-level score can hide class-dependent variation: a state that performs well on average while remaining unreliable for several classes.

We revisit diffusion-state selection from a classwise perspective. Instead of reducing each state to one quality score, we evaluate it through the complete frozen segmentation pipeline and retain its reliability for every class. This analysis reveals two complementary structures. First, globally reliable states occupy a compact region of the layer–timestep space. Second, class preferences remain heterogeneous within that region. On Cityscapes, 17 of 19 classes prefer a state other than the strongest shared state. Their classwise optima are distributed across multiple states, rather than concentrated

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: Qualitative comparisons between Seg4Diff and CSR. The examples are drawn from VOC20, Cityscapes, ADE20K, and PC59, covering both object-centric images and structured scene-level layouts. Compared with Seg4Diff, CSR generally recovers object regions more completely, better separates adjacent instances and produces more spatially coherent segmentation masks.

at a single global optimum. This Cityscapes result shows that a state with mean performance need not be a universal semantic expert, while the degree of class specialization may vary across datasets.

These observations motivate separating state-pool selection from response fusion. We introduce **Classwise State Routing (CSR)**, which uses a small labeled calibration set to estimate a class-by-state downstream reliability matrix. CSR selects a compact shared layer–timestep region by complementary class coverage and converts the fitted reliabilities into fixed class-conditioned fusion weights over that common pool. For a new image, the frozen diffusion backbone produces responses at the selected states and the precomputed weights fuse them before unchanged backend-native decoding. For the class-indexed OVSS branch, response-concentration calibration is subsequently applied to the fused channels. CSR therefore changes the fusion weights over one shared pool rather than selecting an exclusive state or rerunning diffusion for each class.

The router is a frozen calibration table rather than a learned network. Here, training-free means that no model parameter is optimized; the method is not annotation-free, and improvements over fixed zero-shot configurations also include the benefit of labeled calibration. The calibration subset remains disjoint from evaluation. In the OVSS experiments, the calibration subsets cover all benchmark categories; these experiments therefore evaluate class-conditioned routing for calibration-supported categories. We also extend the response-level framework to category-agnostic segmentation. Removing the class axis yields scalar state reliability for compact state selection and fusion.

We evaluate the framework on four benchmarks and six backend–task settings. The classwise OVSS formulation improves mIoU in 10 of 12 controlled comparisons, ranging from -0.66 to +12.79 mIoU percentage points, whereas the scalar category-agnostic specialization improves all 12 comparisons, with changes from +0.07 to +6.88 mIoU percentage points. Our contributions are three-fold:

- We characterize diffusion states by their downstream segmentation reliability over the joint layer–timestep space and find that class-dependent reliability differences remain within a compact region of high-performing states.

- We introduce Classwise State Routing, a label-calibrated framework that selects a compact shared state pool and applies fixed class-conditioned response fusion before unchanged backend-native decoding, without updating model parameters. Removing the class dimension yields a scalar variant for category-agnostic segmentation.

- We evaluate CSR across four benchmarks and six backend–task settings. With each dataset's selected state pool held fixed, controlled ablations show that class-conditioned weighting outperforms both shared-reliability weighting and uniform fusion on all four datasets.

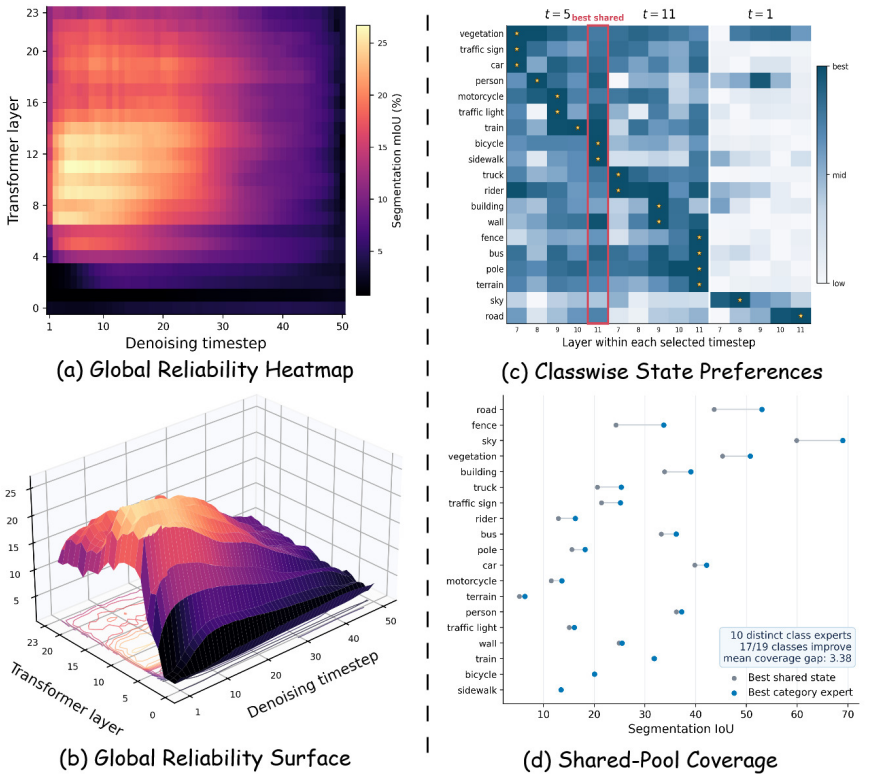

Figure 2: (a) Segmentation mIoU over 1,200 layer–timestep states; (b) Three-dimensional view of the same performance landscape. (c) Classwise reliability over the selected states; stars mark classwise maxima and red marks the strongest shared state. (d) Class IoU of the strongest shared state and each class's preferred state.

2 Related Work

Diffusion Representations across States. Pretrained diffusion models provide dense features for segmentation, correspondence, and depth, but their utility varies across timesteps, layers, and activations Baranchuk et al. (2021); Tang et al. (2023a); Zhao et al. (2023); Ke et al. (2024). Diffusion Hyperfeatures aggregates layers and timesteps at the task level, whereas A Tale of Two Features combines diffusion and DINO features Luo et al. (2023); Zhang et al. (2023). Broad activation analyses construct shared subsets or probe where semantics emerge Meng et al. (2024a); Kim et al. (2025b), while feature adaptation can reduce content shift and noise dependence Meng et al. (2024b); Stracke et al. (2025). Recent methods fuse several Stable Diffusion 3 (SD3) states, select one task-level timestep, or route timesteps per pixel Zuo et al. (2025); Liu et al. (2026); Nakhli et al. (2026). Together, they establish state heterogeneity, but selection remains task-, model-, or pixel-level. CSR instead estimates classwise reliability after the complete frozen decoder.

Diffusion Representations for Segmentation. Diffusion segmentation uses learned decoders or grounding modules, or derives masks directly from frozen attention and features Baranchuk et al. (2021); Xu et al. (2023); Zhao et al. (2023). DAAM and OVAM extract cross-attention maps, while early MM-DiT work studies SD3 cross-attention Tang et al. (2023b); Marcos-Manchón et al. (2024); Yamaguchi & Yanai (2024). DiffSegmenter and iSeg combine cross- and self-attention; MaskDiffusion, From Text to Mask, and SeeDiff refine frozen responses Wang et al. (2025); Sun et al. (2026); Kawano & Aoki (2024); Xiao et al. (2024); Park et al. (2025). DiffSeg and DiffCut produce category-agnostic partitions, whereas ConceptAttention modifies the attention readout Tian et al. (2024); Couairon et al. (2024); Helbling et al. (2025). Seg4Diff selects one shared expert layer and GoCA aggregates heads and layers before pixelwise rescaling Kim et al. (2025a); Meng et al. (2026); OVDiff and FreeDA instead use generated support images or prototypes Karazija et al. (2024); Barsellotti et al. (2024). None estimates class-dependent downstream reliability over the joint layer–timestep space, the focus of CSR.

3 METHOD

We introduce **Classwise State Routing** (CSR), a training-free framework that uses a small labeled calibration set to select a compact state pool and derive fixed class-conditioned fusion weights from downstream reliability. CSR then proceeds in three stages: (1) classwise downstream reliability estimation, (2) compact state-pool selection by complementary class coverage, and (3) classwise state routing followed by response concentration calibration. The first two stages determine which states are retained, while the third determines which retained states each class should trust. All reliability scores, routing weights, and calibration choices are computed on the calibration set and fixed during evaluation; no parameter of the diffusion backbone or segmentation decoder is updated.

3.1 CLASSWISE DOWNSTREAM STATE RELIABILITY

Diffusion-based segmentation requires converting internal denoising responses into a spatial prediction. Let $s = (t, l)$ denote a diffusion state at denoising timestep t and network layer l, and let $R_{s,c}$ be its spatial response to semantic class c. Existing readouts either extract a fixed state or select and aggregate activations using a rule shared across classes Meng et al. (2024a); Liu et al. (2026); Meng et al. (2026); Marcos-Manchón et al. (2024); Yamaguchi & Yanai (2024). Although these strategies differ in how responses are combined, they do not explicitly model classwise downstream reliability over the joint layer–timestep space. Even with the response extractor and segmentation decoder fixed, changing (t, l) changes which classes are localized reliably. We therefore ask not only which states are reliable on average, but which states each class should trust.

To expose this issue, we evaluate each candidate state through the complete frozen segmentation procedure. Given a labeled calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}$, let $\mathcal{C}_{\text{cal}}$ contain its C_{cal} supported semantic classes. For each $c \in \mathcal{C}_{\text{cal}}$, let $q_{s,c}$ denote the reliability of state s for class c, measured on the final segmentation prediction. Each state is represented by a reliability vector, while the conventional scalar score averages that vector:

$$\mathbf{q}_s = [q_{s,c}]_{c\in\mathcal{C}_{\text{cal}}}^{\top}, \qquad q_s = \frac{1}{C_{\text{cal}}} \sum_{c\in\mathcal{C}_{\text{cal}}} q_{s,c}. \tag{1}$$

Collecting these vectors over the candidate state space $\mathcal{S}_{\text{cand}}$ gives the class-by-state reliability matrix $\mathbf{Q} \in \mathbb{R}^{C_{\text{cal}}\times|\mathcal{S}_{\text{cand}}|}$, with $Q_{c,s} = q_{s,c}$. The scalar score indicates whether a state is globally strong, but discards which classes make it strong and which classes it fails to represent.

We first examine reliability at the global level. Fig. 2(a) evaluates all 50×24 layer–timestep states using the same frozen segmentation pipeline. The resulting mIoU varies from 0.78 to 26.68, with high-performing states forming a localized region rather than covering the complete denoising network. The heatmap identifies the location and extent of this region, while the corresponding three-dimensional surface makes its peak-and-ridge structure explicit.

Global variation explains why state selection matters, but it does not determine whether one globally strong state is sufficient. To visualize classwise variation at a compact operating budget, Fig. 2(c) and (d) examines the Cityscapes region retained by the selection procedure described below, with timesteps $\{5, 11, 1\}$ and layers $\{7, \ldots, 11\}$. As shown in Fig. 2(c), the row-wise maxima of the 19 Cityscapes classes occupy 10 of these 15 states rather than concentrating in the strongest shared-state column. Figure 2(d) quantifies the resulting loss: 17 classes have a more reliable state elsewhere in the same region, and the average gap between the classwise-best state and the strongest shared state is 3.38 IoU points. Thus, even a globally reliable region remains heterogeneous across classes.

Fig. 2 shows that a compact high-reliability region contains complementary class evidence that a single scalar-ranked state discards. CSR therefore selects one compact shared pool and assigns class-specific routing distributions over it, avoiding both a universal state and separate diffusion executions for individual classes (Fig. 3).

3.2 COMPACT STATE-POOL SELECTION BY CLASS COVERAGE

We instantiate the classwise reliability introduced above as dataset-level class IoU. For each candidate state s, we pass its response through the frozen segmentation pipeline and accumulate a single confusion matrix over all valid calibration pixels:

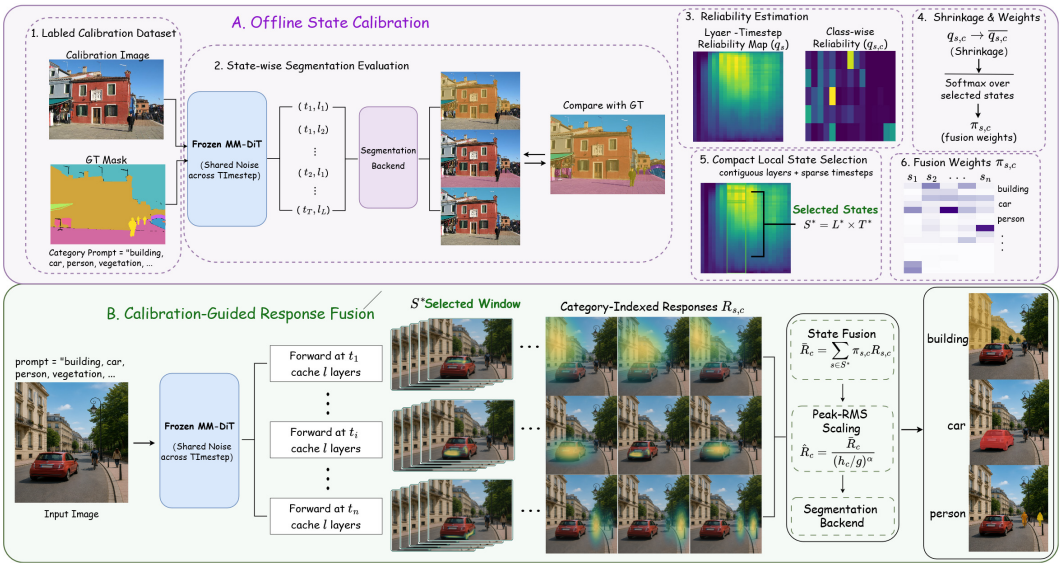

Figure 3: Overview of Classwise State Routing. **Top:** A labeled calibration set estimates downstream class-by-state reliability, selects a compact shared state pool, and fixes class-conditioned routing weights. **Bottom:** For a new image, the frozen backbone extracts selected-state responses, which are fused and rescaled before unchanged backend-native decoding.

$$q_{s,c} = \frac{\sum_{i,p} \mathbf{1}[y_i(p) = c \wedge \widehat{y}_{s,i}(p) = c]}{\sum_{i,p} \mathbf{1}[y_i(p) = c \vee \widehat{y}_{s,i}(p) = c]}, \qquad c \in \mathcal{C}_{\text{cal}}, \tag{2}$$

where $\widehat{y}_{s,i}$ is the segmentation predicted from state s. Computing one dataset-level confusio n matrix, rather than averaging per-image IoUs, ensures that false-positive predictions of classes absent from an image remain included.

Class-specific estimates can be unreliable when a class occurs in only a few calibration images. We therefore regularize the score of a supported class toward the class-averaged state reliability:

$$\widetilde{q}_{s,c} = \frac{n_c}{n_c + \kappa} q_{s,c} + \frac{\kappa}{n_c + \kappa} q_s, \qquad c \in \mathcal{C}_{\text{cal}}, \tag{3}$$

where n_c is the number of calibration images containing class c, and κ controls the regularization strength. As n_c increases, $\widetilde{q}_{s,c}$ approaches the observed class IoU; with limited support, it places more weight on q_s. For a query class $c \notin \mathcal{C}_{\text{cal}}$, class-specific reliability is unavailable, so we define $\widetilde{q}_{s,c} = q_s$. Unseen classes therefore fall back to the shared state reliability.

Motivated by the localized global structure in Fig. 2, we constrain the selected states to a contiguous layer window $\mathcal{L}$ and a sparse timestep set $\mathcal{T}$. Their class-coverage score is

$$F(\mathcal{L}, \mathcal{T}) = \frac{1}{C_{\text{cal}}} \sum_{c \in \mathcal{C}_{\text{cal}}} \max_{(t,l) \in \mathcal{T} \times \mathcal{L}} \widetilde{q}_{(t,l),c}. \tag{4}$$

The maximization over states is performed separately for each class. Consequently, F rewards pools whose members cover different class preferences rather than pools containing several globally strong but redundant states.

We constrain $\mathcal{L}$ to contain N_l consecutive layers and $\mathcal{T}$ to contain N_t timesteps. Let $s_0 = (t_0, l_0) \in \arg\max_s q_s$ be the globally strongest individual state. Its timestep t_0 initializes every candidate layer window with $\mathcal{T}_{\mathcal{L}}^{(1)} = \{t_0\}$. The remaining $N_t - 1$ timesteps are added according to their marginal class-coverage gain:

$$t^*_{\mathcal{L},k} = \arg\max_{t \notin \mathcal{T}_{\mathcal{L}}^{(k-1)}} F\left(\mathcal{L}, \mathcal{T}_{\mathcal{L}}^{(k-1)} \cup \{t\}\right), \qquad \mathcal{T}_{\mathcal{L}}^{(k)} = \mathcal{T}_{\mathcal{L}}^{(k-1)} \cup \{t^*_{\mathcal{L},k}\}, \qquad k = 2, \ldots, N_t. \tag{5}$$

This anchored greedy search is deterministic but is not an exact optimizer of the combinatorial objective. The anchor retains the timestep of the strongest individual state and reduces search variation, although it can be suboptimal when l_0 lies outside a candidate layer window.

Let $\mathcal{W}_{N_l}$ be the set of all contiguous N_l-layer windows. After completing timestep selection for every candidate window, we choose

$$\mathcal{L}^* = \arg\max_{\mathcal{L}\in\mathcal{W}_{N_l}} F\Big(\mathcal{L}, \mathcal{T}_{\mathcal{L}}^{(N_t)}\Big), \qquad \mathcal{T}^* = \mathcal{T}_{\mathcal{L}^*}^{(N_t)}, \qquad \mathcal{S}^* = \mathcal{T}^* \times \mathcal{L}^*. \tag{6}$$

Unlike independently ranking states by global reliability, this procedure adds timesteps according to the class coverage they contribute to the current pool. The resulting $\mathcal{S}^*$ is shared across classes and serves as the compact support of the router.

3.3 Classwise State Routing and Calibrated Fusion

Once the shared state pool is fixed, we convert its classwise reliability into a routing distribution. Unlike aggregation with one shared state weight, this router allows two classes queried in the same image to rely on different subsets of $\mathcal{S}^*$. Let μ_c and σ_c be the mean and standard deviation of $\{\widetilde{q}_{s,c} : s \in \mathcal{S}^*\}$. We standardize the reliability scores across the selected states as

$$z_{s,c} = \frac{\widetilde{q}_{s,c} - \mu_c}{\sigma_c + \varepsilon}, \qquad \pi_{s,c} = \frac{\exp(\rho z_{s,c})}{\sum_{j\in\mathcal{S}^*}\exp(\rho z_{j,c})}, \tag{7}$$

where ε is a small constant for numerical stability and ρ controls the concentration of the state weights. The vector $\boldsymbol{\pi}_c = [\pi_{s,c}]_{s\in\mathcal{S}^*}$ is the dense routing distribution for class c. To control the active routing budget, let TopK_c denote the K states with the largest $\pi_{s,c}$, where $K \in \{1, \ldots, |\mathcal{S}^*|\}$. The fixed routing weights and the corresponding continuous-space response are

$$w_{s,c} = \frac{\pi_{s,c}\,\mathbf{1}[s \in \mathrm{TopK}_c]}{\sum_{j\in\mathrm{TopK}_c}\pi_{j,c}}, \qquad \overline{R}_c(p) = \sum_{s\in\mathcal{S}^*} w_{s,c} R_{s,c}(p). \tag{8}$$

Classwise routing aggregates complementary evidence across diffusion states, but the routed responses may still differ in scale and spatial concentration. Because downstream decoders are sensitive to this cross-class contrast, we use a **response concentration score** for calibrated channel scaling. For a response R over spatial positions $\mathcal{P}$, define $\mathrm{RMS}(R) = \sqrt{|\mathcal{P}|^{-1}\sum_{p\in\mathcal{P}} R(p)^2}$. For each class, we compute

$$h_c = \frac{\max(\|\overline{R}_c\|_\infty, \varepsilon_{\mathrm{num}})}{\max(\mathrm{RMS}(\overline{R}_c), \varepsilon_{\mathrm{num}})}, \qquad g = \mathrm{gmean}(h_1, \ldots, h_C), \qquad \widehat{R}_c(p) = \frac{\overline{R}_c(p)}{(h_c/g)^\alpha}. \tag{9}$$

Here, $\varepsilon_{\mathrm{num}} = 10^{-8}$ prevents an all-zero or underflowed channel from producing an undefined ratio, and g is evaluated in log space for numerical stability. The signed exponent α controls the scaling strength and direction: $\alpha > 0$ suppresses channels with $h_c > g$, $\alpha < 0$ reverses the effect, and $\alpha = 0$ disables scaling.

3.4 Category-Agnostic Specialization

Category-agnostic segmentation follows the same reliability–selection–fusion principle after removing the class axis. It is therefore a scalar specialization of CSR rather than a classwise router. For state $s = (t, l)$, input image x, and sampled diffusion noise z, let $R_s = \mathrm{PreProcess}(\mathrm{Extract}(x, z, s))$ be the aligned continuous response before grouping. Here, $\mathrm{Extract}$ obtains the frozen diffusion response, and $\mathrm{PreProcess}$ converts it into a representation that can be fused across states. Applying $\mathrm{PostProcess}_\theta$ produces an unlabeled partition. Since its cluster IDs are permutation-dependent, Hungarian matching is used only on the calibration set to compute the scalar reliability q_s.

The candidate states use the same timestep–layer space as the compact pool selection. Without classwise coverage, we replace the Cartesian constraint by an admissible family

$$\mathfrak{A} = \{\mathcal{S} \subseteq \mathcal{S}_{\mathrm{cand}} : 1 \le |\mathcal{S}| \le B,\ \mathrm{Compat}(\mathcal{S}) = 1\}, \tag{10}$$

where B is the state budget and Compat requires aligned response dimensions, indexing, semantics, and normalization. For a candidate $\mathcal{S} \in \mathfrak{A}$, we reuse the scalar form of the preceding fusion rule:

$$\pi_s \propto \exp\left(\rho\frac{q_s - \mu}{\sigma + \varepsilon}\right), \qquad \overline{R} = \sum_{s\in\mathcal{S}} w_s R_s, \tag{11}$$

Table 1: Main results for (a) classwise CSR on OVSS, using VOC20, and (b) scalar CSR for category-agnostic partition recovery, using VOC21 (%). Frequency-weighted IoU (FWIoU) and mean IoU (mIoU) are reported. DiffSegm. denotes DiffSegmenter. The non-SD3 rows retain their original source architectures; SD3 denotes the controlled comparison setting. Bold values indicate the better result within each controlled SD3 baseline/CSR pair. Rows labeled Δ report absolute percentage-point changes relative to the corresponding SD3 baseline.

Method	Arch.	VOC		Cityscapes		PC59		ADE	
		FWIoU	mIoU	FWIoU	mIoU	FWIoU	mIoU	FWIoU	mIoU
(a) Classwise OVSS (VOC20)									
DiffSegm.	SD1.5	87.34	88.05	44.36	19.28	54.87	44.97	44.75	26.59
iSeg	SD2.1	78.10	78.46	50.82	21.63	64.08	52.45	49.64	32.34
DiffSegm.	SD3	**90.60**	**90.92**	35.52	23.07	56.06	47.03	44.34	33.38
+ CSR	SD3	89.77	90.36	**48.22**	**35.86**	**60.45**	**51.84**	**50.34**	**37.04**
Δ *vs. SD3*	–	-0.83	-0.56	$+12.70$	$+12.79$	$+4.39$	$+4.81$	$+6.00$	$+3.66$
iSeg	SD3	**88.74**	**89.36**	48.06	21.10	61.68	49.48	50.91	33.38
+ CSR	SD3	88.18	88.70	**57.25**	**25.84**	**63.37**	**51.30**	**53.84**	**36.44**
Δ *vs. SD3*	–	-0.56	-0.66	$+9.19$	$+4.74$	$+1.69$	$+1.82$	$+2.93$	$+3.06$
Seg4Diff	SD3	88.08	89.43	49.35	28.34	56.70	48.73	43.79	33.88
+ CSR	SD3	**88.85**	**90.48**	**56.11**	**33.37**	**60.32**	**52.31**	**48.92**	**37.35**
Δ *vs. SD3*	–	$+0.77$	$+1.05$	$+6.76$	$+5.03$	$+3.62$	$+3.58$	$+5.13$	$+3.47$
(b) Category-agnostic partition recovery (VOC21)									
DiffSeg	SD1.5	56.12	44.41	65.56	22.17	59.43	48.41	52.24	35.96
DiffCut	SSD-1B	72.32	58.17	66.38	26.48	66.58	54.26	46.70	41.07
DiffSeg	SD3	51.08	38.86	49.90	24.30	57.28	46.07	38.41	39.01
+ CSR	SD3	**53.87**	**40.41**	**52.49**	**25.42**	**58.36**	**47.13**	**42.67**	**39.87**
Δ *vs. SD3*	–	$+2.79$	$+1.55$	$+2.59$	$+1.12$	$+1.08$	$+1.06$	$+4.26$	$+0.86$
DiffCut	SD3	72.90	48.69	55.38	24.92	59.74	47.55	**53.02**	39.12
+ CSR	SD3	**73.56**	**50.50**	**60.56**	**27.74**	**59.98**	**48.00**	51.79	**40.52**
Δ *vs. SD3*	–	$+0.66$	$+1.81$	$+5.18$	$+2.82$	$+0.24$	$+0.45$	-1.23	$+1.40$
Seg4Diff	SD3	49.13	38.99	61.18	27.50	53.13	42.76	**44.36**	35.14
+ CSR	SD3	**60.77**	**45.87**	**65.72**	**27.57**	**57.05**	**46.03**	42.65	**36.90**
Δ *vs. SD3*	–	$+11.64$	$+6.88$	$+4.54$	$+0.07$	$+3.92$	$+3.27$	-1.71	$+1.76$

where μ and σ are computed over $\{q_s : s \in \mathcal{S}\}$ and w_s is the normalized top-K truncation of π_s. As $\overline{R}$ has no class axis, classwise weighting becomes scalar weighting without channel scaling.

Finally, the states and backend-native parameters are selected and fixed for evaluation:

$$(\mathcal{S}^*, \theta^*) = \underset{\mathcal{S}\in\mathfrak{A},\, \theta\in\Theta}{\arg\max}\, \mathcal{M}\big(\mathrm{PostProcess}_\theta(\overline{R})\big)\,, \tag{12}$$

where Θ is the finite parameter grid and $\mathcal{M}$ is the partition-quality metric. Backend-specific response construction, candidate constraints, and parameter grids are reported in the appendix.

4 Experiments

4.1 Experimental Setup

We evaluate classwise CSR for OVSS and its scalar specialization for category-agnostic partition recovery on PASCAL VOC Everingham et al. (2010), Cityscapes Cordts et al. (2016), PC59 Mottaghi et al. (2014), and ADE20K Zhou et al. (2019); OVSS uses VOC20 and category-agnostic evaluation uses VOC21. All controlled comparisons use SD3 Esser et al. (2024). We use class-balanced subsets of 50, 30, 100, and 50 labeled training images, respectively. These subsets fix state selection, routing, and calibration choices. This subset remains disjoint from evaluation, and no model parameter is updated. Each paired comparison shares preprocessing, inputs, noise, valid pixels, and evaluation. We report pAcc, FWIoU, mAcc, and mIoU, using per-image Hungarian matching only for category-agnostic outputs. Further implementation details are provided in the appendix.

Table 2: Controlled Seg4Diff/SD3 ablation for classwise OVSS (%). All multi-state variants use the same calibration-selected 15-state pool. Uniform, shared-reliability, and class-conditioned fusion differ in their weighting schemes, while Full CSR further applies response calibration.

	VOC20		Cityscapes		PC59		ADE20K	
Variant	FWIoU	mIoU	FWIoU	mIoU	FWIoU	mIoU	FWIoU	mIoU
Calibrated single state	88.08	89.43	49.35	28.34	56.70	48.73	43.79	33.88
Uniform fusion	87.71	89.34	51.60	30.36	59.41	51.15	46.68	36.87
Shared-reliability fusion	88.12	89.65	49.62	30.10	58.69	50.62	45.93	36.25
Class-conditioned fusion	88.32	89.88	54.86	31.86	60.26	51.74	48.78	37.27
Full CSR	**88.85**	**90.48**	**56.11**	**33.37**	**60.32**	**52.31**	**48.92**	**37.35**

Table 3: Effect of scalar state fusion across three frozen segmentation backends on VOC21 and Cityscapes (%). Pixel accuracy (pAcc) and Frequency-weighted IoU (FWIoU) are reported.

		Single State		+ State Fusion		Gain	
Backend	Dataset	pAcc	FWIoU	pAcc	FWIoU	pAcc	FWIoU
DiffSeg	VOC21	59.58	52.08	62.06	53.87	+2.47	+1.79
	Cityscapes	60.71	49.90	63.48	52.49	+2.77	+2.59
DiffCut	VOC21	83.59	73.32	83.74	73.56	+0.15	+0.25
	Cityscapes	69.25	59.30	71.19	60.56	+1.94	+1.26
Seg4Diff	VOC21	67.81	57.82	71.40	60.77	+3.59	+2.95
	Cityscapes	69.63	60.29	76.54	65.72	+6.91	+5.44

4.2 Main Results

Classwise routing for OVSS. The upper block of Table 1 reports FWIoU and mIoU consistently across all datasets. DiffSegmenter-SD1.5 and iSeg-SD2.1 are their original results; the SD3 rows form the controlled baseline/CSR comparisons. The largest gains occur on Cityscapes, reaching 12.79, 4.74, and 5.03 points for DiffSegmenter, iSeg, and Seg4Diff, respectively. These structured scenes contain classes with different spatial extent and semantic roles, consistent with the heterogeneous preferences in Fig. 2. The complete framework also improves every backend on PC59 and ADE20K, although these end-to-end gains do not by themselves isolate the contribution of classwise routing. As shown below, classwise routing contributes more strongly on Cityscapes, whereas shared-state pooling accounts for most of the ADE20K improvement. On VOC20, DiffSegmenter and iSeg slightly outperform their routed variants, while CSR still improves Seg4Diff.

Category-agnostic specialization. Removing the class axis reduces CSR to scalar reliability-based state selection and fusion. Since cluster identities are permutation-dependent, we evaluate partition recovery using per-image Hungarian matching. The lower block of Table 1 reports FWIoU and mIoU on every dataset. Under the controlled SD3 protocol, the scalar specialization improves mIoU for all three backends on every dataset. The largest improvement is obtained by Seg4Diff on VOC21 (+6.88), followed by Seg4Diff on PC59 (+3.27) and DiffCut on Cityscapes (+2.82). These results support calibration-guided scalar state selection and fusion when class identities are unavailable.

4.3 Ablation Studies

Separating pooling, reliability weighting, and class conditioning. Table 2 extends the Seg4Diff/SD3 ablation to all four datasets and adds shared-reliability fusion as a control for isolating class conditioning. Uniform fusion improves both FWIoU and mIoU over the calibrated single state on Cityscapes, PC59, and ADE20K, while the two metrics change by -0.37 and -0.09 points, respectively, on VOC20. Shared-reliability fusion improves both metrics over uniform fusion on VOC20 but decreases them on the other three datasets, showing that a single reliability weight per state does not consistently improve uniform fusion.

With the selected pool, ρ, and top-K held fixed, and with response calibration disabled, class-conditioned fusion improves over shared-reliability fusion on both metrics for every dataset. The mIoU gains are 0.24, 1.76, 1.13, and 1.02 points on VOC20, Cityscapes, PC59 and ADE20K; FWIoU gains range from 0.20 to 5.24 points. Adding response calibration yields the best results, supporting class-conditioned weighting within the selected pool with dataset-dependent effect sizes.

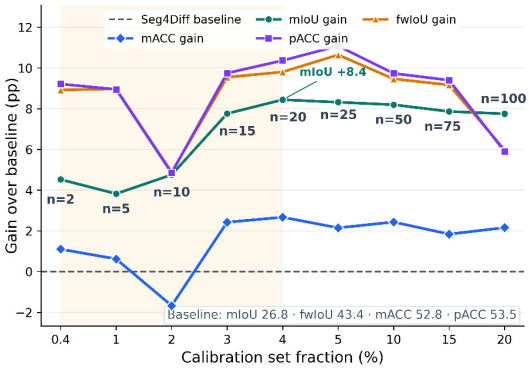

Figure 4: Effect of calibration set size on Cityscapes. Curves show full-validation performance gains over Seg4Diff as the number of labeled training calibration images increases.

Table 4: Online inference cost of CSR on ADE20K. The 3T-5L state budget is used for the main OVSS experiments. Latency per forward is total latency divided by the number of SD3 forwards. Measurements use FP16, batch size 1, and an NVIDIA A100, with 50 measured images after 5 warm-up images.

Metric	Baseline	1T-5L	3T-5L
States ↓	1	5	15
SD3 forwards ↓	1	1	3
Latency (ms/img) ↓	204.15	248.98	737.46
Throughput (img/s) ↑	4.898	4.016	1.356
Latency/forward (ms) ↓	204.15	248.98	245.82
Peak memory (MiB) ↓	8350.4	8620.0	8643.5

Scalar specialization across backends. Table 3 isolates the change from a single state to reliability-weighted multi-state fusion for category-agnostic segmentation. All other components remain unchanged. Complete four-metric results for all datasets are provided in the appendix. Scalar state fusion improves every displayed endpoint. The largest gain is obtained by Seg4Diff on Cityscapes (+6.91/+5.44). Together with Table 2, these results distinguish the broader effect of multi-state fusion from the additional effect of class-conditioned weighting when class identities are available.

4.4 Further Analysis

Calibration efficiency. Figure 4 shows that only two class-balanced Cityscapes training images already improve mIoU and FWIoU by 4.52 and 8.92 points. Increasing this train-only calibration set to 20 images raises the mIoU gain to 8.44 points, while 25 images yield gains of 11.08 pAcc and 10.64 FWIoU points. The mIoU and FWIoU gains remain positive across all evaluated budgets, showing that useful classwise routing can be calibrated from a small labeled subset.

Calibration-subset stability. We use five content-verified unique training manifests per dataset and rerun all 1,200 state scores and pool selection while holding the routing hyperparameters fixed. All ten runs yield positive full-split mIoU gains: 4.60 ± 0.82 points on Cityscapes and 3.15 ± 0.25 on ADE20K. Mean pairwise Jaccard is 0.89 and 0.39 for image manifests but only 0.47 and 0.20 for selected states. Thus, different calibration subsets can produce distinct yet similarly effective pools; the procedure does not identify a unique state pool.

Portability across diffusion backbones. Replacing SD3 with FLUX.1-dev or SD 3.5 and recalibrating each frozen backbone improves all six FLUX.1-dev and eight SD3.5 dataset–metric pairs (Appendix), supporting portability beyond SD3 after backbone-specific recalibration.

Inference efficiency. Table 4 profiles two state budgets on ADE20K. 1T-5L extracts five layer responses in one SD3 forward with 248.98 ms versus 204.15 ms per image and 3.23% more peak memory. The main 3T-5L setting uses three forwards and reaches 737.46 ms per image, while retaining similar amortized latency to 1T-5L (245.82 versus 248.98 ms per forward) with a 3.51% memory increase. These results indicate that the additional cost mainly follows the number of timestep evaluations, whereas multi-layer response extraction introduces modest memory overhead.

5 Conclusion

CSR estimates downstream classwise reliability to select a compact shared state pool and fuse its responses with fixed class-conditioned weights before frozen decoding. Across four benchmarks and six backend–task settings, it improves mIoU in 10 of 12 OVSS comparisons, while its scalar specialization improves all 12 category-agnostic comparisons. Ablations isolate the effects of complementary pooling, class-conditioned weighting, and response calibration, while backbone-specific recalibration yields gains with FLUX.1-dev and SD3.5. CSR uses labeled data only for offline calibration; resulting choices remain fixed during evaluation, and no model parameter is optimized.

AI USE STATEMENT

Generative AI tools were used to assist with drafting parts of the manuscript, polishing and improving the clarity of the writing, retrieving and discovering relevant literature, and generating or refining code. All AI-generated or AI-assisted text, references, and code were carefully reviewed, verified, and, when necessary, revised by the authors before inclusion in the paper or implementation. The authors take full responsibility for the final manuscript, references, experimental code, and reported results.

REFERENCES

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3689–3698, 2024.

Black Forest Labs. FLUX: Official Inference Code for FLUX.1 Models, 2024. URL `https://github.com/black-forest-labs/flux`. FLUX.1-dev model and inference implementation.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. *Advances in Neural Information Processing Systems*, 37:13548–13578, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 22946–22963, 2025.

Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 299–317. Springer, 2024.

Yasufumi Kawano and Yoshimitsu Aoki. Maskdiffusion: Exploiting pre-trained diffusion models for semantic segmentation. *IEEE Access*, 12:127283–127293, 2024.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–9502, 2024.

Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4diff: Unveiling open-vocabulary semantic segmentation in text-to-image diffusion transformers. *Advances in Neural Information Processing Systems*, 38:71685–71724, 2025a.

Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4659–4669, 2025b.

Changyu Liu, James Chenhao Liang, Wenhao Yang, Yiming Cui, Jinghao Yang, Tianyang Wang, Qifan Wang, Dongfang Liu, and Cheng Han. A-select: Automatic timestep selection for diffusion transformer representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2026) Proceedings*, 2026.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C. SanMiguel, and José M. Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9242–9252, 2024.

Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. *Advances in Neural Information Processing Systems*, 37:55141–55177, 2024a.

Benyuan Meng, Qianqian Xu, Zitai Wang, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Suppress content shift: Better diffusion features via off-the-shelf generation techniques. *Advances in Neural Information Processing Systems*, 37:18910–18939, 2024b.

Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, Longtao Huang, and Qingming Huang. Making training-free diffusion segmentors scale with the generative power. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35861–35871, 2026.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, 2014.

Ramin Nakhli, Mahesh Ramachandran, and Luca Ballan. Unlocking diffusion hierarchies: Adaptive timestep selection for zero-shot segmentation. *arXiv preprint arXiv:2606.15590*, 2026.

Joon Hyun Park, Kumju Jo, and Sungyong Baik. Seediff: Off-the-shelf seeded mask generation from diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6406–6415, 2025.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 117–127, 2025.

Lin Sun, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. iseg: An iterative refinement-based framework for training-free segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023a.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 5644–5659, 2023b.

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3554–3563, 2024.

Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *IEEE Transactions on Image Processing*, 2025.

Changming Xiao, Qi Yang, Feng Zhou, and Changshui Zhang. From text to mask: Localizing entities using the attention of text-to-image diffusion models. *Neurocomputing*, 610:128437, 2024.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2955–2966, 2023.

Rento Yamaguchi and Keiji Yanai. Exploring cross-attention maps in multi-modal diffusion transformers for training-free semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision Workshops*, pp. 260–274, 2024.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa F. Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5729–5739, 2023.

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.

Jing Zuo, Jiaqi Wang, Yonggang Qi, and Yi-Zhe Song. Fuse2match: Training-free fusion of flow, diffusion, and contrastive models for zero-shot semantic matching. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

A Additional Details

A.1 Calibration details

We use the notation of the main method. Let $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}$ be the calibration set and $s = (t, l)$ a diffusion state. For OVSS, $R_{s,c}(p)$ denotes the response for category c at spatial position p. Applying the frozen downstream segmentation procedure to state s produces $\widehat{y}_{s,i}$ for image x_i.

Let Ω_i be the valid calibration pixels of image i. The aggregate category counts are

$$\mathrm{TP}_{s,c} = \sum_i \sum_{p \in \Omega_i} \mathbf{1}\begin{bmatrix} y_i(p) = c, \\ \widehat{y}_{s,i}(p) = c \end{bmatrix}, \tag{13}$$

$$\mathrm{FP}_{s,c} = \sum_i \sum_{p \in \Omega_i} \mathbf{1}\begin{bmatrix} y_i(p) \neq c, \\ \widehat{y}_{s,i}(p) = c \end{bmatrix}, \tag{14}$$

$$\mathrm{FN}_{s,c} = \sum_i \sum_{p \in \Omega_i} \mathbf{1}\begin{bmatrix} y_i(p) = c, \\ \widehat{y}_{s,i}(p) \neq c \end{bmatrix}. \tag{15}$$

where $q_{s,c}$ is defined only for categories supported by the calibration set. Predictions of a supported category that is absent from an image contribute to $\mathrm{FP}_{s,c}$. The support count n_c is used only in the regularized estimate $\widetilde{q}_{s,c}$. For a query category $c \notin \mathcal{C}_{\text{cal}}$, category-specific reliability is unavailable, and we follow the main method by setting $\widetilde{q}_{s,c} = q_s$.

For the category-agnostic special case, we retain the main-method definition

$$R_s = \mathrm{PreProcess}(\mathrm{Extract}(x_i, z_i, s)), \qquad \widehat{y}_{s,i} = \mathrm{PostProcess}_\theta(R_s). \tag{16}$$

Table S1: Response-level interface for integrating the proposed method with different segmentation backends.

Integration	$\mathrm{PreProcess}_H$ output	$\mathrm{PostProcess}_H$
OVSS backends	Category-indexed I2T response	Native semantic decoding or mask completion
DiffSeg	Head-aggregated, nonnegative, row-normalized I2I affinity	Anchor-mask generation and KL-based merging
DiffCut	Diffusion-feature graph affinity	Recursive normalized-cut partitioning
Seg4Diff grouping	I2T-derived pre-grouping response	Native response grouping and mask merging

Here, R_s is the aligned continuous response before grouping and $\widehat{y}_{s,i}$ is an unlabeled partition. Per-image Hungarian matching resolves its label permutation only when computing the scalar calibration reliability q_s.

OVSS selects a Cartesian region $\mathcal{S} = \mathcal{T} \times \mathcal{L}$: consecutive layers at one timestep share a diffusion forward, while each additional timestep requires another forward. Category-agnostic segmentation instead selects $\mathcal{S} \in \mathcal{A}$ from the admissible family defined in the main method.

Similar category reliabilities yield near-uniform fusion weights. In Peak-RMS scaling, $|\alpha|$ controls strength, its sign controls direction, and $\alpha = 0$ disables scaling. Category-agnostic responses omit this operation. Under argmax decoding, the shared factor g^α preserves the ordering of existing channels; this property does not hold for fixed-threshold decoders.

A.2 Common response-level interface and specialization boundary

All integrations share the following response-level interface:

$$A_s = \mathrm{PreProcess}_H(\mathrm{Extract}(x, z, s)), \qquad \widehat{y} = \mathrm{PostProcess}_H(\mathrm{Fuse}\{A_s : s \in \mathcal{S}^*\}). \tag{17}$$

Following the calibration definitions above, fusion is deliberately placed between $\mathrm{PreProcess}_H$ and $\mathrm{PostProcess}_H$, before discrete grouping decisions. Table S1 summarizes the continuous response passed between these two stages for each segmentation backend.

Algorithm 1 Calibration-guided category-agnostic adapter

Require: $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}$, $\mathcal{A}$, Θ,θ_0 Extract, PreProcess, $\text{PostProcess}_\theta$, ρ, and K

Calibration

Cache one noise tensor z_i for every x_i
for all $s \in \mathcal{S}_{\text{cand}}$ **do**
  for all $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ **do**
    $R_{s,i} \leftarrow \text{PreProcess}\big(\text{Extract}(x_i, z_i, s)\big)$
    $\widehat{y}_{s,i} \leftarrow \text{PostProcess}_{\theta_0}(R_{s,i})$
    $\widetilde{y}_{s,i} \leftarrow \text{HungarianMatch}(\widehat{y}_{s,i}, y_i)$
  end for
  $q_s \leftarrow \text{mIoU}\big(\{\widetilde{y}_{s,i}\}, \{y_i\}\big)$
end for
$\mathcal{C} \leftarrow \text{BuildCandidates}(\mathcal{A}, \Theta)$
for all $(\mathcal{S}, \theta) \in \mathcal{C}$ **do**
  $w_s \leftarrow \text{ScalarWeights}\big(\{q_s : s \in \mathcal{S}\}, \rho, K\big)$
  for all $(x_i, y_i) \in \mathcal{D}_{\text{ation-guidedcategory-acal}}$ **do**
    $\overline{R}_i \leftarrow \sum_{s \in \mathcal{S}} w_s R_{s,i}$
  end for
  $\text{score}(\mathcal{S}, \theta) \leftarrow \text{mIoU}\left(\left\{\text{PostProcess}_\theta(\overline{R}_i)\right\}, \{y_i\}\right)$
end for
$(\mathcal{S}^*, \theta^*) \leftarrow \arg\max_{(\mathcal{S},\theta)\in\mathcal{C}} \text{score}(\mathcal{S}, \theta)$

Inference for image x

Sample one z and share it across all $s \in \mathcal{S}^*$
for all $s \in \mathcal{S}^*$ **do**
  $R_s \leftarrow \text{PreProcess}\big(\text{Extract}(x, z, s)\big)$
end for
$\overline{R} \leftarrow \sum_{s \in \mathcal{S}^*} w_s R_s$
return $\text{PostProcess}_{\theta^*}(\overline{R})$

Across integrations, calibration scores states through the task-native decoder, and fusion occurs in an aligned continuous space before discrete prediction. OVSS uses Cartesian regions, category coverage, category-conditioned weights, and Peak-RMS scaling; category-agnostic adapters use scalar reliability and backend-specific state selection. Discrete masks are not fused because their cluster identifiers are permutation-dependent.

A.3 Backend-specific state search

For a regular layer–timestep grid, candidate regions are ranked by

$$J_{\text{peak}}(\mathcal{L}, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \max_{l \in \mathcal{L}} q_{(t,l)}, \tag{18}$$

and resolve ties using the plateau score

$$J_{\text{plateau}}(\mathcal{L}, \mathcal{T}) = \frac{1}{|\mathcal{T}||\mathcal{L}|} \sum_{t \in \mathcal{T}} \sum_{l \in \mathcal{L}} q_{(t,l)}. \tag{19}$$

J_{peak} measures the strongest state at each timestep, while J_{plateau} favors reliability across the complete region. DiffSeg ranks Cartesian regions by $(J_{\text{peak}}, J_{\text{plateau}})$. Seg4Diff ranks its fixed 12-state grid by calibration mIoU and retains the selected state budget. DiffCut ranks its local scan, forms compatible prefixes containing two or three states, and evaluates the graph-parameter grid. These constructions instantiate the admissible family $\mathcal{A}$ defined in the main method. Algorithm 1 summarizes the complete calibration and inference procedure for the category-agnostic adapter.

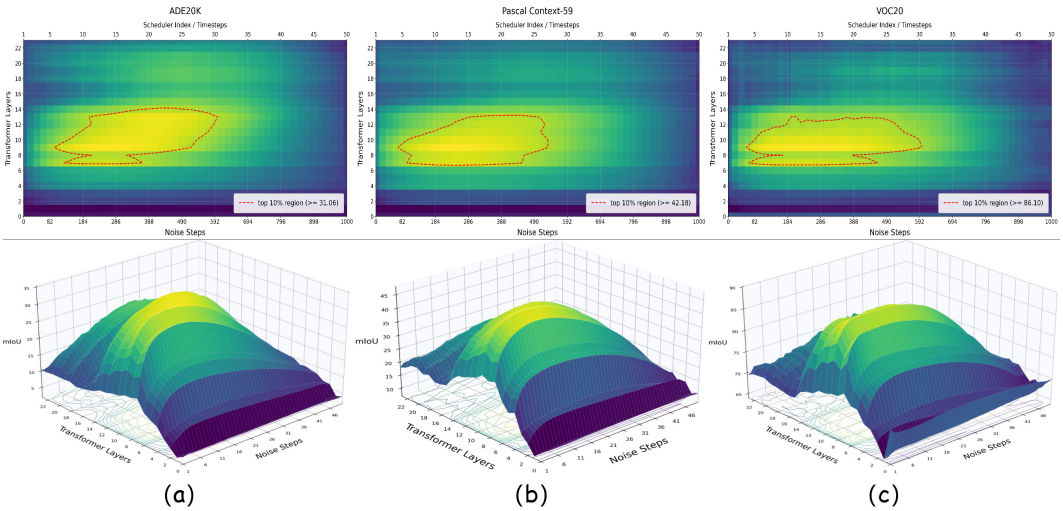

Figure S1: Cross-dataset state-reliability heatmaps and three-dimensional surfaces.

The scalar weights follow the normalized top-K rule in the main method; explicit candidate weights are used only when specified by a backend. Hungarian matching is restricted to calibration and evaluation, while inference directly returns the predicted partition. The selected $(\mathcal{S}^*, \theta^*)$ is frozen for evaluation.

A.4 Category-dependent state reliability

The analyses in main Fig. 2 use Stable Diffusion 3 Medium with the frozen Seg4Diff OVSS pipeline. For each layer–timestep state, Seg4Diff constructs category-indexed image-to-text responses and applies its native hard-assignment semantic decoder. We keep this segmentation procedure fixed across states, so the measured overall mIoU q_s and class IoU $q_{s,c}$ reflect only the choice of diffusion state.

The 3×5 region displayed in main Fig. 2 is the Cityscapes configuration retained by the category-coverage selector: $\mathcal{T}^* = \{5, 11, 1\}$ and $\mathcal{L}^* = \{7, \ldots, 11\}$. The figure reports raw class IoUs $q_{s,c}$ rather than the regularized scores $\widetilde{q}_{s,c}$. The best shared state is $(5, 11)$, whereas the category maxima occupy 10 of the 15 states. Seventeen of 19 categories favor another state, producing a mean gap of 3.38 points. These statistics characterize state-space heterogeneity rather than the gain of the subsequent fusion method.

We additionally visualize the complete 50×24 single-state grids on VOC20, PASCAL Context 59, and ADE20K using the same Seg4Diff OVSS procedure. Across all three datasets, high-mIoU responses remain confined to a limited layer–timestep region, but the location and extent of the high-reliability region vary with the dataset. Thus, the global localization observed on Cityscapes is not dataset-specific, although these category-averaged plots do not by themselves establish category-level heterogeneity.

Cross-dataset localization. Figure S1 shows **(a)** ADE20K, **(b)** PASCAL Context-59, and **(c)** VOC20. Each column pairs category-averaged mIoU over all 50 timesteps and 24 transformer layers (top) with its three-dimensional surface (bottom); red dashed contours mark the top 10% of states.

B Reproducibility Details

B.1 Candidate grids and common settings

Stable Diffusion 3 Medium is the common backbone for the controlled comparisons. The upper block of main Table 1 evaluates category-conditioned OVSS with DiffSegmenter Wang et al. (2025), iSeg Sun et al. (2026), and Seg4Diff Kim et al. (2025a) under the same calibration and evaluation protocol. The category-agnostic specialization is evaluated through separate DiffSeg Tian et al.

Table S2: Dataset-specific calibration and fusion configurations for Seg4Diff OVSS.

Dataset	Calibration Size	$\mathcal{T}^*$	$\mathcal{L}^*$	κ	ρ	K	α
VOC20	50	$\{18, 22, 16\}$	8–12	1	0.6	15	1
City.	30	$\{5, 11, 1\}$	7–11	40	2.5	15	1
PC59	100	$\{13, 21, 8\}$	7–11	40	1.5	15	1
ADE20K	50	$\{20, 16, 9\}$	9–13	45	1.875	5	1

Table S3: Additional OVSS metrics corresponding to the upper block of main Table 1 (%).

Method	Arch.	VOC		City.		PC59		ADE20K	
		pAcc	mAcc	pAcc	mAcc	pAcc	mAcc	pAcc	mAcc
DiffSegmenter	SD1.5	92.91	95.64	53.56	37.88	69.61	63.67	57.86	46.18
iSeg	SD2.1	86.97	88.76	61.69	35.44	77.13	68.43	63.95	48.74
DiffSegmenter	SD3	94.87	96.31	46.24	51.40	70.25	65.99	58.32	56.03
+ CSR	SD3	94.41	96.06	58.07	56.05	74.13	71.81	65.40	50.68
iSeg	SD3	93.78	95.69	55.95	37.77	74.65	67.42	64.22	51.06
+ CSR	SD3	93.41	95.29	65.26	47.47	76.15	69.19	66.98	58.11
Seg4Diff	SD3	93.45	95.54	60.27	52.37	70.79	70.51	59.06	55.45
+ CSR	SD3	93.94	96.03	66.85	55.69	74.07	69.77	63.93	58.23

(2024), DiffCut Couairon et al. (2024), and Seg4Diff Kim et al. (2025a) adapters that share scalar reliability calibration and pre-decoding response fusion but retain different admissible state families and decoder parameters.

Seg4Diff/DiffSegmenter OVSS and the category-agnostic DiffSeg/DiffCut adapters use a 50-step candidate schedule with $t \in \{1, \ldots, 50\}$. iSeg OVSS and the category-agnostic Seg4Diff adapter retain their 28-step schedules. All experiments scan the 24 SD3 transformer layers $l \in \{0, \ldots, 23\}$. Unless stated otherwise, we use

$$N_t = 3, \qquad N_l = 5, \qquad \varepsilon = 10^{-6},$$

with FP16 inference and random seed 42.

The Seg4Diff OVSS configurations used in the main experiments are:

The shrinkage coefficient κ governs the regularization of category-wise calibration statistics. With $N_t = 3$ and $N_l = 5$, $K = 15$ denotes dense fusion over the complete selected state region. Table S3 reports the pAcc and mAcc results corresponding to the FWIoU and mIoU comparisons in the upper block of main Table 1.

The OVSS search first scores all $50 \times 24 = 1{,}200$ states on the training-only calibration set. The region selector then evaluates category-coverage timestep sets with every contiguous five-layer window. Table S2 reports detailed settings.

B.2 Backend-specific frozen configurations

The calibration budget is shared across integrations on the same dataset. The frozen configurations are organized below by task: OVSS integrations consume category-indexed responses, whereas category-agnostic integrations recover partitions without semantic prompts.

OVSS integrations. Table S2 reports the Seg4Diff OVSS configuration. Table S4 reports the exact iSeg configurations used in the upper block of main Table 1, selected from the finite train-calibration candidate pools for that comparison.

Here, α is the signed exponent defined in the main paper, not a quantity constrained to $[0, 1]$. Positive values apply inverse concentration scaling, negative values apply direct scaling, and $\alpha = 0$ disables the operation. The Cityscapes region used in the upper block of main Table 1 is $\{9, 2, 20\} \times \{7, \ldots, 11\}$; it is a separate post-hoc diagnostic that expands the admissible candidate pool to the complete 28×24 state grid and does not replace the frozen Table S4 configuration.

Table S4: Dataset-specific calibrated configurations for the iSeg OVSS integration.

Dataset	States (t, l)	κ	ρ	K	α
VOC20	$\{15, 4, 11\} \times \{8, \ldots, 12\}$	5	2.0	all	0.5
City.	$\{9, 2, 20\} \times \{7, \ldots, 11\}$	40	2.5	all	-0.5
PC59	$\{13, 21, 8\} \times \{9, \ldots, 13\}$	40	1.5	all	-0.25
ADE20K	$\{15, 25, 7\} \times \{9, \ldots, 13\}$	45	0.5	all	-0.5

Table S5: Frozen settings for the DiffSeg integration. $\mathcal{T}^*$ and $\mathcal{L}^*$ define the candidate state region, while Top-K is the number of calibration-ranked states retained for fusion (`all` retains all 15 states). The KL threshold, anchor grid, and affinity are native DiffSeg decoder parameters.

Dataset	$\mathcal{T}^*$	$\mathcal{L}^*$	ρ	Top-K	KL threshold	Anchor grid	Affinity
VOC21	$\{18, 12, 15\}$	5–9	1.0	5	1.50	16	symmetric
Cityscapes	$\{15, 18, 24\}$	5–9	0.5	5	0.80	16	direct
PC59	$\{9, 12, 15\}$	7–11	2.0	3	1.75	16	direct
ADE20K	$\{12, 9, 15\}$	7–11	1.0	all	1.30	16	direct

DiffSegmenter also consumes the category-indexed I2T response. Fusion is inserted before its self-attention mask-completion stage while preserving prompt construction, category filtering, and mask-completion parameters. Its main-table results use the same calibration budgets, evaluation splits, and metric implementation as the other controlled OVSS comparisons.

Category-agnostic integrations. Table S5 reports the frozen dataset-specific state regions, fusion settings, and decoder parameters for DiffSeg.

For DiffCut, fusion is applied to the diffusion features used to construct the graph affinity. Table S6 reports the calibrated sparse state sets and frozen graph-partitioning parameters.

The symbols τ and graph α are DiffCut adapter parameters controlling recursive normalized-cut partitioning and are unrelated to the response-scaling exponent used for OVSS. They are selected jointly with the sparse state set from the finite calibration candidate list. PC59 uses the calibration-selected state weights $0.65/0.35$ instead of scalar-reliability softmax weights; this exception is part of the DiffCut adapter and not part of the common fusion core.

For the category-agnostic Seg4Diff adapter, the backend consumes I2T-derived pre-grouping responses without semantic prompts. Table S7 reports its frozen sparse state sets and decoder parameters.

These category-agnostic configurations are expressed as sparse state sets rather than OVSS Cartesian regions. The fixed Seg4Diff state grid is

$$\{6, 8, 10, 12\} \times \{7, 9, 11\},$$

which contains 12 ordered states. States are ranked by training-calibration mIoU, then pAcc, with ties resolved in favor of lower timestep and layer indices; the top dataset-specific budget is retained.

Output power and the KL threshold are backend-native adapter parameters and are reported separately from state selection. DiffSeg retains the Cartesian $(J_{\text{peak}}, J_{\text{plateau}})$ search described under backend-specific state search, whereas DiffCut uses its enumerated sparse candidate list. The resulting category-agnostic comparison measures each complete frozen adapter under its native representation and decoder.

The controlled OVSS ablation in the main paper fixes its decoder while introducing the category-conditioned components sequentially; the factorial study below separately controls the state response and decoder for all three category-agnostic backends.

Factorial ablation of multi-state fusion and decoder calibration. In the lower block of main Table 1, each baseline uses the calibration-selected single state with the original decoder, while + `Ours` uses the selected multi-state response with the calibrated decoder. Tables S8–S10 report all four factorial cells and the native single-state Seg4Diff reference.

Table S6: Calibrated state sets and frozen graph-partitioning settings for the DiffCut integration. "Selected states" lists the final calibration-ranked Top-K sparse state set retained for evaluation.

Dataset	Selected states (t, l)	ρ	τ	Graph α
VOC21	$\{(10,7),(11,8),(11,7)\}$	0.5	0.45	5
City.	$\{(9,7),(8,7),(8,8)\}$	0.1	0.50	5
PC59	$\{(12,6),(12,8)\}$	fixed	0.10	10
ADE20K	$\{(15,7),(15,8),(18,8)\}$	0.25	0.15	10

Table S7: Frozen configurations for the category-agnostic Seg4Diff integration.

Dataset	Selected states (t, l)	ρ	Output power	KL threshold
VOC21	$\{(12,7),(10,7),(12,9),(8,7)\}$	0	0.75	0.050
Cityscapes	$\{(6,7),(8,7)\}$	0	1.00	0.030
PC59	$\{(8,7),(10,7)\}$	2.0	0.65	0.035
ADE20K	$\{(10,9),(6,9),(8,9),(12,9)\}$	1.0	0.65	0.015

In Table S10, the first two rows of each dataset measure state selection without changing the native decoder. The remaining rows form the same 2×2 design used in Tables S8–S10: the calibration-selected single state versus the selected multi-state set, crossed with the original versus calibrated decoder.

The dataset-level cells expose substantial interactions that a cross-dataset mean would conceal. Main Table 3 reports the VOC21 and Cityscapes portions of the `Calibrated single/Calibrated` and `Multi-state/Calibrated` comparison; the corresponding PC59 and ADE20K remain in Tables S8–S10. The response type changes while the decoder setting remains fixed. With the original decoder, Seg4Diff lowers ADE20K mAcc/mIoU by $2.37/0.38$ points even as other endpoints improve. These results motivate reporting complete adapter performance separately from the fixed-decoder fusion effect and selecting finite state/decoder pairs only on calibration data.

B.3 Calibration and evaluation protocol

We evaluate on PASCAL VOC, Cityscapes, PASCAL Context, and ADE20K. OVSS uses VOC20 with 20 foreground categories, while category-agnostic segmentation uses VOC21 with background retained; Cityscapes, PC59, and ADE20K contain 19, 59, and 150 categories, respectively.

Before diffusion feature extraction, each image is decoded in RGB format and resized in the dataset space while preserving its aspect ratio. The standard OVSS and DiffCut pipelines set the shorter side to 640 pixels and cap the longer side at 2,560 pixels; the official Seg4Diff category-agnostic adapter uses 1,024 and 4,096 pixels, respectively, followed by its deterministic square center crop. The resulting uint8 tensor is converted to floating point in $[0, 1]$ and bilinearly resampled to 1024×1024 immediately before entering the frozen diffusion backbone. No ImageNet/CLIP mean–standard-deviation normalization, random crop, color augmentation, or horizontal flip is applied to the diffusion input at evaluation time. Ground-truth labels are never passed to the diffusion backbone: they follow exactly the same backend-native geometry using nearest-neighbor interpolation, and the predicted responses are restored to the evaluated annotation geometry before metric computation. Each baseline/`+ Ours` pair receives exactly the same preprocessed image tensor and valid-pixel mask.

Reference implementations in this area do not use a single evaluation path: they differ in whether assignment is restricted to present class rows, how unmatched proposals are handled, and whether a center crop excludes peripheral pixels. We therefore do not use cross-paper absolute differences to support our method claim. The original-architecture rows in the lower block of main-paper Table 1 and Table S11 retain their source protocols only for context. Every controlled SD3 baseline/`+ Ours` comparison instead uses the same full-class, per-image one-to-one assignment and the same metric implementation. Backend-native image geometry is held fixed within each pair, so a gain cannot arise from evaluating fewer boundary pixels or changing the crop for `+ Ours`.

The calibration subsets are sampled once with seed 42 exclusively from each dataset's official training split, using budgets of 50, 30, 100, and 50 images for VOC, Cityscapes, PC59, and ADE20K, respectively. Selection is category-balanced: it first covers missing categories and then favors underrepresented categories by inverse-frequency balancing. For PASCAL Context, the training pool

Table S8: Complete factorial ablation for category-agnostic DiffSeg with SD3 (%).

Dataset	State response	Decoder	pAcc	FWIoU	mAcc	mIoU
VOC21	Calibrated single	Original	58.10	51.08	67.65	38.86
	Calibrated single	Calibrated	59.58	52.08	69.71	39.64
	Multi-state	Original	61.32	53.42	70.91	40.12
	Multi-state	**Calibrated**	**62.06**	**53.87**	**71.56**	**40.41**
Cityscapes	Calibrated single	Original	60.71	49.90	44.72	24.30
	Calibrated single	Calibrated	60.71	49.90	44.72	24.30
	Multi-state	Original	63.48	52.49	44.49	25.42
	Multi-state	**Calibrated**	**63.48**	**52.49**	**44.49**	**25.42**
PC59	Calibrated single	Original	69.09	57.28	65.74	46.07
	Calibrated single	Calibrated	71.11	58.34	65.88	47.15
	Multi-state	Original	69.55	57.48	65.72	46.28
	Multi-state	**Calibrated**	**71.39**	**58.36**	**65.41**	**47.13**
ADE20K	Calibrated single	Original	54.53	38.41	60.17	39.01
	Calibrated single	Calibrated	54.53	38.41	60.17	39.01
	Multi-state	Original	58.98	42.67	61.95	39.87
	Multi-state	**Calibrated**	**58.98**	**42.67**	**61.95**	**39.87**

Table S9: Complete factorial ablation for category-agnostic DiffCut with SD3 (%).

Dataset	State response	Decoder	pAcc	FWIoU	mAcc	mIoU
VOC21	Calibrated single	Original	83.50	72.90	63.29	48.69
	Calibrated single	Calibrated	83.59	73.32	66.44	49.90
	Multi-state	Original	83.60	72.98	62.78	48.55
	Multi-state	**Calibrated**	**83.74**	**73.56**	**67.38**	**50.50**
Cityscapes	Calibrated single	Original	67.40	55.38	39.68	24.92
	Calibrated single	Calibrated	69.25	59.30	44.48	27.42
	Multi-state	Original	68.26	56.14	38.55	25.47
	Multi-state	**Calibrated**	**71.19**	**60.56**	**42.93**	**27.74**
PC59	Calibrated single	Original	72.99	59.74	63.76	47.55
	Calibrated single	Calibrated	72.99	59.74	63.76	47.55
	Multi-state	Original	73.08	59.98	64.71	48.00
	Multi-state	**Calibrated**	**73.08**	**59.98**	**64.71**	**48.00**
ADE20K	Calibrated single	Original	68.21	53.02	54.74	39.12
	Calibrated single	Calibrated	66.22	50.90	56.39	39.64
	Multi-state	Original	69.01	53.96	55.95	40.28
	Multi-state	**Calibrated**	**67.04**	**51.79**	**57.77**	**40.52**

is formed from the official train portion of the trainval annotations after excluding every identifier in the validation list. The selected calibration images cover all 20 VOC foreground categories, all 19 Cityscapes categories, all 59 PASCAL Context categories, and all 150 ADE20K categories; VOC21 additionally retains background.

The calibration and benchmark evaluation sets are strictly disjoint. We verify zero image-identifier overlap between every calibration manifest and the complete evaluation split. Consequently, no benchmark validation/test image or annotation contributes to candidate-state scoring, region construction, reliability weighting, response-scaling selection, or any other hyperparameter decision. This protocol prevents information leakage from the benchmark evaluation set into model selection. Calibration estimates one dataset-level configuration for each response type and backend adapter; each configuration is frozen before evaluation and shared across all evaluation images of that experiment. The reported benchmark numbers are then computed once on the complete held-out evaluation split; evaluation labels are used only for final metric computation.

Different benchmark implementations name their held-out evaluation partitions either *validation* or *test*. Throughout the paper, *benchmark evaluation split* denotes the standard partition used for comparison by the corresponding segmentation literature: 1,449 images for VOC, 500 for Cityscapes,

Table S10: Complete factorial ablation for category-agnostic Seg4Diff with SD3 (%).

Dataset	State response	Decoder	pAcc	FWIoU	mAcc	mIoU
VOC21	Native single	Original	53.17	45.56	63.46	35.39
	Calibrated single	Original	57.66	49.13	68.61	38.99
	Calibrated single	Calibrated	67.81	57.82	73.56	44.13
	Multi-state	Original	63.25	54.43	73.63	43.57
	Multi-state	**Calibrated**	**71.40**	**60.77**	**74.07**	**45.87**
Cityscapes	Native single	Original	63.94	53.88	44.91	25.61
	Calibrated single	Original	71.28	61.18	44.19	27.50
	Calibrated single	Calibrated	69.63	60.29	**46.50**	**27.70**
	Multi-state	Original	**76.82**	65.41	38.60	27.08
	Multi-state	**Calibrated**	76.54	**65.72**	41.29	27.57
PC59	Native single	Original	63.20	50.83	61.14	41.34
	Calibrated single	Original	65.82	53.13	61.50	42.76
	Calibrated single	Calibrated	70.22	57.02	64.05	45.95
	Multi-state	Original	65.92	53.24	61.52	42.81
	Multi-state	**Calibrated**	**70.26**	**57.05**	**64.12**	**46.03**
ADE20K	Native single	Original	60.12	44.04	53.94	35.88
	Calibrated single	Original	60.37	44.36	53.30	35.14
	Calibrated single	Calibrated	56.16	40.44	55.96	35.74
	Multi-state	Original	**61.30**	**45.19**	50.93	34.76
	Multi-state	**Calibrated**	58.52	42.65	**57.18**	**36.90**

Table S11: Pixel accuracy (pAcc) and mean accuracy (mAcc) for the original-architecture category-agnostic results (%).

Source method	Arch.	VOC21		City.		PC59		ADE	
		pAcc	mAcc	pAcc	mAcc	pAcc	mAcc	pAcc	mAcc
DiffSeg	SD1.5	64.91	72.68	75.98	35.35	72.84	65.46	67.65	49.42
DiffCut	SSD-1B	79.74	84.54	78.85	36.79	78.21	71.86	62.36	59.98

5,104 for PC59, and 2,000 for ADE20K. Regardless of naming, this split is separate from the official training split that supplies calibration data.

OVSS follows each dataset's official label space and ignore mask. Predictions are restored to annotation resolution before confusion-matrix evaluation. Its calibration score uses overall mIoU for q_s and category IoU for $q_{s,c}$. For category-agnostic segmentation, geometric transforms are shared between images and labels, using bilinear interpolation for images and nearest-neighbor interpolation for labels.

Permutation-invariant partition evaluation. Let P_i be the predicted partition of image i, with local proposal IDs $k \in \{1, \dots, K_i\}$. We form C dataset-class rows and $\widetilde{K}_i = \max(C, K_i)$ proposal columns, zero-padding columns when $K_i < C$, and define

$$M_i(c, k) = |\{p : y_i(p) = c,\ P_i(p) = k\}| \tag{20}$$

as the non-padded entries of its full overlap matrix. We compute one injective assignment $\pi_i : \{1, \dots, C\} \to \{1, \dots, \widetilde{K}_i\}$,

$$\pi_i^* = \arg\max_{\pi_i} \sum_{c=1}^{C} M_i(c, \pi_i(c)) \tag{21}$$

for the complete image. All C dataset-class rows participate, including classes absent from that image, and unmatched or extra proposals remain errors. After assignment, per-class TP, FP, and FN are accumulated over the complete evaluation split before pAcc, FWIoU, mAcc, and mIoU are computed; we do not average independently optimized per-image IoUs. Ground truth therefore

Table S12: GoCA reproduction under its source protocol, compared with our Seg4Diff-based aggregation strategy (%).

Method	Arch.	Aggregation	VOC20. mIoU	ADE20K mIoU
GoCA	Flux.1-dev	Layer weighting + token rescaling	71.23	29.32
Ours	Flux.1-dev	Seg4Diff	85.08	32.65

resolves only the arbitrary integer naming of an already produced partition during offline calibration or evaluation.

Hungarian matching is applied to the integer overlap matrix with dataset classes and proposals ordered by ascending index. This deterministic ordering resolves exact ties consistently. Because absent-class rows can admit multiple optimal assignments, conclusions sensitive to such ties require separate analysis.

A single dataset-level assignment is not well-defined for these outputs because proposal ID k is reinitialized for every image, K_i can vary, and the backends expose no persistent cross-image cluster prototype. Treating, for example, cluster `2` from two images as the same object would impose a correspondence absent from the predictor. Constructing such a correspondence would require an additional global clustering or semantic-labeling module and would evaluate cross-image category discovery rather than the category-agnostic partition task studied here. Conversely, for methods that do expose globally persistent cluster identities, dataset-level matching is the appropriate protocol; their numbers should not be mixed with per-image partition scores without relabeling the task. The lower block of Table 1 therefore applies the per-image protocol uniformly to all controlled SD3 methods and interprets it explicitly as partition quality, not cross-image semantic-label consistency.

Each baseline/`+ Ours` pair shares the held-out evaluation split, preprocessing, valid pixels, and sampled noise; the same noise tensor is also shared across selected states for each image. OVSS pairs keep the downstream decoder fixed. For category-agnostic pairs, only the reported adapter parameters may differ; these are selected on the training-only calibration set and frozen before evaluation. Across methods, we use the same dataset definitions, label handling, matching rule, and metric computation. The reported metrics are pixel accuracy, frequency-weighted IoU, mean class accuracy, and mean IoU; all values and gains are expressed in percentage points.

B.4 Results under source protocols

Tables S11 and S12 report reference results obtained with each method's source architecture and evaluation protocol. They provide context only; all controlled baseline/`+ Ours` claims use the unified SD3 protocol described above.

Original DiffSeg and DiffCut. The lower block of Table 1 reports FWIoU and mIoU for DiffSeg-SD1.5 and DiffCut-SSD-1B; Table S11 provides the complementary pAcc and mAcc.

DiffSeg is evaluated on the complete splits using SD1.5, a 512×512 center crop, timestep 300, a 16×16 anchor grid, no refinement, and per-image Hungarian matching. Cityscapes follows its 27-class evaluation with KL thresholds $(0.9, 0.9, 0.9)$; the other datasets use $(1.1, 1.1, 1.1)$. DiffCut retains its SSD-1B source configuration.

We run GoCA Meng et al. (2026) official implementation with Flux.1-dev Black Forest Labs (2024), timestep 150, 1024×1024 input, and its low-memory layer configuration on Pascal VOC and ADE20K validation images. The corresponding source-protocol results are reported in Table S12.

GoCA uses validation annotations to form present-class prompt candidates and selects the best background offset on the evaluation split. Our controlled protocol instead queries the benchmark vocabulary and freezes all reusable choices using a disjoint training calibration subset. Because the backbones and evaluation paths differ, Table S12 is not used for direct gains or significance claims.

Table S13: Full-split OVSS results of Seg4Diff with and without our method using FLUX.1-dev.

Dataset	Metric	Seg4Diff	+ Ours	Gain
VOC20	mIoU	83.95	**85.08**	+1.13
VOC20	mACC	91.53	**92.21**	+0.67
Cityscapes	mIoU	19.84	**21.90**	+2.05
Cityscapes	mACC	46.42	**49.14**	+2.72
ADE20K	mIoU	24.86	**32.65**	+7.79
ADE20K	mACC	47.57	**53.04**	+5.47

Table S14: Full-split OVSS results of Seg4Diff with and without our method using SD3.5.

Dataset	Metric	Seg4Diff	+ Ours	Gain
VOC20	mIoU	81.58	**83.89**	+2.30
VOC20	fwIoU	80.26	**82.09**	+1.83
Cityscapes	pACC	57.40	**71.65**	+14.26
Cityscapes	fwIoU	47.48	**61.71**	+14.23
ADE20K	pACC	55.30	**62.73**	+7.43
ADE20K	fwIoU	40.81	**47.88**	+7.07
PC59	mIoU	40.25	**43.52**	+3.27
PC59	pACC	64.11	**67.35**	+3.24

C More Results

C.1 Transfer to alternative diffusion backbones

We evaluate procedure portability beyond the SD3 Medium backbone used in the main experiments. These experiments replace the frozen diffusion backbone in the Seg4Diff OVSS pipeline and rerun the same train-only calibration-guided selection and fusion procedure. Dataset- and backbone-specific configurations are selected on the disjoint training-split calibration sets, frozen, and then evaluated on the complete benchmark evaluation splits. They therefore test whether the procedure can be reapplied after recalibration, not whether an SD3-selected configuration transfers unchanged.

FLUX.1-dev. Black Forest Labs (2024) We first transfer the pipeline to FLUX.1-dev and freeze each dataset-specific configuration before evaluation on 1,449 VOC20, 500 Cityscapes, and 2,000 ADE20K images. Table S13 reports the full-split results after backbone-specific recalibration.

All six reported FLUX.1-dev metrics improve. The gains are particularly pronounced on ADE20K, where mIoU and mACC increase by 7.79 and 5.47 points, respectively.

Stable Diffusion 3.5. Esser et al. (2024) We next replace the main SD3 Medium backbone with SD3.5 and evaluate VOC20, Cityscapes, ADE20K, and PC59 under the same frozen-backbone protocol. Table S14 reports the corresponding full-split results after backbone-specific recalibration.

SD3.5 yields improvements on all eight displayed dataset–metric pairs. The largest gains occur on Cityscapes, with improvements of 14.26 pACC and 14.23 fwIoU, followed by gains of 7.43 pACC and 7.07 fwIoU on ADE20K. Improvements on VOC20 and PC59 further show that the procedure can be reapplied beyond dense scene-level benchmarks. Together with the FLUX.1-dev results, these findings support procedure portability after backbone-specific recalibration; they do not establish configuration transfer from SD3 to another backbone.

C.2 Computational characteristics

The calibration scan is amortized over subsequent inference images, with cost linear in the number of candidate states and calibration samples. The online profile below measures the reusable configuration after state discovery.

At inference, all N_l selected layers for one timestep are extracted during the same diffusion forward. Therefore, the number of diffusion forwards is governed by N_t, not by $N_t N_l$. The structured OVSS

Table S15: Additional asymptotic cost of response fusion and normalization.

Task	Fusion complexity	Normalization complexity
OVSS	$\mathcal{O}(N_t N_l\, C\lvert\mathcal{P}\rvert)$	$\mathcal{O}(C\lvert\mathcal{P}\rvert)$
Token-affinity grouping	$\mathcal{O}(N_t N_l\, \lvert\mathcal{P}\rvert^2)$	Backend-dependent affinity normalization

Table S16: Seed-wise variation and paired significance on the complete Cityscapes evaluation split.

Metric	Seed 40	Seed 41	Seed 42	Seed 43	Seed 44	Mean $\pm$ std	95% CI of mean gain	Paired p
mIoU	+4.69	+4.64	+5.03	+5.01	+5.13	**+4.90 $\pm$ 0.22**	$[+4.63, +5.17]$	9.75×10^{-7}
fwIoU	+6.58	+6.78	+6.76	+6.84	+6.89	**+6.77 $\pm$ 0.12**	$[+6.62, +6.92]$	2.29×10^{-8}
mACC	+2.79	+2.88	+3.32	+2.97	+3.56	**+3.10 $\pm$ 0.33**	$[+2.70, +3.51]$	2.91×10^{-5}
pACC	+6.43	+6.60	+6.59	+6.71	+6.73	**+6.61 $\pm$ 0.12**	$[+6.46, +6.76]$	2.77×10^{-8}

setting with three timesteps and five layers produces 15 responses from three forwards, while the efficient one-timestep variant produces five responses from one forward.

Table S15 summarizes the additional arithmetic cost of response aggregation, excluding backend-specific decoding.

Table S18 reports the complete latency and peak-memory comparison. Ours-1T-5L retains the five layers associated with the strongest calibrated timestep and obtains all responses in one backbone forward. Ours-3T-5L is the full 15-state OVSS configuration used for the main accuracy results.

All measurements use FP16, batch size 1, and an NVIDIA A100. Each value is averaged over 50 measured images after 5 warm-up images; timing is CUDA-synchronized and excludes data loading and metric evaluation. The full configuration requires approximately one forward per selected timestep, explaining its 3.61–$3.64\times$ latency. In contrast, peak memory grows by only 3.10–3.88% because layer responses are aggregated at the feature level rather than retained as separate backbone replicas.

The full-split segmentation accuracy of the profiled Ours-1T-5L variant is reported in Table S19. This variant retains the strongest calibrated timestep and all five associated layers, together with category-conditioned weighting, shrinkage, and Peak-RMS scaling. These results are measured directly on the complete evaluation splits rather than inherited from the three-timestep configuration.

Ours-1T-5L therefore preserves a practical accuracy–efficiency operating point while requiring only one SD3 forward. Relative to the corresponding single-state baselines evaluated under the same protocol, it improves mIoU by 1.03, 4.25, 1.08, and 1.84 points on VOC20, Cityscapes, PC59, and ADE20K, respectively, for a macro-average gain of 2.05 points.

C.3 Randomness and statistical significance

We evaluate sensitivity to diffusion randomness using five paired runs with seeds $\{40, 41, 42, 43, 44\}$. For each seed, Seg4Diff and `+ Ours` receive the same sampled noise, preprocessing, complete evaluation split, and frozen train-calibrated configuration. We therefore analyze the paired improvement, rather than either method's absolute performance. All values below are percentage-point gains of `+ Ours` over Seg4Diff. This experiment fixes the calibration subset and varies only diffusion noise. It is a secondary implementation check, not evidence about calibration-set selection uncertainty; the independent calibration-subset experiment reported next addresses the latter.

The frozen configurations are those in Table S2. Seed 42 corresponds to the main results; the other runs change only the shared diffusion-noise seed. Tables S16 and S17 report the paired gains for Cityscapes and ADE20K, respectively.

The improvement is positive under every tested seed and for every reported metric. In particular, the mIoU gain varies by only 0.49 points from its minimum to maximum, with a standard deviation of 0.22 points. With only five noise seeds, the confidence intervals and paired tests are descriptive checks rather than the paper's main statistical evidence; they do not turn seeds into independent image samples and are not used to rank methods.

Table S17: Seed-wise variation and paired significance on the complete ADE20K evaluation split.

Metric	Seed 40	Seed 41	Seed 42	Seed 43	Seed 44	Mean $\pm$ std	95% CI of mean gain	Paired p
mIoU	+3.52	+3.22	+3.47	+3.71	+3.32	**+3.45 $\pm$ 0.19**	$[+3.21, +3.68]$	2.12×10^{-6}
fwIoU	+5.02	+5.04	+5.13	+5.00	+5.03	**+5.04 $\pm$ 0.05**	$[+4.98, +5.11]$	2.16×10^{-9}
mACC	+2.92	+2.55	+2.79	+3.44	+2.81	**+2.90 $\pm$ 0.33**	$[+2.49, +3.31]$	3.86×10^{-5}
pACC	+4.73	+4.76	+4.86	+4.75	+4.77	**+4.77 $\pm$ 0.05**	$[+4.71, +4.84]$	3.26×10^{-9}

Table S18: Online latency and peak-memory comparison.

Dataset	Variant	States	SD3 forward	Latency (ms/image)	Peak memory (MiB)
VOC20	Baseline	1	1	203.08	8350.17
	1T-5L	5	1	249.65	8625.78
	3T-5L	15	3	733.48	8634.61
Cityscapes	Baseline	1	1	203.68	8348.73
	1T-5L	5	1	252.00	8658.47
	3T-5L	15	3	736.26	8673.00
PC59	Baseline	1	1	201.87	8379.31
	1T-5L	5	1	247.69	8620.83
	3T-5L	15	3	735.24	8639.08
ADE20K	Baseline	1	1	204.15	8350.24
	1T-5L	5	1	248.98	8620.05
	3T-5L	15	3	737.46	8643.45

ADE20K shows the same conclusion under fixed calibration: every metric improves for every tested noise seed, and the mIoU gain has a standard deviation of 0.19 points. Calibration-subset resampling is a separate, more demanding robustness test because the selected states must be re-estimated for each category-balanced training subset.

C.4 Calibration-subset resampling

We use five content-verified unique category-balanced training manifests per dataset: Cityscapes seeds $\{42, 45, 46, 48, 53\}$ with 30 images and ADE20K seeds $\{40, 42, 43, 44, 45\}$ with 50 images. Duplicate manifests from the earlier seed set are excluded. For each manifest, all $50 \times 24 = 1{,}200$ states are rescored and the three-timestep, five-layer region is reselected. The fusion hyperparameters $(\kappa, \rho, K, \alpha)$ remain fixed at their Table S2 values; this experiment measures calibration-image and state-region uncertainty, not uncertainty from a nested fusion-hyperparameter search. Table S20 reports every reselection result, while Table S21 summarizes the mean, variation, and overlap statistics across the five manifests.

The mIoU gain is positive in all ten runs, ranging from +3.69 to +5.31 on Cityscapes and +2.89 to +3.47 on ADE20K; auxiliary metrics show greater subset sensitivity.

C.5 Shared versus backend-specific OVSS regions

After freezing the finite-pool iSeg configuration in Table S4, we run a separate region-transfer diagnostic on the same 30-image Cityscapes training calibration manifest. This diagnostic expands the admissible candidate family to all $28 \times 24 = 672$ iSeg states and applies the anchored category-coverage selector, yielding $\{11, 6, 8\} \times \{10, \ldots, 14\}$. We compare this full-grid diagnostic region with the Seg4Diff train-calibrated region $\{5, 11, 1\} \times \{7, \ldots, 11\}$ while holding iSeg's shrinkage, fusion strength, top-K, response scaling, preprocessing, noise seed, and decoder fixed, then evaluate both candidates jointly on all 500 Cityscapes validation images. Table S22 compares these two calibrated regions with the native single-state baseline on the complete Cityscapes validation split.

The decoder-specific scan confirms that iSeg and Seg4Diff rank the state space differently. Under fixed non-region settings, the transferred Seg4Diff region exceeds the iSeg-specific region by 1.82 pAcc, 1.67 FWIoU, 0.34 mAcc, and 1.09 mIoU points. This result supports transfer for the tested I2T pair but does not establish that a shared region is generally optimal.

Table S19: Full-split OVSS performance of the efficient 1T-5L variant (%).

Dataset	pAcc	FWIoU	mAcc	mIoU
VOC20	93.99	88.94	96.18	90.57
Cityscapes	60.29	48.99	51.95	31.08
PC59	71.09	56.83	68.68	49.81
ADE20K	61.22	45.86	52.48	34.40
Macro average	**71.65**	**60.15**	**67.32**	**51.47**

Table S20: Calibration-subset state reselection and full-split gains (%). Each row reruns state scoring and region selection, then evaluates the frozen configuration on the complete validation split.

Dataset	Seed	Selected timesteps	Layers	pAcc gain	FWIoU gain	mAcc gain	mIoU gain
Cityscapes	42	$\{5, 11, 1\}$	7–11	+6.59	+6.76	+3.32	+5.03
	45	$\{5, 12, 21\}$	9–13	+1.81	+2.41	+2.37	+3.73
	46	$\{5, 11, 1\}$	7–11	+7.13	+7.08	+3.39	+5.25
	48	$\{5, 11, 1\}$	7–11	+7.21	+7.22	+3.43	+5.31
	53	$\{5, 12, 21\}$	9–13	+1.76	+2.41	+2.54	+3.69
ADE20K	40	$\{14, 21, 10\}$	9–13	+4.86	+4.56	-0.34	+2.90
	42	$\{20, 16, 9\}$	9–13	+4.86	+5.13	+2.79	+3.47
	43	$\{22, 13, 10\}$	9–13	+3.89	+4.08	+3.47	+3.26
	44	$\{13, 21, 8\}$	9–13	+5.67	+5.35	-0.17	+3.24
	45	$\{14, 21, 10\}$	9–13	+4.73	+4.41	-0.56	+2.89

D Method Boundaries and Failure Cases

D.1 Additional qualitative comparisons

Figure S2 extends the main-paper visualization to three category-agnostic segmentation backends and two datasets. The examples are selected by per-image mIoU gain from the same deterministic 40-image evaluation prefix used for qualitative screening. They are explicitly best-case illustrations and are not representative samples or evidence of aggregate performance. For visualization only, permutation-dependent cluster IDs are aligned to the ground truth by per-image Hungarian matching, without changing the predicted regions.

Additional qualitative comparisons for category-agnostic segmentation. Columns show the input, ground truth, original backend prediction, and the same backend augmented with our method. Rows compare DiffSeg on VOC21, DiffCut on VOC21, and Seg4Diff on Cityscapes; Δ denotes the per-image mIoU gain in percentage points.

The three rows expose complementary effects of the calibrated adapters. For DiffSeg, the fused affinity improves the spatial support of adjacent people in a crowded object-centric scene. For DiffCut, the calibrated state set and graph parameters better recover riders and thin bicycle structures instead of absorbing them into coarse appearance-based partitions. In the Cityscapes example, the Seg4Diff baseline fragments the road and assigns scattered labels across the scene, whereas the calibrated adapter restores a more coherent street layout while retaining the central pedestrian and vehicle. Together, these examples show that the common response-level interface described above can support different backend-specific adapters; they do not imply an identical decoder-independent correction.

D.2 Failure analysis and scope

The method assumes that the candidate diffusion states contain complementary task-relevant evidence and that this evidence is reasonably represented by the calibration set. Its benefit may be limited when all candidate states produce nearly identical responses, when calibration categories are

Table S21: Mean ± sample standard deviation over five unique calibration manifests (%).

Dataset	pAcc gain	FWIoU gain	mAcc gain	mIoU gain	Manifest image Jaccard	State-set Jaccard
Cityscapes	4.90 ± 2.85	5.18 ± 2.53	3.01 ± 0.51	4.60 ± 0.82	0.89 mean (0.82 min)	0.47 mean (0.11 min)
ADE20K	4.80 ± 0.63	4.71 ± 0.52	1.04 ± 1.93	3.15 ± 0.25	0.39 mean (0.30 min)	0.20 mean (0.00 min)

Table S22: Controlled full-grid region-transfer diagnostic for iSeg on Cityscapes (%). Region selection uses the same disjoint training calibration subset; evaluation uses the complete validation split. This expanded-grid diagnostic holds the non-region iSeg settings fixed and is not a reselection of the finite-pool Table S4 configuration used in the upper block of main Table 1.

iSeg response region	pAcc	FWIoU	mAcc	mIoU
Native single state $(8, 9)$	55.95	48.06	37.77	21.10
iSeg-specific region	64.51	56.65	46.84	25.53
Shared Seg4Diff region	**66.34**	**58.32**	**47.18**	**26.62**

severely underrepresented, or when the downstream backend is insensitive to changes in its input response.

Figure S2 illustrates that local improvements do not guarantee a better complete partition. DiffCut recursively partitions a global affinity graph; a calibration-induced change that strengthens a missing foreground region can also alter downstream cuts over large background areas. This behavior is most visible when several states provide conflicting affinities or when the graph decoder amplifies a modest response change into a different partition topology.

The approach also introduces a calibration/inference trade-off. Increasing N_t expands semantic coverage but requires additional diffusion forwards; increasing N_l is less expensive when multiple layer responses can be cached in a single forward, but an excessively wide window may introduce redundant or weak states. The structured 3-timestep, 5-layer OVSS region provides a compact operating point rather than claiming that a single state budget is universally optimal.

Finally, calibration reliability is task- and representation-specific. A configuration selected for an I2T response should not be transferred blindly to an I2I affinity or a different diffusion backbone. When the response construction, backbone, or downstream decoder changes materially, reliability estimation and state selection should be repeated, while the model itself remains frozen.

Figure S3 provides the gain-selected extended gallery used during qualitative screening. It covers OVSS and category-agnostic segmentation across object-centric, indoor, and urban-scene images, and includes Seg4Diff, DiffSeg, and DiffCut integrations. Because selection uses ground-truth gain, the gallery visualizes possible positive effects but does not estimate their frequency. Each group compares the input, ground truth, original backend prediction, and the corresponding `+ Ours` result. The gallery includes OVSS examples from Cityscapes, VOC20, PC59, and ADE20K, together with category-agnostic results for DiffSeg, DiffCut, and Seg4Diff on VOC21 and Cityscapes.

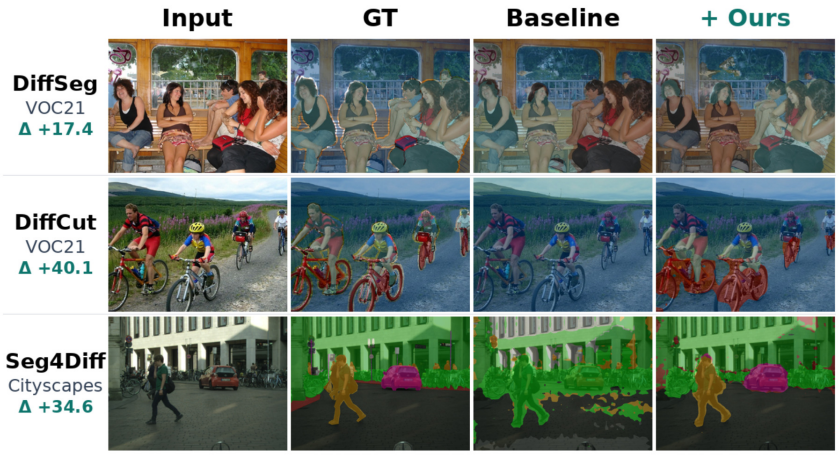

Figure S2: Qualitative comparisons across category-agnostic segmentation backends.

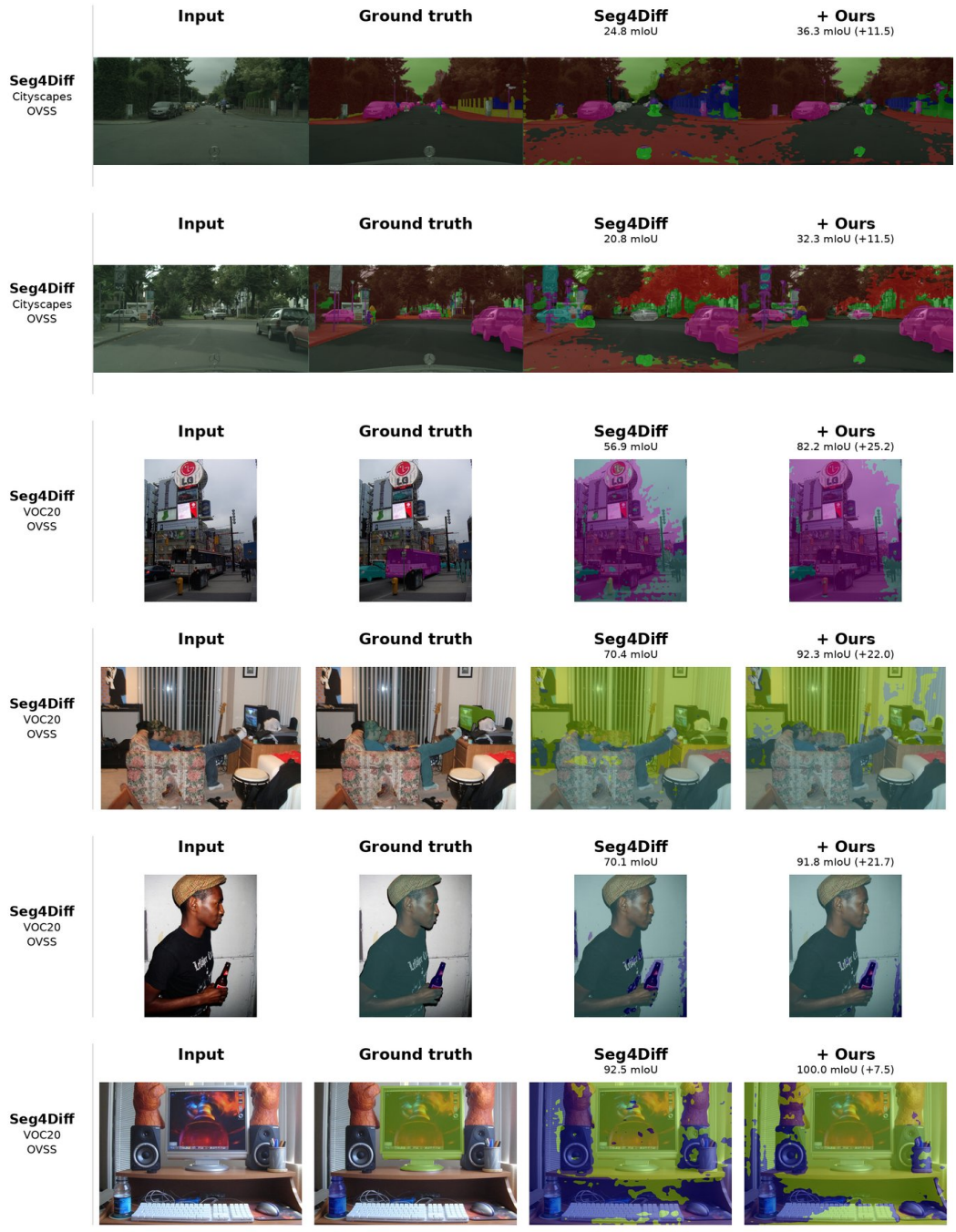

Figure S3: Extended qualitative gallery across tasks, methods, and datasets.

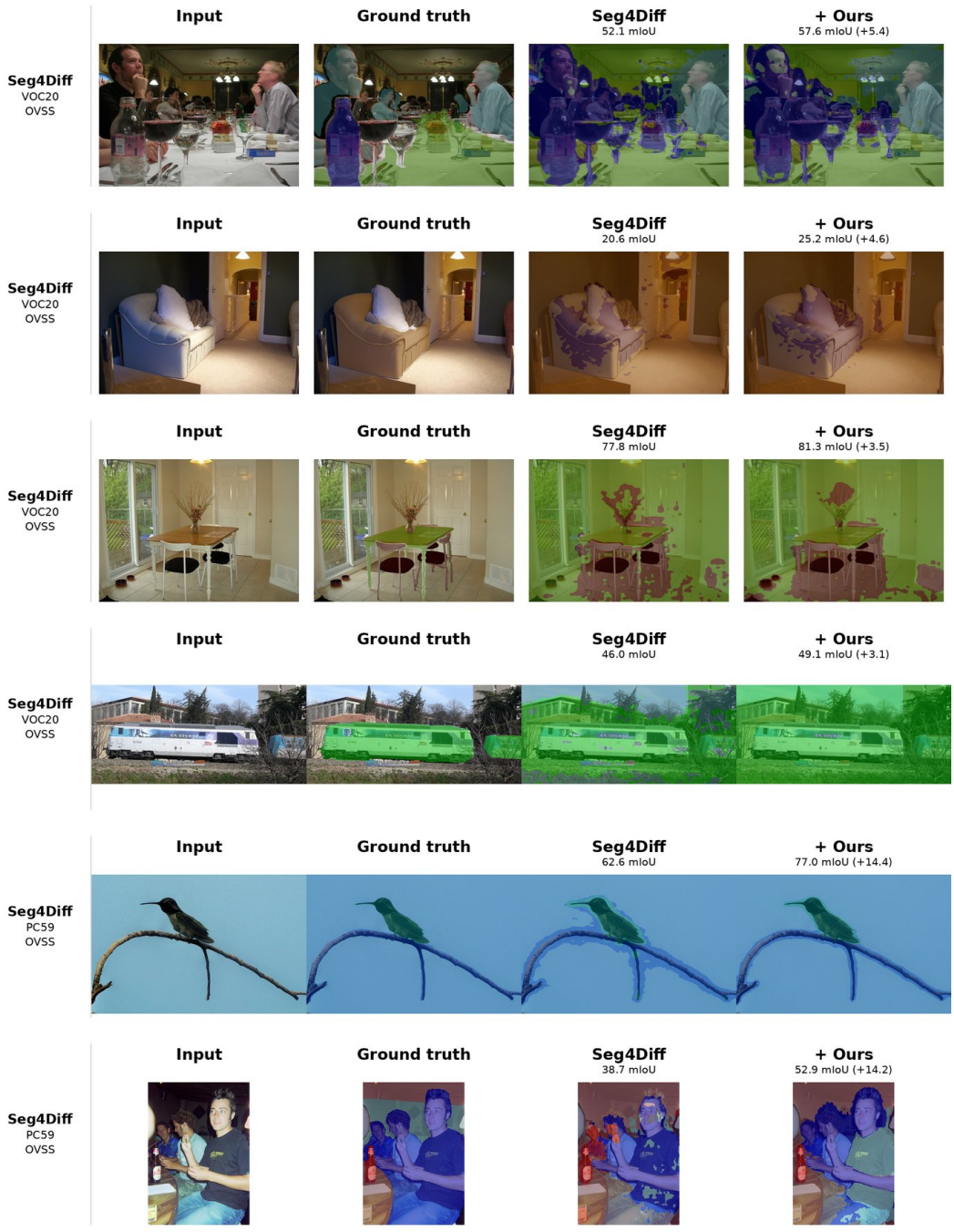

Figure S4: Extended qualitative gallery across tasks, methods, and datasets.

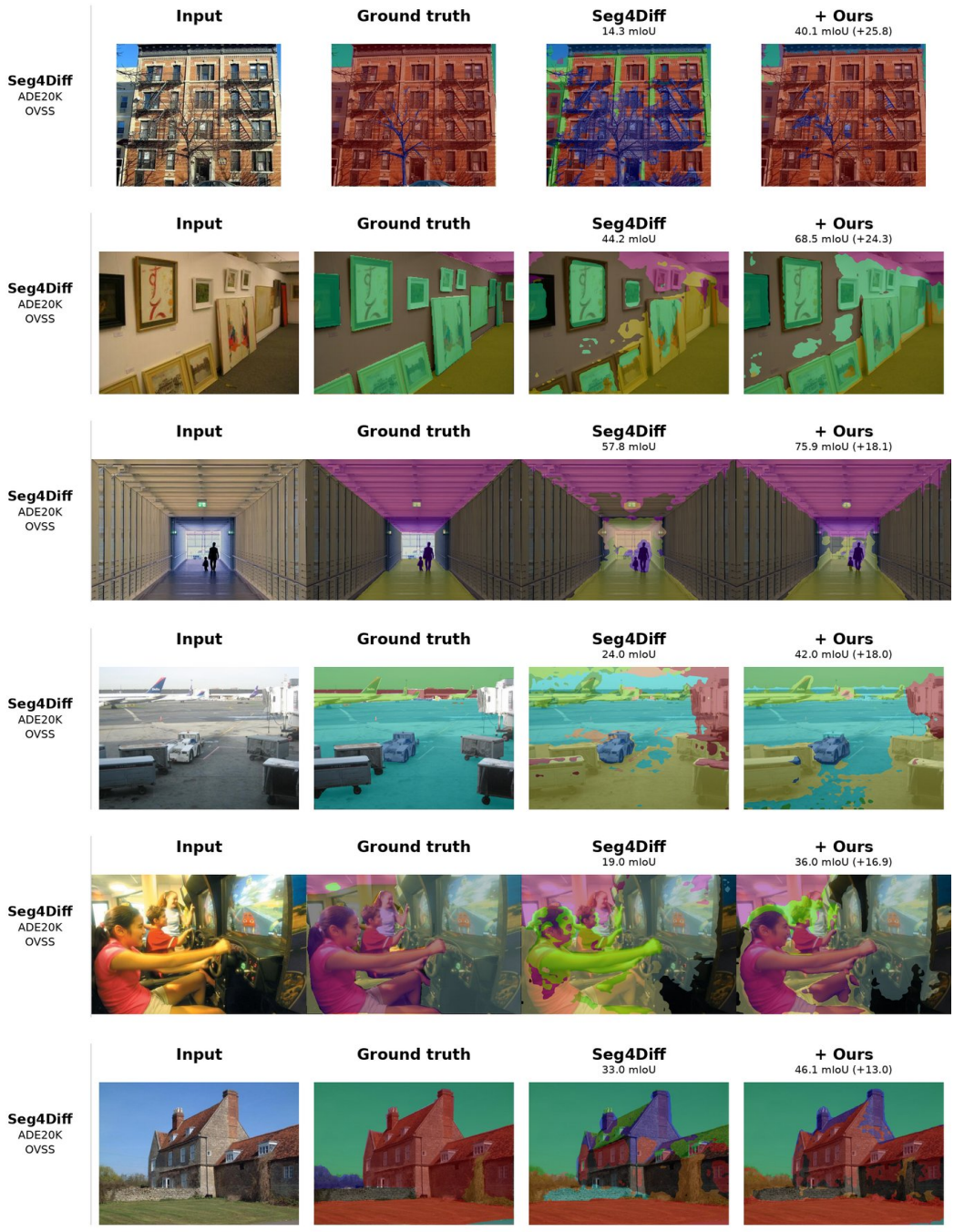

Figure S5: Extended qualitative gallery across tasks, methods, and datasets.

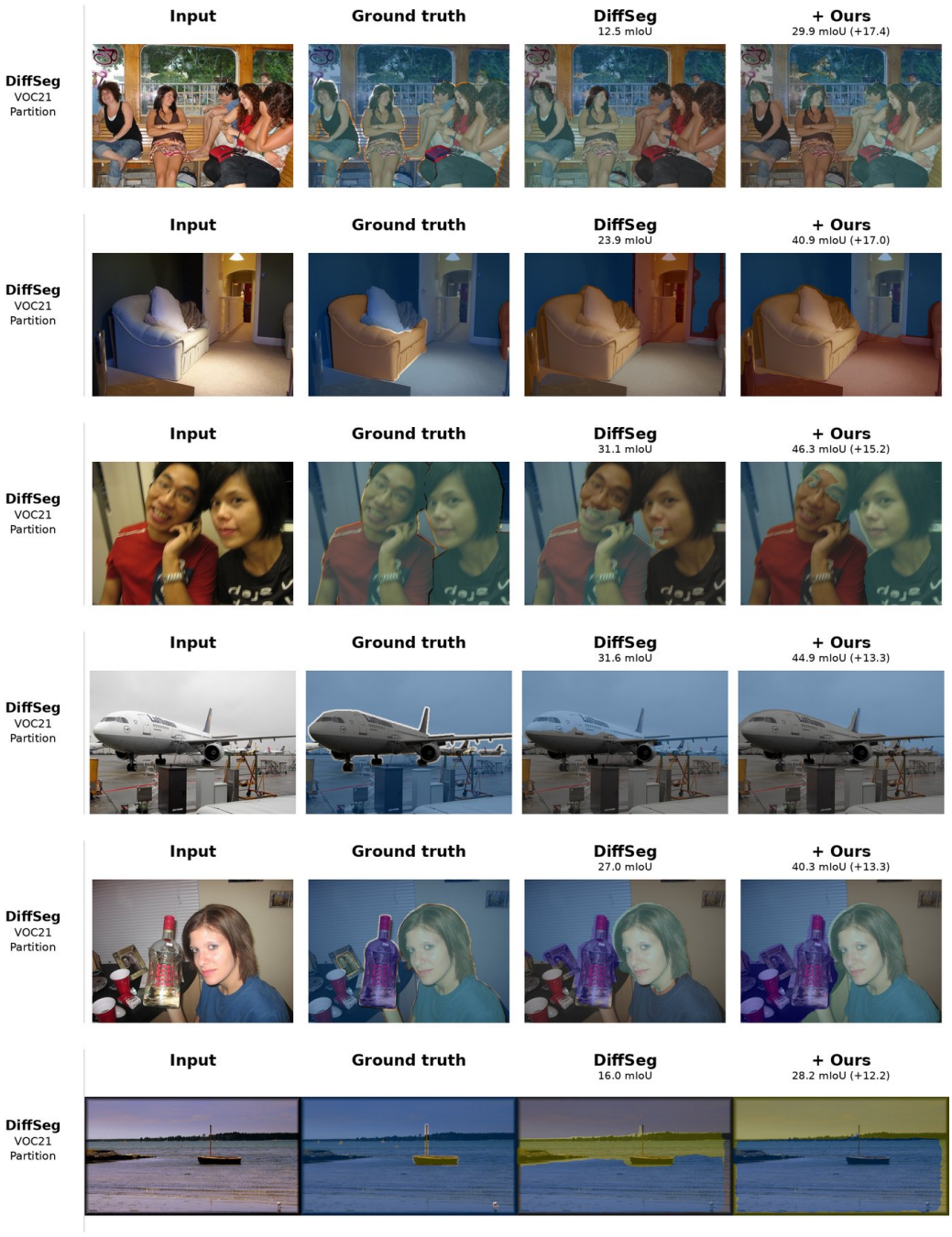

Figure S6: Extended qualitative gallery across tasks, methods, and datasets.

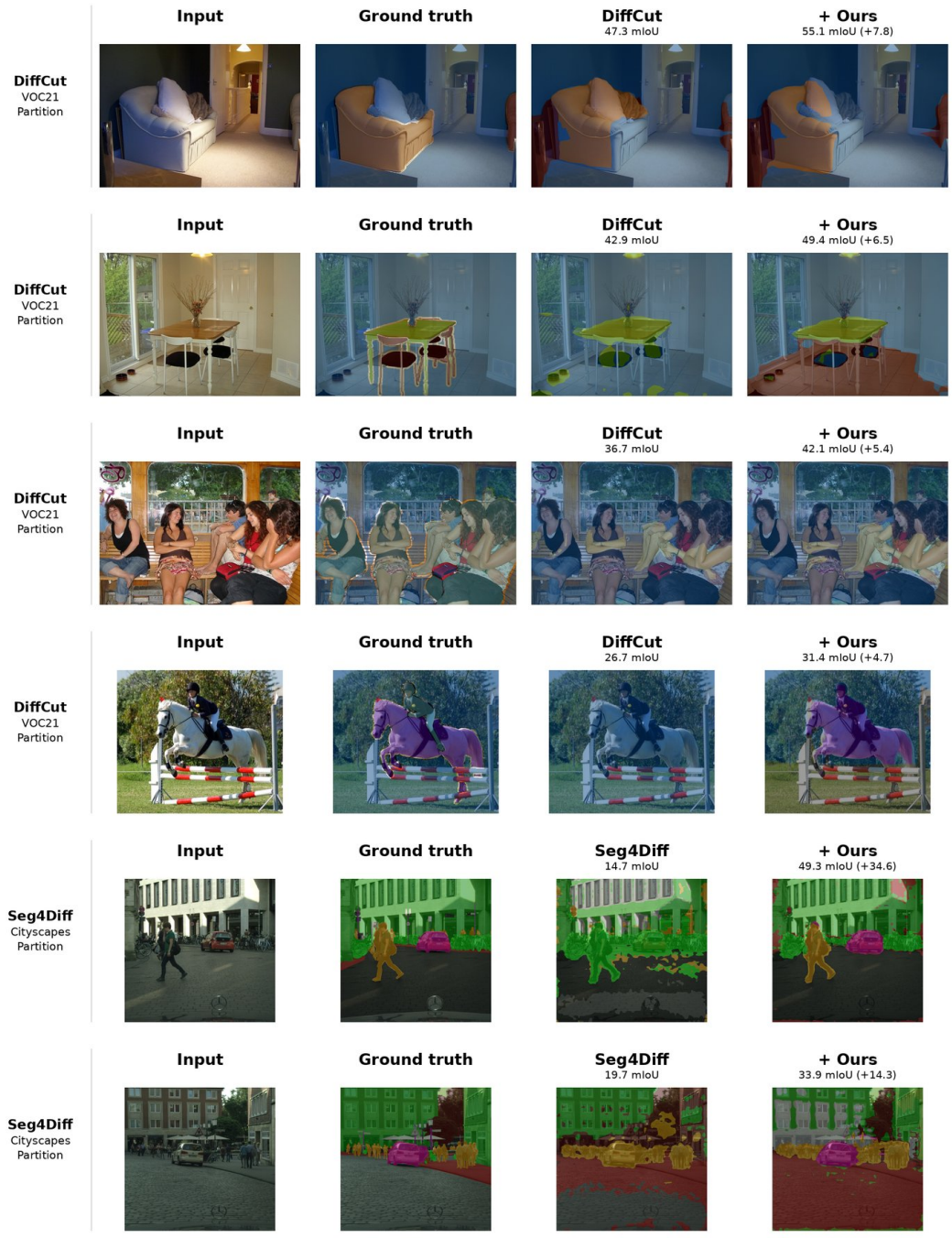

Figure S7: Extended qualitative gallery across tasks, methods, and datasets.