VISA: VLM-GUIDED INSTANCE SEMANTIC AUDITING FOR 3D SEMANTIC OCCUPANCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic 3D occupancy provides a compact scene representation for autonomous driving and robotic decision making, yet closed-set object semantics remain challenging for rare and visually confusable classes. A common way to transfer vision-language model (VLM) knowledge is to align 3D features with language embeddings; however, we find that such alignment can improve text-space objectives without reliably improving voxel-level semantic prediction. We introduce **VISA**, a task-aligned VLM supervision framework that uses an offline VLM as a structured semantic auditor rather than an embedding target. For each physical object instance, VISA produces a reliability-aware audit containing a closed-set class hypothesis, plausible confusions, and visual attributes, propagates the audit along the object track, and grounds it only to matched 3D object voxels. The resulting structured supervision is distilled directly into semantic logits, while the underlying occupancy architecture and inference pipeline remain unchanged. On nuScenes, VISA yields 3.75–5.72% relative mIoU improvements across two backbones, with 6.16–9.92% relative object-mIoU gains and 8.46–11.70% relative rare-class gains. Beyond autonomous driving, VISA achieves 1.47–2.68% relative mIoU improvements on NYUv2 across four monocular SSC architectures, extending its benefits to indoor scene completion from single RGB images without video input. These results support the applicability of VISA across outdoor and indoor settings, covering both temporal occupancy models and single-image scene-completion architectures.

1 INTRODUCTION

3D semantic occupancy prediction provides a volumetric scene representation for autonomous driving and indoor robotic perception, jointly describing foreground objects, background regions, and free space Song et al. (2017); Cao & de Charette (2022); Huang et al. (2023); Wei et al. (2023); Tian et al. (2023). Such representations provide spatial information for scene understanding, future-state prediction, and planning Zheng et al. (2024b); Zuo et al. (2025b). Their usefulness depends on both geometric accuracy and semantic reliability: a correctly reconstructed occupied region may still be assigned an incorrect category. Distinguishing buses, trucks, and trailers, or bicycles and motorcycles, can be difficult when their visual evidence is incomplete or ambiguous. Occlusion, limited image resolution, and imbalanced class frequencies further complicate these distinctions, particularly for rare objects. Despite progress in camera-based occupancy reconstruction Li et al. (2023b); Huang et al. (2023); Wei et al. (2023), improving object-level semantic reliability remains an important challenge. We investigate how additional visual-semantic knowledge can complement dense occupancy labels while preserving the deployed predictor.

Vision-language models (VLMs) offer transferable semantic knowledge through image-language pretraining Radford et al. (2021); Zhai et al. (2023); Li et al. (2023a); Liu et al. (2023). Prior work has demonstrated the value of language-aligned features and region-level language supervision for open-vocabulary 3D scene understanding Peng et al. (2023); Ding et al. (2023); Yang et al. (2024); Vobecký et al. (2023). Building on this direction, we study how VLM knowledge can supplement existing labels in image-based, closed-set semantic occupancy prediction. One possible adaptation is to generate captions for object crops and align the corresponding 3D features with their language embeddings. However, optimizing feature–caption similarity does not directly specify how the transferred information should influence voxel-level decisions under a fixed task taxonomy. Moreover,

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.

dense occupancy labels already identify the target categories, raising the question of what complementary information a VLM should provide beyond another class prediction. We therefore examine both the form and grounding of this auxiliary supervision: *should crop-level VLM knowledge be transferred through feature alignment, or explicitly organized into instance-specific semantic guidance for the corresponding 3D objects?*

We propose **VISA**, a training-time semantic auditing framework that uses an offline VLM as a *structured auditor* rather than an embedding target. VISA extracts closed-set class hypotheses, plausible visual confusions, visual attributes, and confidence and reliability estimates from representative object crops. These audits are converted into auxiliary supervision over semantic logits through reliability-weighted taxonomy, attribute-factor, and scene-level audit graph objectives, while the original occupancy labels remain supervision anchors. Physical-instance audit-to-voxel grounding uses training-time object regions and voxel labels to restrict each audit to matched object voxels, rather than indiscriminately supervising surrounding background, free space, or neighboring instances. When temporal associations are available, an audit is reused along the same physical object track. Audits are generated offline and cached; the occupancy architecture and inference inputs remain unchanged, with no VLM, text encoder, or additional module required at inference. Our main contributions are:

1. We investigate language-based supervision for closed-set occupancy through task-adapted caption-alignment, class-text prototype, and calibrated crop-teacher controls. Our occupancy-specific diagnostics show that improved caption alignment does not necessarily translate into improved object and rare-class voxel predictions.
2. We propose VISA, which converts offline VLM judgments into structured semantic audits and distills their class ambiguity, visual attributes, and cross-instance relations into semantic logits through reliability-aware auxiliary supervision.
3. We develop physical-instance audit-to-voxel grounding, restricting sparse crop-level supervision to matched 3D object voxels and reusing audits across associated frames when tracks are available, without modifying the inference-time occupancy model.
4. We validate VISA across six architectures on nuScenes occupancy prediction and indoor NYUv2 semantic scene completion (SSC). On nuScenes, VISA achieves 3.75–5.72% relative mIoU improvements, with 6.16–9.92% relative object-mIoU gains and 8.46–11.70% relative rare-class-mIoU gains. On NYUv2, VISA achieves 1.47–2.68% relative mIoU improvements across four monocular SSC architectures, demonstrating its applicability to indoor scene completion from single RGB images without video input.

2 Related Work

3D Semantic Occupancy Prediction: Semantic scene completion jointly infers scene geometry and semantic categories Song et al. (2017); Cao & de Charette (2022), while surround-view occupancy benchmarks extend volumetric prediction to autonomous-driving environments Wang et al. (2023); Tian et al. (2023). Representative camera-based architectures construct BEV or voxel representations through view transformation, spatiotemporal feature aggregation, and 3D decoding Li et al. (2022b); Zhang et al. (2023); Li et al. (2023c); Wang et al. (2024). Other approaches reduce dependence on dense 3D annotations through multi-view self-supervision and differentiable rendering from 2D observations Huang et al. (2024a); Pan et al. (2024). Occupancy representations also support temporal modeling: OccWorld predicts future occupancy and ego trajectories Zheng et al. (2024b), while GaussianWorld exploits scene evolution for streaming occupancy prediction Zuo et al. (2025b). VISA complements these architectural and temporal advances with training-time semantic supervision. It operates on the semantic predictions of existing occupancy and scene-completion models rather than introducing a new scene representation or transition model.

Vision-Language Supervision for 3D Perception: Language supervision has been transferred to 3D perception through both feature distillation and caption-based learning. OpenScene distills language-aligned 2D features into 3D point representations Peng et al. (2023). PLA associates image-derived captions with scene-, view-, and entity-level point sets and learns point–language embeddings through contrastive supervision Ding et al. (2023). RegionPLC constructs regional 3D–language pairs from multiple 2D foundation models and introduces region-aware,

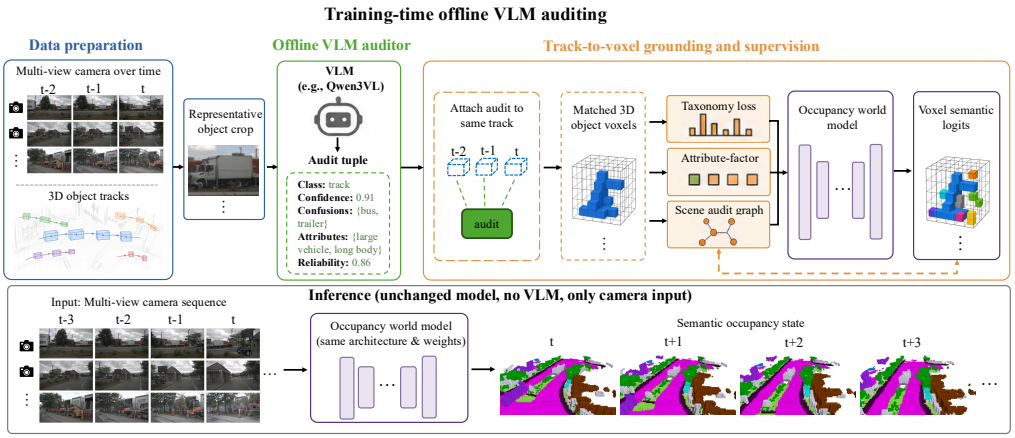


Figure 1: Overview of **VISA**. During training, object tracks are converted into representative crops and audited by an offline VLM to obtain closed-set class hypotheses, plausible confusions, attributes, reliability, and evidence. The audit is propagated to 3D boxes of the same physical instance, grounded to matched object voxels, and distilled into semantic occupancy logits through taxonomy, attribute, and scene-level audit graph supervision. At inference, the original occupancy world model is unchanged and requires no VLM, text encoder, crop extraction, or object-track input.

point-discriminative contrastive learning Yang et al. (2024). These methods primarily target open-vocabulary semantic and instance segmentation of input point clouds. For occupancy prediction, OVO transfers knowledge from a 2D open-vocabulary segmentation teacher through pixel–voxel filtering Tan et al. (2023); POP-3D and LangOcc learn language-aligned voxel representations through feature distillation and volume-rendering supervision, respectively Vobecký et al. (2023); Boeder et al. (2025). LOcc constructs dense voxel–text supervision through semantic transitive labeling Yu et al. (2025), while VEON adapts depth and CLIP priors for open-vocabulary occupancy Zheng et al. (2024a). VISA instead studies auxiliary VLM supervision for closed-set occupancy with dense training labels retained. It converts crop-level class ambiguity, attributes, and reliability into structured supervision over matched object logits, rather than learning a language-embedding output space. Our crop-caption baseline is a task-adapted control of the alignment mechanism, not a reproduction of these complete systems.

Taxonomy- and Attribute-Aware Supervision: Semantic structure has been incorporated into dense prediction through hierarchical class relations Li et al. (2022a). More broadly, class-balanced objectives Cui et al. (2019), decoupled representation and classifier learning Kang et al. (2020), and supervised contrastive learning Khosla et al. (2020) address class imbalance and discriminative representation learning. Within occupancy and scene completion, class-guided sampling and contextual semantic relations also improve learning Zhang et al. (2023); Cao & de Charette (2022). VISA builds on the use of semantic structure but obtains auxiliary targets from instance-level visual audits. It retains a task-specific class-to-attribute mapping while supervising instance-specific attributes, plausible class confusions, and relations derived from audited objects. Dense occupancy labels remain supervision anchors, and audit reliability controls filtering and weighting. GT-class attribute and calibrated crop-teacher controls examine whether the audit signal contributes beyond fixed class-level structure and uncertainty-weighted soft targets.

3 VISA

We advise readers to review Appendix A first. In this section, we begin by analyzing crop-caption alignment to motivate the design of VISA. We then detail how structured VLM audits are grounded to object regions and converted into auxiliary supervision for semantic occupancy prediction.

From Caption Alignment to Semantic Auditing: Motivated by caption-based 3D language supervision Ding et al. (2023); Yang et al. (2024), we study a controlled adaptation to closed-set occupancy. For each object crop, a VLM generates a caption, which is encoded into a text embedding and aligned with the corresponding pooled 3D feature. Let $\mathbf{h}_{i,t}$ denote the pooled feature of object i

at frame t, $\mathbf{g}_i^{\text{text}}$ its caption embedding, and ϕ a learned projection head. The alignment objective is

$$\mathcal{L}_{\text{cap}} = 1 - \cos\big(\phi(\mathbf{h}_{i,t}), \mathbf{g}_i^{\text{text}}\big). \tag{1}$$

This is a task-adapted supervision control rather than a reproduction of any single prior method.

Fig. 2 contrasts text-space learning with downstream occupancy prediction. Paired caption cosine, nearest class-prototype accuracy, and caption retrieval measure the learned feature–language correspondence. Although the caption branch improves these auxiliary metrics, the illustrated experiment does not show corresponding improvements in semantic mIoU or occupied IoU. Thus, optimizing feature–caption alignment alone does not guarantee better closed-set voxel predictions. This diagnostic motivates transferring VLM information through task-specific semantic targets while retaining dense occupancy labels as anchors. VISA organizes instance-level ambiguity, visual attributes, and reliability into structured audits and grounds their supervision to the corresponding 3D object regions.

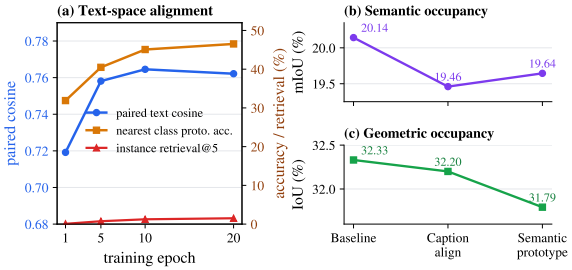


Figure 2: Diagnostic of generic language alignment for occupancy. Left: paired text cosine measures projected 3D feature–caption similarity, nearest class-prototype accuracy uses closed-set class text prototypes, and instance retrieval@5 retrieves the paired caption from the caption bank. Right: caption alignment uses crop captions, while semantic prototype uses class text prototypes; both remain below the occupancy baseline.

Instance Semantic Audit Tuple: For each representative object crop, VISA queries the offline VLM with a closed-set class vocabulary and a predefined visual-factor vocabulary. The response is parsed into the audit tuple $\mathcal{A}_i$ defined in Eq. 13. The vocabularies are specified for the task taxonomy, while the audit fields are predicted from the individual crop. For example, a vehicle crop may yield a `bus` hypothesis, plausible `truck` or `trailer` confusions, and attributes such as `large_vehicle` and `long_body`. Unlike a class-level attribute lookup, the audit records the visual evidence and ambiguity of the particular instance. The evidence string e_i is retained for inspection but is not used as a training target.

Physical-Instance Audit-to-Voxel Grounding: VLM audits describe sparse image regions, whereas the occupancy model predicts semantic logits throughout a 3D volume. VISA connects these representations through the matched object-voxel set $\mathcal{V}_{i,t}$ in Eq. 14. Training-time region annotations and voxel labels restrict supervision to class-matched voxels within the associated object region, reducing leakage to unrelated scene regions. When temporal associations are available, the same audit is reused along the physical object's track; otherwise, it is applied to the corresponding single-frame object region.

For each valid region, we pool the predicted voxel probabilities:

$$\bar{\mathbf{p}}_{i,t} = \frac{1}{|\mathcal{V}_{i,t}|} \sum_{v \in \mathcal{V}_{i,t}} \text{softmax}(\mathbf{z}_{v,t}). \tag{2}$$

The following audit losses operate on this region-level prediction and backpropagate to the matched voxel logits. The original occupancy objective continues to supervise the full valid voxel grid.

Reliability-Weighted Taxonomy Distillation: The taxonomy objective supplements the annotated class with instance-specific ambiguity from the audit. The VLM-reported class confidence q_i and visual reliability r_i are used for filtering and weighting, not treated as calibrated probabilities. An audit is eligible for taxonomy supervision when $q_i \geq \tau_q$ and $r_i \geq \tau_r$, where τ_q and τ_r are the confidence and reliability thresholds.

For each eligible audit, we construct a normalized distribution $\mathbf{s}_i$ from its class hypothesis and scored confusion set:

$$s_{i,c} \propto q_i \mathbb{1}[c = c_i^{\text{vlm}}] + \sum_{(c',\gamma) \in \mathcal{C}_i} \gamma \mathbb{1}[c = c']. \tag{3}$$

Here, γ is the reported confidence of a confusion class, and $\mathbf{s}_i$ is normalized across the semantic classes, with zero mass outside the audited object vocabulary. We blend this distribution with the annotated object class:

$$\mathbf{t}_i^{\text{tax}} = (1-\alpha)\,\text{onehot}(y_i^{\text{obj}}) + \alpha\,\mathbf{s}_i, \tag{4}$$

where $\alpha \in [0,1]$ controls the contribution of the audit distribution. Let $\mathcal{I}_{\text{tax}}$ contain valid object-frame pairs whose audits pass the thresholds. Using $w_i = r_i \max(q_i, 0.25)$, the taxonomy loss is

$$\mathcal{L}_{\text{tax}} = -\frac{1}{|\mathcal{I}_{\text{tax}}|} \sum_{(i,t)\in\mathcal{I}_{\text{tax}}} w_i \sum_{c=1}^{C} t_{i,c}^{\text{tax}} \log \bar{p}_{i,t,c}. \tag{5}$$

This target represents plausible class confusions without replacing the original occupancy annotation.

Attribute-Factor Distillation: Class hypotheses and confusions do not explicitly represent the visual properties shared by related categories. VISA therefore also supervises the audited attribute vector $\mathbf{u}_i$. We define a fixed class-to-attribute matrix $\mathbf{M} \in \{0,1\}^{C_{\text{obj}}\times A}$, where C_{obj} is the number of object classes and A is the number of attributes. Each row specifies the class-level attribute relations for the task taxonomy; for example, bicycles and motorcycles share `two_wheeler`. The matrix is fixed within a dataset, whereas the target $\mathbf{u}_i$ is obtained from the individual crop rather than from its annotated class.

Let $\bar{\mathbf{p}}_{i,t}^{\text{obj}}$ denote the pooled object-class probability vector used by the attribute branch. Using column-vector notation, the predicted attributes are

$$\hat{\mathbf{u}}_{i,t} = \mathbf{M}^\top \bar{\mathbf{p}}_{i,t}^{\text{obj}}. \tag{6}$$

We match these predictions to the audited attributes using binary cross-entropy:

$$\mathcal{L}_{\text{attr}} = \frac{1}{|\mathcal{I}_{\text{attr}}|} \sum_{(i,t)\in\mathcal{I}_{\text{attr}}} r_i\ \text{BCE}(\hat{\mathbf{u}}_{i,t}, \mathbf{u}_i), \tag{7}$$

where $\mathcal{I}_{\text{attr}}$ contains valid object-frame pairs with reliable audits and valid attribute targets. This objective encourages semantic predictions to reflect instance-level visual properties in addition to the annotated category.

Scene-Level Audit Graph: Taxonomy and attribute supervision act on individual object regions. To incorporate relations between co-occurring objects, VISA constructs an audit graph at each frame. Nodes are reliable audited objects with valid matched voxels and non-empty attribute vectors. Edges are derived deterministically from the audit fields rather than predicted as additional VLM edge scores.

For nodes i and j, let $\text{sim}_a(\mathbf{u}_i, \mathbf{u}_j)$ denote the Jaccard similarity of their binary attribute vectors. Let $\text{compat}(i,j)$ hold when the audits share a class hypothesis or either hypothesis appears in the other's confusion set. We define positive and negative edges by

$$(i,j) \in \mathcal{E}_t^+ \quad \text{if} \quad \text{sim}_a(\mathbf{u}_i, \mathbf{u}_j) \geq \tau_+ \quad \text{or} \quad \text{compat}(i,j), \tag{8}$$

$$(i,j) \in \mathcal{E}_t^- \quad \text{if} \quad \text{sim}_a(\mathbf{u}_i, \mathbf{u}_j) \leq \tau_- \quad \text{and} \quad \neg\,\text{compat}(i,j), \tag{9}$$

where $\tau_- < \tau_+$ are similarity thresholds. Pairs satisfying neither condition are omitted.

The predicted relation is the cosine similarity between the model's attribute vectors: $\rho_{ij} = \cos(\hat{\mathbf{u}}_{i,t}, \hat{\mathbf{u}}_{j,t})$. With edge weight $w_{ij} = \min(r_i, r_j)$, the graph objective is

$$\mathcal{L}_{\text{graph}} = \frac{1}{Z}\Big[\sum_{(i,j)\in\mathcal{E}_t^+} w_{ij}(1-\rho_{ij}) + \sum_{(i,j)\in\mathcal{E}_t^-} w_{ij} \max(0, \rho_{ij} - m)\Big], \tag{10}$$

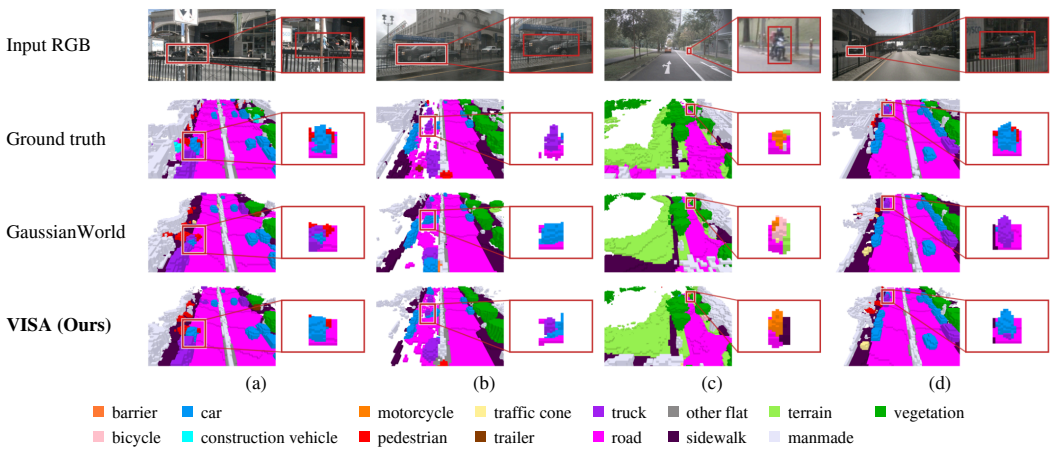


Figure 3: Qualitative comparisons on nuScenes. Each column shows one scene, with input RGB, ground truth, GaussianWorld, and GaussianWorld + VISA (ours), from top to bottom. Red boxes and connecting lines indicate the regions shown in the enlarged views. VISA improves semantic consistency with the ground truth in the highlighted regions.

Table 1: NuScenes semantic occupancy benchmark. IoU denotes binary occupied-space IoU and mIoU denotes semantic mIoU over 16 foreground classes. All numbers are percentages. For reproduced world-model backbones, IoU and mIoU are reported as mean ± standard deviation over three seeds, and per-class IoUs are averaged over the same seeds.

Method	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene Cao & de Charette (2022)	23.96	7.31	4.03	0.35	8.00	8.04	2.90	0.28	1.16	0.67	4.01	4.35	27.72	5.20	15.13	11.29	9.03	14.86
Atlas Murez et al. (2020)	28.66	15.00	10.64	5.68	19.66	24.94	8.90	8.84	6.47	3.28	10.42	16.21	34.86	15.46	21.89	20.95	11.21	20.54
BEVFormer Li et al. (2022b)	30.50	16.75	14.22	6.58	23.46	28.28	8.66	10.77	6.64	4.05	11.20	17.78	37.28	18.00	22.88	22.17	13.80	22.21
TPVFormer Huang et al. (2023)	30.86	17.10	15.96	5.31	23.86	27.32	9.79	8.74	7.09	5.20	10.97	19.22	38.87	21.25	24.26	23.15	11.73	20.81
OccFormer Zhang et al. (2023)	31.39	19.03	18.65	10.41	23.92	30.29	10.31	14.19	13.59	10.13	12.49	20.77	38.78	19.79	24.19	22.21	13.48	21.35
SurroundOcc Wei et al. (2023)	31.49	20.30	20.59	11.68	28.06	30.86	10.70	15.14	14.09	12.06	14.38	22.26	37.29	23.70	24.49	22.77	14.89	21.86
GaussianFormer Huang et al. (2024b)	29.83	19.10	19.52	11.26	26.11	29.78	10.47	13.83	12.58	8.67	12.74	21.57	39.63	23.28	24.46	22.99	9.59	19.12
GaussianFormer-2 Huang et al. (2025)	31.74	20.82	21.39	13.44	28.49	30.82	10.92	15.84	13.55	10.53	14.04	22.92	40.61	24.36	26.08	24.27	13.83	21.98
QuadricFormer Zuo et al. (2025a)	32.13	21.11	21.38	13.41	28.40	31.01	11.32	17.10	13.94	11.28	14.75	22.66	40.81	24.71	26.51	25.22	13.54	21.78
OccWorld Zheng et al. (2024b)	31.64	19.06	15.55	13.38	25.16	27.49	10.08	13.67	12.44	9.52	9.16	17.83	**42.34**	24.58	25.32	23.95	10.97	**23.48**
OccWorld + VISA	**33.51**	**20.15**	**17.53**	**15.61**	**27.83**	**28.03**	**10.47**	**15.94**	**13.12**	**10.54**	**10.07**	**20.44**	41.52	**25.42**	**25.79**	**24.20**	**12.81**	23.01
GaussianWorld Zuo et al. (2025b)	32.00	21.33	20.01	12.88	26.94	31.12	12.64	17.10	13.31	10.44	14.33	22.93	42.52	**25.33**	27.56	25.77	14.22	24.23
GaussianWorld + VISA	**34.51**	**22.13**	**21.47**	**15.47**	**27.92**	**31.64**	**14.21**	**19.03**	**14.01**	10.33	**15.31**	**23.46**	**42.97**	24.56	**28.05**	**26.24**	**14.84**	**24.60**

where m is the negative-edge similarity margin and

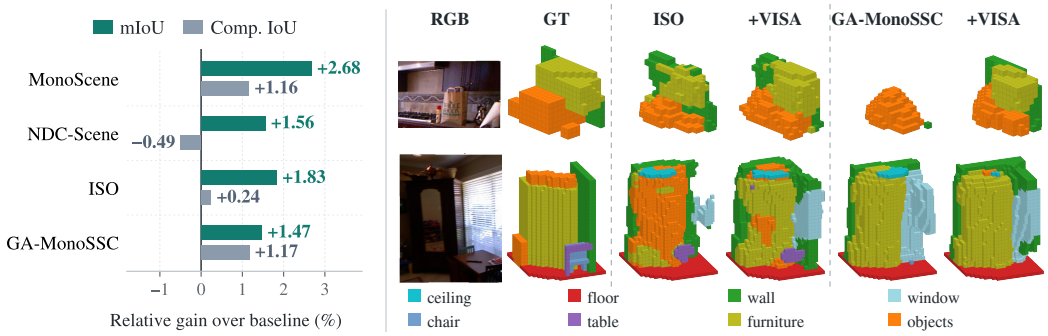
$$Z = \sum_{(i,j)\in\mathcal{E}_t^+} w_{ij} + \sum_{(i,j)\in\mathcal{E}_t^-} w_{ij} \tag{11}$$

is the total valid edge weight. The objective encourages compatible audited objects to share predicted semantic factors while separating incompatible pairs.

Training Objective and Inference: The VISA term in Eq. 15 combines the three auxiliary objectives:

$$\mathcal{L}_{\text{VISA}} = \lambda_{\text{tax}}\mathcal{L}_{\text{tax}} + \lambda_{\text{attr}}\mathcal{L}_{\text{attr}} + \lambda_{\text{graph}}\mathcal{L}_{\text{graph}}, \tag{12}$$

where the λ coefficients control their respective contributions. These losses supplement the backbone's original occupancy objective. All audits are generated offline from training-set crops. At inference, the model uses only its original inputs and architecture; audit records, grounding annotations, and auxiliary losses are not needed, and no VLM, text encoder, or additional module is introduced.



(a) Quantitative generalization **(b) Selected indoor predictions**

Figure 4: **Results on NYUv2.** (a) Relative gains in completion IoU and mIoU across four backbone/VISA pairs. (b) Qualitative examples showing input RGB, ground truth, and predictions from ISO and GA-MonoSSC with and without VISA. Colors denote semantic classes.

4 Experiments

4.1 Experimental Setup and Main Results

Experimental setup. We evaluate VISA on nuScenes (700/150 training/validation scenes) and NYUv2 (795/654 training/test samples) using OccWorld and GaussianWorld on nuScenes, and MonoScene, NDC-Scene, ISO, and GA-MonoSSC on NYUv2, respectively (Caesar et al., 2020; Cao & de Charette, 2022). All backbones retain their native occupancy objectives and inference inputs; VLM audits use training images only. We report occupied-space/completion IoU and semantic mIoU. On nuScenes, Obj. and Rare denote unweighted mIoU over the ten object classes and six rare classes (bicycle, bus, construction vehicle, motorcycle, traffic cone, and trailer), respectively. Implementation and evaluation details are in the appendix. All IoU scores are percentages.

Results on nuScenes. Table 1 reports the overall and per-class results. VISA improves semantic mIoU by 5.72% on OccWorld and 3.75% on GaussianWorld, with corresponding occupied-space IoU gains of 5.91% and 7.84%. Object mIoU improves by 9.92% and 6.16%, while rare-class mIoU improves by 11.70% and 8.46%, respectively. The larger relative semantic gains on object and rare classes are consistent with VISA's object-centric supervision. Figure 3 complements these quantitative results with comparisons of GaussianWorld with and without VISA. In the highlighted regions, VISA reduces local category confusions and produces object semantics that more closely match the ground truth.

Results on NYUv2. We evaluate VISA on NYUv2 to test whether its benefits extend to indoor scene completion from a single RGB image, without video input or temporal modeling. Each backbone and its VISA-augmented variant are trained on the NYUv2 training split and evaluated on the test split. Figure 4(a) shows consistent semantic improvements across all four backbones, with relative mIoU gains of 1.47–2.68%. Completion IoU also improves on three backbones, with relative gains of 0.24–1.17%. For NDC-Scene, mIoU improves by 1.56% despite a 0.49% decrease in completion IoU, indicating that semantic gains do not always coincide with improvements in binary occupied-space reconstruction. Figure 4(b) provides complementary qualitative comparisons for ISO and GA-MonoSSC. These results demonstrate that VISA also benefits single-image indoor models, extending its applicability beyond temporal occupancy modeling on nuScenes.

4.2 Alternative Supervision

GT-region and crop-teacher controls. Table 2(a) examines auxiliary supervision from ground-truth labels and a supervised crop classifier. GT-region supervision applies an additional GT class loss to matched object regions, using either uniform instance weights or VLM-derived audit weights. The latter retains VLM information through weighting. Our crop teacher uses ConvNeXt (Liu et al., 2022) with temperature scaling (Guo et al., 2017). Its taxonomy variant supervises class predictions using calibrated teacher probabilities, while its full-structure variant additionally derives attribute and relation targets through fixed class-level mappings.

Table 2: Alternative supervision on GaussianWorld (%). Panels group methods by the source of their auxiliary targets. Crop teachers are calibrated, and audit-weighted GT supervision retains VLM-derived weights. All results use the same nuScenes validation evaluation.

(a) GT and crop-teacher targets

Supervision	IoU	mIoU	Obj.	Rare
GaussianWorld	32.00	21.33	18.17	15.72
GT-region (uniform)	32.63	21.53	18.95	16.62
GT-region (audit weights)	32.56	21.60	19.04	16.70
Crop teacher: taxonomy	32.89	21.71	18.85	16.47
Crop teacher: full structure	33.01	21.76	18.93	16.58
Full VISA	**34.51**	**22.13**	**19.29**	**17.05**

(b) Language and audit targets

Supervision	IoU	mIoU	Obj.	Rare
GaussianWorld	32.00	21.33	18.17	15.72
Caption alignment	32.57	21.56	18.57	16.08
Class-text prototypes	32.70	21.59	18.67	16.21
VLM class/confusion	33.36	21.67	18.58	15.98
Full VISA	**34.51**	**22.13**	**19.29**	**17.05**

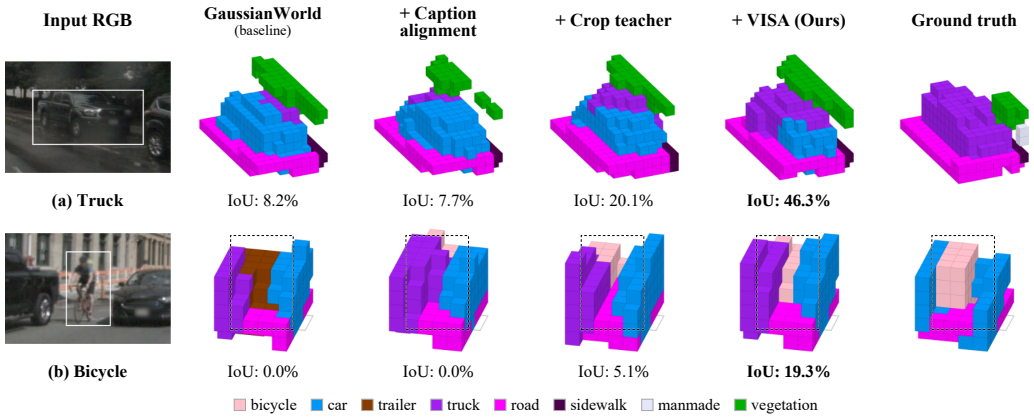


Figure 5: Local semantic occupancy predictions under different training-time supervision strategies on GaussianWorld. Rows (a) and (b) show selected truck and bicycle examples, respectively. Colors denote semantic classes, and the scores below the predictions report the IoU for each displayed example. Boxes in the RGB images identify the target objects; dashed outlines in (b) are visual guides rather than evaluation masks.

Language and audit controls. Table 2(b) compares language and audit supervision. Motivated by caption-based 3D learning (Ding et al., 2023; Yang et al., 2024), our caption-alignment control aligns pooled 3D features with embeddings of crop descriptions generated by Qwen3-VL (Bai et al., 2025), using a frozen SigLIP text encoder (Zhai et al., 2023). Class-text prototypes instead use fixed class-name embeddings, following the use of text-defined class representations in vision-language models (Radford et al., 2021). VLM class/confusion uses audited class hypotheses and plausible confusion classes without attribute or graph supervision. Full VISA combines taxonomy, attribute, and graph objectives.

Effect of supervision design. Table 2 examines whether additional supervision or simpler teacher targets can match the semantic benefits of VISA. GT-region supervision, caption alignment, and class-text prototypes improve over the GaussianWorld reference, but remain below Full VISA on all three semantic metrics. Compared with VLM class/confusion supervision, the separately trained Full VISA reference achieves 3.82% higher Obj. and 6.70% higher Rare. It also exceeds the full-structure crop teacher by 1.70%, 1.90%, and 2.83% in mIoU, Obj., and Rare, respectively, despite this control providing attribute and relation targets through fixed class-level mappings. These comparisons support the value of structured, instance-specific visual audits beyond the simpler supervision alternatives tested here. Figure 5 provides complementary qualitative evidence: in the selected examples, VISA reduces car–truck confusion and recovers more of the bicycle region than caption alignment and the displayed crop-teacher variant. The component study in Section 4.3 further examines the contributions of the individual objectives.

4.3 Components and Audit Correspondence

Components and attributes. Table 3(a,b) tests whether attribute supervision benefits from visual information beyond a fixed GT-class-to-attribute mapping (GT Attr.). Instance-derived attribute targets (Attr.) improve Obj. by 6.72% on GaussianWorld and 2.58% on OccWorld, supporting the value of crop-specific attribute evidence. Combining taxonomy (Tax.) and attributes improves mIoU over either component alone, although Obj. and Rare do not improve uniformly. Adding the graph objective to Tax.+Attr. improves all three metrics on both backbones, including relative Rare gains of 4.28% and 1.89%, respectively. This supports relational supervision between audited instances as a useful complement to individual taxonomy and attribute targets.

Instance correspondence. Table 3(c) tests whether correct audit-to-instance assignment contributes beyond class membership. We permute audit semantics or attributes within each class while preserving receiving-instance eligibility and weights. Both permutations reduce mIoU, Obj., and Rare, with relative Rare declines of 1.23% and 2.11% for semantic and attribute shuffling, respectively. The decreases support a contribution from matching audit content to its originating instance: retaining within-class audit information alone does not preserve the full performance of VISA.

Table 3: Components and instance correspondence. (a) GaussianWorld component study. (b) OccWorld component study. (c) Within-class instance-pairing permutation.

(a) GaussianWorld Components

Variant	mIoU	Obj.	Rare
Backbone	21.33	18.17	15.72
Tax.	21.51	18.17	16.05
Attr.	21.44	18.91	16.57
Tax.+Attr.	21.78	18.66	16.35
GT Attr.	21.22	17.72	15.98
Full VISA	**22.13**	**19.29**	**17.05**

(b) OccWorld Components

Variant	mIoU	Obj.	Rare
Backbone	19.06	15.43	13.50
Tax.	19.53	16.02	14.28
Attr.	19.61	16.31	14.21
Tax.+Attr.	19.86	16.55	14.80
GT Attr.	19.39	15.90	14.00
Full VISA	**20.15**	**16.96**	**15.08**

(c) Instance Correspondence

Assignment	mIoU	Obj.	Rare
Original	**22.13**	**19.29**	**17.05**
Shuffle semantics	22.02	19.14	16.84
Shuffle attributes	21.96	19.05	16.69

4.4 Offline Audit Quality and Cost

Table 4 summarizes the offline audit bank on the nuScenes training split. The VLM's closed-set class predictions agree with ground-truth labels on 87.7% of evaluated object audits. Agreement varies substantially with crop visibility: 92.9% for clear crops, 88.6% for partially visible crops, and 22.9% for poor-visibility crops. These diagnostics motivate filtering and weighting audit supervision according to visual reliability, while retaining dense occupancy labels as the primary training signal.Appendix H reports the effect of audit coverage, offline generation costs, and student training overhead.

Table 4: Offline VLM audit diagnostics on the nuScenes training split.

Statistic	**Value**
Training-scene coverage	700 / 700
Audited 3D-box instances	51,420
Evaluated object-class audits	49,568
Closed-set agreement	87.7%
Clear / partial / poor crops	20.7 / 76.6 / 2.7%
Rare-class audit fraction	15.1%

5 Conclusion

We presented VISA, a training-time framework that converts offline VLM judgments into reliability-aware taxonomy, attribute-factor, and graph supervision grounded to matched object voxels. Dense occupancy labels remain supervision anchors, with no additional module required at inference. Experiments across six backbones on nuScenes and NYUv2 support structured auditing for closed-set occupancy prediction and scene completion. Limitations include offline audit costs, inherited VLM biases, and reliance on training-time object grounding, dense voxel labels, and taxonomy-specific attribute mappings. Audits cover visible objects, while background regions retain the original occupancy supervision. Future work will explore lightweight auditors, selective querying, weaker grounding, and broader scene-context supervision.

ETHICS STATEMENT

This work uses the existing nuScenes and NYUv2 datasets and involves no new human-subject data collection. Although improved semantic occupancy prediction may benefit autonomous driving and embodied perception, prediction errors remain a concern in safety-critical applications. VLM-generated audits may contain incorrect judgments or reflect biases in their pretraining data; reliability weighting mitigates but does not eliminate these risks. Our benchmark results do not establish deployment safety. Any downstream use should respect dataset licenses and privacy requirements and undergo validation under representative operating conditions.

REPRODUCIBILITY STATEMENT

The method and experimental sections describe the VISA objectives, backbone integration, datasets, and evaluation protocols. Reproducing the offline auditing stage additionally requires the VLM configuration, audit prompts and output schema, instance selection and grounding rules, and reliability weighting procedure, which are detailed in the appendix. These details specify how the audits are generated, converted into training supervision, and evaluated without modifying backbone inference.

AI USE STATEMENT

As detailed in Section 3, VISA incorporates a pretrained vision-language model to generate offline, instance-level semantic audits that serve as auxiliary training supervision rather than ground-truth annotations. Generative AI tools were also used to assist with manuscript drafting, LaTeX typesetting, figure and table formatting, and organizing discussions of experimental results. All AI-generated text, code, scripts, figures, and comparative analyses were rigorously reviewed, verified, and edited by the human authors. The authors take full responsibility for the scientific claims, experimental findings, and final contents of this paper.

REFERENCES

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

Simon Boeder, Fabian Gigengack, and Benjamin Risse. LangOcc: Open vocabulary occupancy estimation via volume rendering. In *3DV*, 2025.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *CVPR*, 2022.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-driven open-vocabulary 3D scene understanding. In *CVPR*, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3D semantic occupancy prediction. In *CVPR*, 2023.

Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. SelfOcc: Self-supervised vision-based 3D occupancy prediction. In *CVPR*, 2024a.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. GaussianFormer: Scene as gaussians for vision-based 3D semantic occupancy prediction. In *ECCV*, 2024b.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. GaussianFormer-2: Probabilistic gaussian superposition for efficient 3D occupancy prediction. In *CVPR*, 2025.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a.

Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *CVPR*, 2022a.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In *CVPR*, 2023b.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022b.

Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M. Alvarez. FB-OCC: 3D occupancy prediction based on forward-backward view transformation. In *CVPR Workshop on End-to-End Autonomous Driving*, 2023c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022.

Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3D scene reconstruction from posed images. In *ECCV*, 2020.

Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Bing Wang, Hongwei Xie, Li Liu, and Shanghang Zhang. RenderOcc: Vision-centric 3D occupancy prediction with 2D rendering supervision. In *CVPR*, 2024.

Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *CVPR*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017.

Zhiyu Tan, Zichao Dong, Cheng Zhang, Weikun Zhang, Hang Ji, and Hao Li. OVO: Open-vocabulary occupancy. *arXiv preprint arXiv:2305.16133*, 2023.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving. In *NeurIPS*, 2023.

Antonín Vobecký, Oriane Siméoni, David Hurych, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. POP-3D: Open-vocabulary 3D occupancy prediction from images. In *NeurIPS*, 2023.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. OpenOccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, 2023.

Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. PanoOcc: Unified occupancy representation for camera-based 3D panoptic segmentation. In *CVPR*, 2024.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In *ICCV*, 2023.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding. In *CVPR*, 2024.

Zhu Yu, Bowen Pang, Lizhe Liu, Runmin Zhang, Qiang Li, Si-Yuan Cao, Maochun Luo, Mingxia Chen, Sheng Yang, and Hui-Liang Shen. Language driven occupancy prediction. In *ICCV*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *ICCV*, 2023.

Jilai Zheng, Pin Tang, Zhongdao Wang, Guoqing Wang, Xiangxuan Ren, Bailan Feng, and Chao Ma. VEON: Vocabulary-enhanced occupancy prediction. In *ECCV*, 2024a.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3D occupancy world model for autonomous driving. In *ECCV*, 2024b.

Sicheng Zuo, Wenzhao Zheng, Xiaoyong Han, Longchao Yang, Yong Pan, and Jiwen Lu. QuadricFormer: Scene as superquadrics for 3D semantic occupancy prediction. In *NeurIPS*, 2025a.

Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. GaussianWorld: Gaussian world model for streaming 3D occupancy prediction. In *CVPR*, 2025b.

A Problem Setup

We study how offline VLM audits can provide auxiliary training-time supervision for 3D semantic occupancy prediction. The occupancy model predicts dense closed-set semantic logits, while the VLM observes sparse object-centric image crops and produces instance-level semantic judgments. Our goal is to convert these audits into voxel-compatible supervision without changing the inference-time model.

Semantic 3D Occupancy Prediction: The occupancy model f_θ takes camera observations $\mathbf{I}_{1:t}$ and predicts dense semantic occupancy logits $\mathbf{Z}_t = f_\theta(\mathbf{I}_{1:t})$, where $\mathbf{Z}_t \in \mathbb{R}^{H\times W\times D\times C}$. Here, H, W, and D specify the voxel-grid dimensions, and C is the number of semantic classes, including object, stuff, and empty categories. Each voxel v at frame t has a class-logit vector $\mathbf{z}_{v,t} \in \mathbb{R}^C$. For single-frame prediction, the input reduces to $\mathbf{I}_t$ and no temporal associations are required. Given dense voxel labels $\mathbf{Y}_t$, we retain the backbone's original training objective $\mathcal{L}_{\text{occ}}$, including its native loss terms and weights. VISA does not assume a particular internal representation, temporal update rule, or form of $\mathcal{L}_{\text{occ}}$; it adds auxiliary supervision to the semantic logits.

Offline VLM Audits and Object Voxels: During training, we additionally use object-level regions and offline VLM audits. For each audited object i, a representative image crop is queried with a closed-set class vocabulary and a predefined visual-factor vocabulary. The VLM returns a JSON response, which we parse into an audit tuple:

$$\mathcal{A}_i = \left(c_i^{\text{vlm}}, q_i, \mathcal{C}_i, \mathbf{u}_i, r_i, e_i\right), \tag{13}$$

where c_i^{vlm} is a closed-set class hypothesis, $q_i \in [0,1]$ is its VLM-reported confidence, $\mathcal{C}_i$ contains plausible confusion classes and their reported confidence scores, $\mathbf{u}_i \in \{0,1\}^A$ encodes A predefined visual attributes, $r_i \in [0,1]$ is the reported visual reliability of the crop, and e_i is a short evidence string used for inspection. Confidence and reliability are audit metadata for filtering and weighting, not calibrated guarantees of semantic correctness. These audits describe visible object crops rather than the full occupancy grid or stuff regions. When instance tracks are available, the same audit is associated with the object across frames; otherwise, it is grounded within the corresponding frame.

Let $\mathcal{R}_{i,t}$ denote the training-time 3D region associated with object i at frame t, instantiated by its annotated 3D box in the nuScenes setting. Let y_i^{obj} be the object's annotated semantic class and $\mathbf{x}_{v,t}$ the center of voxel v in the same coordinate frame as $\mathcal{R}_{i,t}$. Using the dense voxel label $y_{v,t}$ from $\mathbf{Y}_t$, we define the matched object-voxel set as

$$\mathcal{V}_{i,t} = \{v \mid \mathbf{x}_{v,t} \in \mathcal{R}_{i,t},\ y_{v,t} = y_i^{\text{obj}}\}. \tag{14}$$

This restriction confines audit supervision to class-matched voxels within the associated object region, reducing leakage to nearby background, free space, and objects of other classes.

Learning Objective: Given dense occupancy labels, training-time object regions, and offline audits, we optimize the backbone's original objective together with VISA supervision:

$$\min_\theta \mathcal{L}_{\text{occ}}(\theta) + \mathcal{L}_{\text{VISA}}\left(\theta; \{\mathcal{A}_i, \mathcal{V}_{i,t}\}_{i,t}\right), \tag{15}$$

where i indexes audited objects and t indexes frames containing their associated 3D regions. The original occupancy objective supervises the full valid voxel grid, while VISA provides auxiliary supervision over matched object regions. Section 3 instantiates $\mathcal{L}_{\text{VISA}}$ with reliability-weighted taxonomy distillation, attribute-factor supervision, and scene-level audit graph regularization. Audits and grounding annotations are used only during training and are not required at inference.

B Implementation Details

Crop extraction and view selection. We generate VLM audits only from the nuScenes training split. For each annotated 3D object box, we project the box into all six surround-view cameras and compute a tight 2D bounding box whenever at least two projected corners are visible. To reduce severe occlusion, we compute inter-box overlap in each camera view and mark a projected crop as occluded when its visible area ratio is below 0.3. We select the best camera view by preferring unoccluded projections with larger visible crop area; if all views are occluded, we fall back to the

best available projection. The final crop is padded by 10% of the tight 2D box size with a minimum padding of 8 pixels, clipped to the image boundary, and saved as a JPEG. We discard crops whose tight 2D box is smaller than 24 pixels on either side. Each crop record stores the scene, frame index, object box index, selected camera, tight and padded 2D boxes, depth, visibility ratio, occlusion flag, number of LiDAR points, and instance token when available.

Blind VLM audit prompt. The VLM is not given the ground-truth class label. Instead, we query Qwen3-VL offline with only the crop image, a closed-set occupancy object vocabulary, and a fixed visual-factor vocabulary. The prompt is designed to elicit structured semantic audits rather than free-form captions. The core prompt is:

> You are auditing a small crop from a vehicle surround-view camera. Choose only from this closed-set class list: barrier, bicycle, bus, car, construction_vehicle, motorcycle, pedestrian, traffic_cone, trailer, truck. Your task is not to write a caption. Output an instance-level taxonomy and visual-factor audit using only visible crop evidence: the most likely closed-set class, confidence, up to three plausible confusion classes, a visual subtype, binary visual attributes useful for 3D semantic occupancy, visual reliability, visibility, motion likelihood, and one short visual reason. Do not use scene context. For visibility, use clear when the class is identifiable from the crop, partial when the crop is truncated, occluded, tiny, or blurred but still contains class evidence, and poor when the crop is too ambiguous to identify from visual evidence. Output only one JSON object with the required fields.

The required JSON fields are: `nuScenes_class`, `class_confidence`, `possible_confusions`, `subtype`, `attributes`, `motion_likelihood`, `visibility`, `visual_reliability`, and `reason`. The attribute keys are fixed to: vehicle-like, large vehicle, long body, box-shaped object, two-wheeler, vulnerable road user, traffic-control object, construction-like object, static obstacle, and thin vertical object. Each attribute is binary and must be supported by visible evidence.

JSON validation and audit statistics. We parse only valid JSON objects and canonicalize class names to the 10 object classes used for object mIoU. Class confidence and reliability are clipped to $[0, 1]$; malformed confusion entries, invalid fields, or missing attributes are replaced by conservative defaults. The full blind audit covers all 700 nuScenes training scenes and produces 51,420 audited 3D-box instances, including 49,568 instances from the 10 evaluated object classes. The parse-valid rate after JSON extraction and validation is 100.0%. The merged audit file is generated once offline and reused for all VISA training runs.

Blind audit quality diagnostic. Because the VLM never receives the ground-truth class label, we can directly evaluate its closed-set visual agreement with nuScenes annotations. Overall, the blind audit achieves 87.7% top-1 agreement with nuScenes labels. Agreement is strongly correlated with the VLM reliability estimate: clear crops have 92.9% agreement, partial crops have 88.6% agreement, and poor crops drop to 22.9% agreement. The corresponding mean confidence/reliability values are $0.95/0.96$ for clear crops, $0.66/0.60$ for partial crops, and $0.18/0.14$ for poor crops. This supports using the VLM as a reliability-weighted semantic auditor rather than an oracle pseudo-labeler.

Voxel grounding and filtering. During training, each audit is attached to the matching physical object using the nuScenes instance token when available; otherwise it is matched by scene, frame index, and box index. VISA applies audit supervision only to voxels inside the corresponding oriented 3D box whose dense occupancy label matches the object class. We enlarge the box by a factor of 1.05 to tolerate discretization error. For taxonomy supervision, we keep audits with visual reliability at least 0.20 and class confidence at least 0.20. For attribute and scene-graph supervision, we keep audits with visual reliability at least 0.20. We require at least 4 valid object voxels per box, supervise at most 80 boxes per frame for taxonomy and attribute losses, at most 60 boxes per frame for the scene graph loss, sample at most 4096 voxels per box for instance losses, and at most 2048 voxels per box for graph loss.

Training losses. VISA is added to the original occupancy objective as auxiliary training supervision. The taxonomy loss constructs a soft target

$$\tilde{\mathbf{y}}_i = (1 - \alpha)\mathbf{e}_{y_i} + \alpha \mathbf{q}_i,$$

Table 5: Robustness to object grounding quality on GaussianWorld. All variants use the same offline VLM audits and training schedule.

Grounding setting	IoU	mIoU	Obj.	Rare
Full VISA: GT boxes/tracks	33.61	21.91	19.16	16.79
w/o track broadcast	32.81	21.34	18.42	15.92
Jittered boxes	33.25	21.67	18.81	16.34
w/o voxel-label filtering	32.07	21.21	17.98	15.47

where $\mathbf{e}_{y_i}$ is the closed-set occupancy label, $\mathbf{q}_i$ is the normalized VLM class/confusion distribution, and $\alpha = 0.20$. The taxonomy loss weight is 0.03. The attribute loss maps predicted object-class probabilities to the fixed attribute vocabulary and applies binary cross-entropy to the VLM attribute vector with weight 0.02. The scene-level audit graph builds pairwise relations between audited objects in the same frame: pairs are positive if their attribute Jaccard similarity is at least 0.50 or their VLM taxonomy/confusion sets are compatible, and negative if their attribute Jaccard similarity is at most 0.05 and the taxonomy sets are incompatible. The negative similarity margin is 0.20, the graph loss weight is 0.005, and at most 512 object pairs are sampled per frame. All VISA losses are applied only during training.

Offline cost and inference. The full audit generation is embarrassingly parallel across crops. In our implementation, Qwen3-VL-8B is run with greedy decoding and up to 340 generated tokens per crop. The audit bank covers 51,420 training 3D-box instances and is generated once offline before occupancy training. At deployment, VISA introduces no additional inputs, parameters, VLM calls, text encoders, crop extraction, object tracks, or latency; inference is exactly the original occupancy world model.

Experimental runs and comparison protocol. Some appendix results differ from the main-paper results because they were obtained from separate training runs. Within each table, methods follow a common training and evaluation protocol, with the factors under investigation varied as specified. Comparisons should therefore be made within each table; numerical differences across tables should not be interpreted as effects of individual components.

C More Ablation Studies

Robustness to object grounding quality. Table 5 studies how VISA depends on the quality of object-level grounding. The full model uses GT 3D boxes, instance tracks, and voxel-label filtering to attach each blind VLM audit to matched object voxels. Removing track broadcast reduces mIoU from 21.91 to 21.34 and rare mIoU from 16.79 to 15.92, showing that propagating an audit along the same physical object track improves supervision coverage, especially for long-tail classes. When the 3D boxes used for VISA grounding are jittered with moderate noise, performance remains close to the full model, with mIoU decreasing by only 0.24 and rare mIoU by 0.45. This suggests that VISA is not overly sensitive to small box localization errors. Finally, removing voxel-label filtering gives the largest degradation in object and rare-class metrics. This confirms that restricting VLM supervision to voxels whose dense label matches the audited object class prevents semantic leakage from nearby stuff regions or neighboring objects inside the 3D box. Overall, the results show that VISA benefits from accurate object grounding, but remains robust to moderate box noise, while track broadcast and matched-voxel filtering are important for clean and effective supervision.

Audit reliability. Table 6 first evaluates the reliability mechanism in VISA. Removing reliability weighting consistently reduces performance, lowering mIoU from 21.91 to 21.63 and rare mIoU from 16.79 to 16.35. This indicates that the VLM's confidence and visual reliability scores provide useful information for modulating the strength of audit supervision. Removing reliability filtering causes a larger degradation, especially on object mIoU, which drops from 19.16 to 17.10. This suggests that low-quality or ambiguous audits can introduce harmful semantic noise if they are allowed to supervise voxel logits. Thus, reliability is not merely a diagnostic output of the VLM; it is an important part of making VLM supervision robust.

Table 6: Ablation of audit reliability and crop selection on GaussianWorld. All variants use the same training schedule and offline VLM audit format.

Variant	IoU	mIoU	Obj.	Rare
VISA full	**33.61**	**21.91**	**19.16**	**16.79**
w/o reliability weighting	33.35	21.63	18.84	16.35
w/o reliability filtering	31.90	20.81	17.10	15.51
Random visible crop	32.08	21.32	18.23	15.68
Largest-area crop	33.48	21.74	17.98	16.53

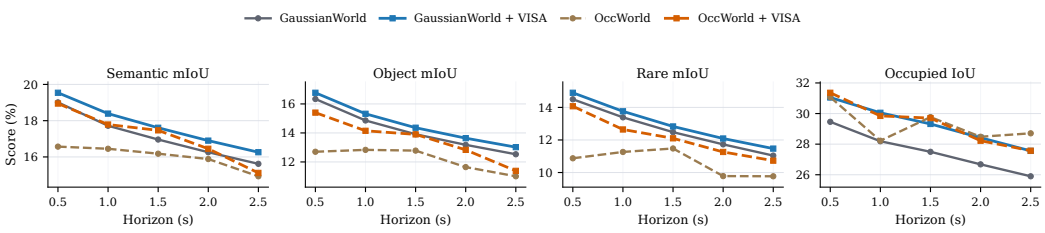


Figure 6: Ego-motion-conditioned world-state rollout on nuScenes validation. Predicted occupancy states are propagated to future ego frames using known future ego poses. VISA consistently improves rollout quality on both GaussianWorld and OccWorld without using future images.

Crop selection. Table 6 also studies how the choice of audited crop affects training. Using a random visible crop degrades IoU and rare mIoU, showing that arbitrary views can produce noisier audits and weaker supervision. The largest-area crop baseline recovers much of the overall mIoU and rare-class performance, but still underperforms full VISA in object mIoU. This suggests that visible area is an important factor for audit quality, but area alone is not sufficient: the full crop-selection strategy also accounts for occlusion and representative visibility, which improves object-level semantic supervision.

D Ego-Motion-Conditioned World-State Rollout

Beyond current-frame occupancy, we evaluate whether the semantic world state predicted at time t remains useful after propagation to future ego frames. Following the rollout setting of occupancy world models, we transform the predicted semantic occupancy state using known future ego poses and compare the propagated state with future occupancy labels. Figure 6 shows that VISA improves rollout quality on both GaussianWorld and OccWorld. The gains are consistent across semantic mIoU, object mIoU, rare-class mIoU, and occupied IoU, indicating that instance-level VLM auditing improves the semantic quality of the world state before rollout. The improvement is most pronounced for object and rare-class metrics, matching the object-centric design of VISA.

E Future Occupancy Forecasting Diagnostics

We further evaluate how the semantic occupancy representations learned with VISA affect future occupancy forecasting. This experiment is intended as an additional temporal diagnostic beyond single-frame occupancy evaluation. Given a stream occupancy backbone, we train the same lightweight forecasting-head architecture to predict future occupancy states from past predicted occupancy states. The forecasting head is trained on predicted states rather than ground-truth occupancy, so the evaluation reflects the compounded errors of the perception backbone and the forecasting module. At test time, we autoregressively roll out predictions from 0.5s to 2.5s and report semantic mIoU, object mIoU, rare-class mIoU, and occupied-space IoU on the nuScenes validation set. All methods use the same forecasting-head training protocol, validation split, horizons, and metric definitions. All metrics are reported in percentages.

Figure 7 provides a qualitative example of the same forecasting setting. In this challenging urban driving scene, VISA produces visibly sharper future occupancy forecasts, with cleaner foreground regions and more coherent semantic boundaries over time. The improvement is consistent across the displayed horizons: both GaussianWorld+VISA and OccWorld+VISA achieve higher semantic and

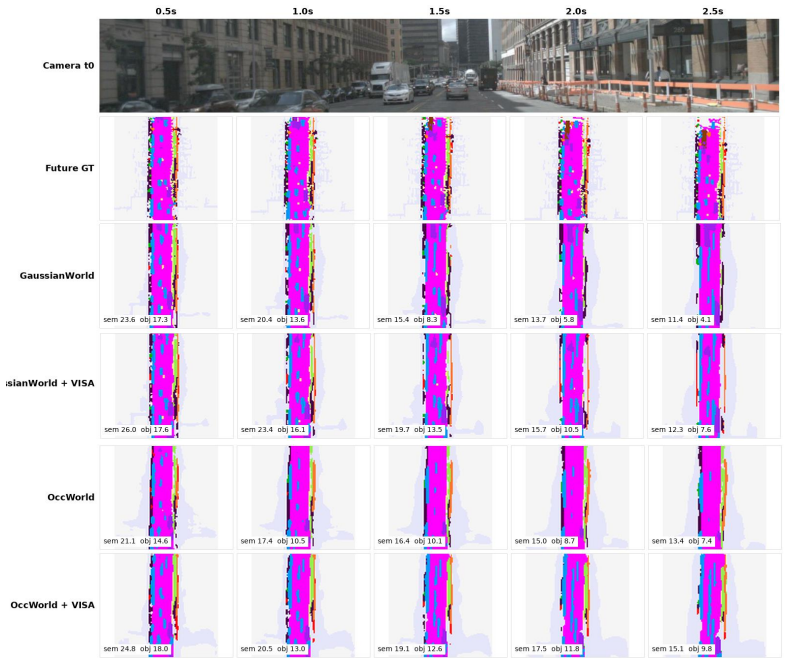


Figure 7: Qualitative future occupancy forecasting visualization on nuScenes validation. The columns show forecasting horizons from 0.5s to 2.5s, and rows compare future ground truth with predictions from GaussianWorld, GaussianWorld+VISA, OccWorld, and OccWorld+VISA. The numbers under each prediction denote semantic and object scores for the displayed forecast. In this representative sequence, VISA preserves sharper semantic boundaries and cleaner object-centric foreground regions over longer horizons.

object scores than their corresponding backbones from 0.5s to 2.5s. This visualization highlights the benefit of injecting instance-level semantic structure into the occupancy representation before autoregressive future prediction.

Table 7 shows that the effect of VISA on future occupancy forecasting is metric- and backbone-dependent. For GaussianWorld, VISA improves semantic mIoU and occupied-space IoU across all horizons. The gains become larger at longer horizons: at 2.5s, GaussianWorld+VISA improves semantic mIoU by 1.31 points and occupied-space IoU by 3.63 points. Object and rare-class forecasting are slightly lower at short horizons, but improve at longer horizons. This suggests that, for GaussianWorld, the occupancy states learned with VISA provide a more favorable input representation for long-horizon forecasting.

For OccWorld, VISA shows a different pattern. It improves object mIoU at every horizon, with gains between 0.80 and 1.28 points, and also improves rare-class mIoU across all horizons. This is consistent with the object-centric design of VISA and indicates that the learned representation helps preserve object and rare-class semantic information during future prediction. However, occupied-space IoU decreases at longer horizons for OccWorld+VISA. This indicates that improvements in object-level semantic discrimination do not necessarily translate into improved coarse occupied support under autoregressive forecasting.

Overall, these results show that VISA can provide benefits beyond static occupancy evaluation, but the effect is not uniform across backbones or metrics. On GaussianWorld, the benefits are strongest for long-horizon semantic and geometric forecasting. On OccWorld, the benefits are concentrated on object and rare-class forecasting, while coarse geometric occupancy becomes worse at longer horizons. This behavior is consistent with VISA's design: it injects instance- and semantic-level structure into occupancy representations, which is most directly useful for preserving object-class

Table 7: Future occupancy forecasting diagnostics from 0.5s to 2.5s on nuScenes validation. Sem., Obj., Rare, and Occ. IoU denote semantic mIoU, object mIoU, rare-class mIoU, and occupied-space IoU, respectively. All metrics are reported in percentages; values in parentheses indicate percentage-point changes relative to the corresponding reproduced backbone. Bold numbers indicate the better result within each backbone pair at the same horizon.

Horizon	Method	Sem. (%) ↑	Obj. (%) ↑	Rare (%) ↑	Occ. IoU (%) ↑
0.5s	GaussianWorld	21.03	**15.60**	**10.12**	37.89
	GaussianWorld + VISA	**21.07** (+0.04)	15.48 (-0.12)	9.89 (-0.23)	**38.56** (+0.67)
	OccWorld	15.74	11.29	6.28	32.37
	OccWorld + VISA	**16.87** (+1.13)	**12.57** (+1.28)	**7.28** (+1.00)	**32.67** (+0.30)
1.0s	GaussianWorld	18.54	**11.94**	**7.01**	36.17
	GaussianWorld + VISA	**18.68** (+0.14)	11.51 (-0.43)	6.89 (-0.12)	**37.42** (+1.25)
	OccWorld	14.17	8.57	4.30	**32.16**
	OccWorld + VISA	**14.97** (+0.80)	**9.54** (+0.97)	**4.80** (+0.50)	32.11 (-0.05)
1.5s	GaussianWorld	16.27	**9.37**	5.02	33.54
	GaussianWorld + VISA	**16.75** (+0.48)	9.04 (-0.33)	**5.32** (+0.30)	**35.75** (+2.21)
	OccWorld	12.78	6.78	3.11	**31.42**
	OccWorld + VISA	**13.19** (+0.41)	**7.58** (+0.80)	**3.43** (+0.32)	30.35 (-1.07)
2.0s	GaussianWorld	14.22	7.34	3.58	30.89
	GaussianWorld + VISA	**15.17** (+0.95)	**7.41** (+0.07)	**4.24** (+0.66)	**33.93** (+3.04)
	OccWorld	11.52	5.43	2.11	**30.41**
	OccWorld + VISA	**11.67** (+0.15)	**6.28** (+0.85)	**2.63** (+0.52)	28.23 (-2.18)
2.5s	GaussianWorld	12.46	5.89	2.80	28.51
	GaussianWorld + VISA	**13.77** (+1.31)	**6.11** (+0.22)	**3.39** (+0.59)	**32.14** (+3.63)
	OccWorld	**10.34**	4.36	1.37	**29.24**
	OccWorld + VISA	10.31 (-0.03)	**5.20** (+0.84)	**1.92** (+0.55)	26.15 (-3.09)

information over time, but it does not directly optimize future free-space geometry or coarse occupied support.

F Open-Loop Planning and Near-Field Semantic Diagnostics

We further evaluate whether the semantic occupancy representations learned with VISA benefit downstream open-loop planning diagnostics and near-field semantic perception. This experiment is intended as a diagnostic rather than a closed-loop safety evaluation. For open-loop planning, we use the same downstream planning protocol for each reproduced backbone and train the same lightweight planning-head architecture. All methods follow the same planning-head training and evaluation protocol. We report trajectory L2 error in meters at the 1s, 2s, and 3s horizons, as well as the average L2 across horizons. For transparency, we also report the mean instantaneous box-collision rate over the same horizons.

To connect the planning diagnostic with semantic occupancy quality, we report near-field object mIoU and rare-class mIoU within 20m of the ego vehicle. These near-field metrics are computed directly from the semantic occupancy predictions and are reported in percentages. All near-field evaluations use the same validation scenes, frame count, and metric definitions.

Table 8 shows that VISA reduces open-loop trajectory L2 for both reproduced occupancy world-model backbones under the same downstream planning protocol. For GaussianWorld, VISA reduces average L2 from 1.500 to 1.426. The improvement is consistent across all horizons, with reductions of 0.097m at 1s, 0.073m at 2s, and 0.051m at 3s. For OccWorld, VISA reduces average L2 from 1.472 to 1.438, with smaller but consistent gains across the three horizons. These results suggest that the semantic occupancy states learned with VISA can improve expert-trajectory imitation in an open-loop planning setting.

Table 8: Open-loop planning and near-field semantic occupancy diagnostics on nuScenes validation. L2 errors are reported in meters at the 1s, 2s, and 3s horizons, and Avg. L2 denotes their average. BoxCol is the mean instantaneous box-collision rate over the same horizons. Near-field metrics are computed within 20m of the ego vehicle and reported in percentages. Values in parentheses indicate changes relative to the corresponding backbone; for L2 columns, changes are in meters, and for percentage metrics, changes are in percentage points. Bold numbers indicate the better result within each backbone pair.

Method	L2@1s ↓	L2@2s ↓	L2@3s ↓	Avg. L2 ↓	BoxCol (%) ↓	Obj.@20 (%) ↑	Rare@20 (%) ↑
GaussianWorld	0.956	1.479	2.064	1.500	**4.79**	28.40	24.93
GaussianWorld + VISA	**0.859** (-0.097)	**1.406** (-0.073)	**2.013** (-0.051)	**1.426** (-0.074)	4.99 (+0.20)	**28.41** (+0.01)	**25.19** (+0.26)
OccWorld	0.882	1.462	2.072	1.472	**4.28**	21.12	17.69
OccWorld + VISA	**0.842** (-0.040)	**1.425** (-0.037)	**2.048** (-0.024)	**1.438** (-0.034)	4.97 (+0.69)	**22.78** (+1.66)	**18.89** (+1.20)

Table 9: VLM audit quality on the rare classes used for rare-class mIoU. Top-1 agreement measures whether the VLM's predicted class matches the nuScenes annotation. In-conf. agreement measures whether the ground-truth class is either the top prediction or appears in the VLM-proposed confusion set. Conf. and Rel. denote the average VLM-reported class confidence and visual reliability.

Class	# Audits	Top-1 (%)	In-conf. (%)	Conf.	Rel.
Bicycle	574	67.3	76.9	0.66	0.63
Bus	523	76.7	93.3	0.76	0.72
Constr. vehicle	529	60.0	95.0	0.86	0.79
Motorcycle	605	86.3	88.2	0.62	0.61
Traffic cone	4617	91.0	97.1	0.78	0.73
Trailer	917	48.3	79.3	0.84	0.83

The near-field semantic diagnostics show a consistent object-centric trend. VISA improves object mIoU within 20m for both backbones, with a small gain on GaussianWorld and a larger gain on OccWorld. It also improves rare-class mIoU within 20m for both backbones. These gains are aligned with the design of VISA, which provides training-time supervision on matched object voxels and is intended to improve object and rare-class semantic quality in planning-relevant regions.

The box-collision diagnostic shows a different trend: mean BoxCol does not improve under this open-loop planning protocol. This reflects the difference between expert-trajectory imitation and collision-aware planning. L2 measures similarity to logged trajectories, whereas box collision depends on footprint clearance, geometric support, and planner safety costs. We therefore interpret this section as an open-loop planning and near-field semantic diagnostic rather than a closed-loop safety evaluation. VISA does not directly optimize planner collision cost, clearance margins, or closed-loop control behavior. The results indicate that VISA improves open-loop trajectory imitation and near-field object-centric semantic occupancy, but should not be interpreted as evidence of improved collision avoidance or closed-loop planning safety.

G Additional VLM Audit Diagnostics

VISA uses an offline VLM auditor to provide training-time instance-level semantic structure. Since this supervision is derived from a vision-language model, we further analyze the quality and ambiguity patterns of the audit signal used during training. For each object crop, the auditor returns a structured response containing a closed-set class hypothesis, plausible confusion classes, visual attributes, class confidence, visual reliability, and short evidence. We report audit agreement with the nuScenes class annotation, as well as reliability-binned audit quality. These diagnostics are intended to characterize the training audit bank used by VISA, rather than to benchmark the VLM as a standalone classifier.

Table 9 shows that the audit signal is informative but not noise-free. For visually distinctive categories such as traffic cone and motorcycle, the VLM often predicts the correct closed-set class. For more ambiguous rare classes such as trailer and construction vehicle, top-1 agreement is lower, but

Table 10: Audit quality as a function of VLM-reported visual reliability over the full training audit set. Reliability is not treated as a calibrated probability, but higher reliability bins correlate with higher audit agreement.

Reliability bin	# Audits	Top-1 (%)	Avg. confidence
$[0.0, 0.2)$	801	29.2	0.10
$[0.2, 0.4)$	3153	79.9	0.29
$[0.4, 0.6)$	8989	82.9	0.30
$[0.6, 0.8)$	21198	90.4	0.76
$[0.8, 1.0]$	17279	91.1	0.94

Table 11: Representative structured VLM audit examples from the training audit set. Ambiguous cases often involve visually similar object categories, but the ground-truth class is frequently preserved in the VLM confusion set or reflected in the visual attributes.

Type	GT	VLM	Confid.	Rel.	Notes
Correct	Constr. vehicle	Constr. vehicle	0.98	0.99	construction-like, large vehicle
Correct	Bicycle	Bicycle	0.95	0.98	two-wheeler, vulnerable road user
Correct	Traffic cone	Traffic cone	0.95	0.98	traffic-control, thin obstacle
Ambiguous	Constr. vehicle	Truck	0.98	0.98	GT appears in confusion set
Ambiguous	Trailer	Truck	0.95	0.98	long-body vehicle ambiguity
Ambiguous	Traffic cone	Barrier	0.95	0.98	traffic-control obstacle ambiguity

the ground-truth class frequently appears in the proposed confusion set. This is important for VISA because the audit is not used as a hard replacement label. Instead, VISA uses the full structured audit, including plausible confusions, visual attributes, confidence, and reliability, as soft object-centric supervision anchored by the original occupancy labels. The confidence and reliability values should therefore be interpreted as audit metadata for weighting and filtering, not as guarantees that the top-1 VLM class is semantically correct.

Table 10 supports the use of reliability filtering and weighting in VISA. Very low-reliability crops have poor agreement, while crops with reliability above 0.6 achieve over 90% top-1 agreement. Reliability is therefore useful as a training-time visual-quality signal, even though it should not be interpreted as a calibrated probability or as a guarantee of semantic correctness. This distinction matters because a crop can be visually clear and thus receive high reliability, while still being semantically ambiguous under the fine-grained closed-set driving taxonomy. In the full training audit bank, we filter or downweight low-quality signals using both class-confidence and visual-reliability thresholds.

Table 11 illustrates why VISA uses structured audits rather than single hard pseudo-labels. The dominant audit errors are not random class assignments, but structured confusions among visually related driving categories, such as trailer versus truck, construction vehicle versus truck, and traffic cone versus barrier. Some of these ambiguous cases can still have high visual reliability because the crop itself is clear, even though the fine-grained closed-set category remains ambiguous. This motivates the VISA design: reliability is used to filter poor visual evidence, while the confusion set and attribute factors capture uncertainty among semantically related classes. The audit signal is used only during training as auxiliary supervision; inference remains VLM-free.

H Audit Coverage and Computational Cost

We examine audit coverage, offline generation cost, and student training overhead on Gaussian-World. Table 12 summarizes these complementary aspects of the computational requirements of VISA.

Offline audit-generation setup. We generate structured audits using Qwen3-VL-8B-Instruct on four nodes, each equipped with four NVIDIA RTX A6000 GPUs, for a total of 16 GPUs. Each GPU

Table 12: Audit coverage and computational cost on GaussianWorld. (a) Nested, class-stratified audit subsets with an unchanged student training schedule. (b) Offline audit generation cost on NVIDIA RTX A6000 GPUs; storage is reported in MB. (c) Student training overhead under matched hardware and batch settings.

(a) Audit coverage				**(b) Generation cost**				**(c) Student overhead**			
Share	mIoU	Obj.	Rare	Share	Queries	GPU-h	MB	Variant	mIoU	ms/step	GB
0%	21.37	18.18	15.60	0%	0	0.0	0.0	Occupancy only	21.37	1,071	22.38
25%	21.53	18.50	15.96	25%	12,855	18.1	23.6	Crop teacher	21.64	1,088	22.53
50%	21.71	18.82	16.38	50%	25,710	35.7	46.9	Tax.+Attr.	21.74	1,111	22.70
100%	**21.91**	**19.16**	**16.79**	100%	51,420	71.8	94.1	Full VISA	**21.91**	1,130	22.85

Inference overhead: VISA introduces no additional inference modules and makes no VLM calls at inference time.

processes an independent crop shard using greedy decoding with at most 340 new tokens per crop. Audits are generated from training images only, validated, and cached in JSONL format for reuse during occupancy training.

Audit coverage. The full audit bank contains 51,420 records. We construct nested, class-stratified subsets containing 25%, 50%, and 100% of this bank while keeping the student training schedule unchanged. The 0% setting uses occupancy supervision without audit losses. Panel (a) reports semantic mIoU, object mIoU, and rare-class mIoU at each coverage level.

Offline generation cost. Panel (b) reports the VLM query count, generation GPU-hours, and JSONL storage for each audit budget. Generation is parallelized across independent crop shards. The resulting audit bank is constructed before student training and can be reused across training runs, amortizing the preprocessing cost. Offline generation is accounted for separately from student training.

Student training overhead. Panel (c) compares occupancy-only training, crop-teacher supervision, taxonomy-plus-attribute supervision, and Full VISA under matched hardware and batch settings. We report training step time and peak GPU memory. During VISA training, cached audit records provide targets for the auxiliary objectives; the VLM itself is not loaded or queried in the student training loop.

Inference cost. The audit bank and auxiliary objectives are used only during training. At inference, VISA retains the original occupancy architecture and input pipeline, requiring no additional VLM calls, text encoders, crop extraction, or object tracking.

I Failure Cases and Boundary Conditions

We provide qualitative failure cases and boundary conditions to clarify what VISA does and does not address. These examples are intentionally selected stress cases rather than typical outcomes. They are meant to illustrate the operating limits of VLM-guided semantic auditing and structured occupancy supervision. VISA uses offline VLM audits as soft training-time semantic guidance, rather than treating the VLM as an oracle or adding inference-time reasoning modules. The examples highlight two complementary sources of residual error: ambiguity in crop-level audit signals and local errors in spatially coherent semantic occupancy prediction.

Figure 8 shows representative failure modes of the offline VLM audit. In the first example, the target instance is a concrete barrier, but a salient traffic cone appears in the foreground; the VLM focuses on the foreground object and predicts *traffic cone* instead of *barrier*. In the second example, the target truck is partially visible behind a fence, and the surrounding sign and construction-like context bias the audit toward *construction vehicle*. The third example is a far-range, low-resolution crop, where the vehicle shape is too blurred to reliably distinguish *bus* from *truck*. The fourth example contains

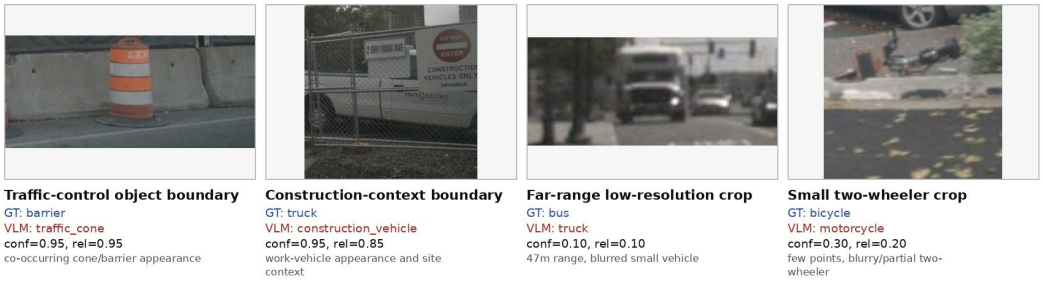


Figure 8: Representative VLM audit boundary cases from the full training audit set. The examples illustrate foreground distractors, contextual bias, low-resolution evidence, and viewpoint ambiguity. We report the VLM class confidence and visual reliability for each crop.

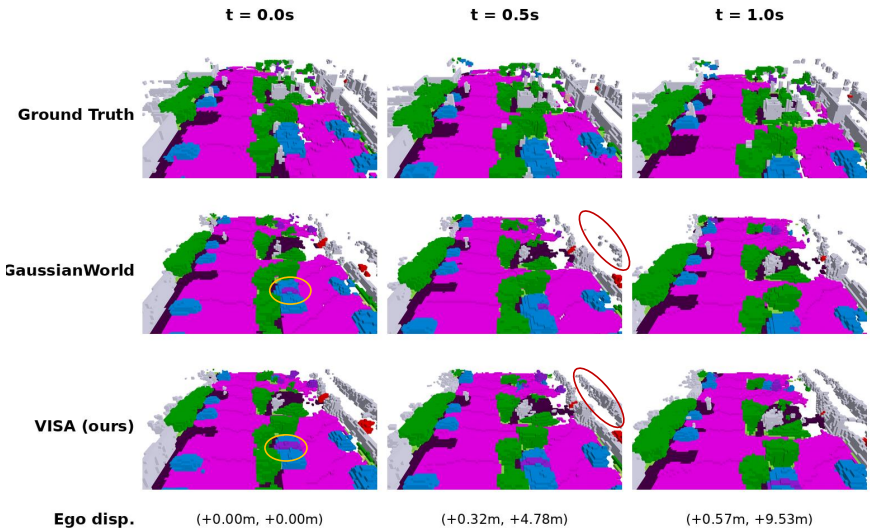


Figure 9: Representative boundary case of structured occupancy completion on nuScenes validation. Red circles highlight over-completed regions where the prediction fills spatially coherent occupancy that is absent from the ground truth. Yellow circles highlight cases where a local semantic misclassification is propagated to nearby voxels due to the same structured completion behavior. Highlighted regions are representative rather than exhaustive.

a fallen two-wheeler viewed from above, where the visible evidence indicates a two-wheeler but is insufficient to reliably separate *bicycle* from *motorcycle*.

These cases show that best-view crop selection improves visibility but cannot completely remove foreground distractors, contextual bias, far-range resolution limits, or viewpoint ambiguity. This is why VISA does not use VLM predictions as hard pseudo-labels. Instead, each audit is used as a soft, reliability-aware training signal, anchored by the original dense occupancy supervision and restricted to matched object voxels. Low-confidence or low-reliability audits are filtered or weakly weighted, while even reliable audits remain auxiliary guidance rather than replacements for ground-truth occupancy labels.

Figure 9 shows a boundary case of spatially coherent semantic occupancy prediction. VISA often produces cleaner and more coherent object or region-level predictions, which is beneficial when the local semantic evidence is correct. However, in locally ambiguous regions, this same coherence can make an incorrect local hypothesis appear as a more contiguous false-positive region, as highlighted by the red circles. Similarly, when a small region receives an incorrect semantic label, the

prediction can become spatially more coherent while remaining semantically incorrect over a local neighborhood, as highlighted by the yellow circles.

This behavior reflects a boundary of consistency-oriented semantic supervision rather than an additional inference-time failure mode. VISA encourages object-centric semantic consistency, but it does not directly solve all sources of geometric uncertainty, stuff-region ambiguity, or local semantic noise. Thus, when the underlying local evidence is incorrect or incomplete, spatial coherence can improve visual consistency without necessarily correcting the underlying semantic error.

Overall, these stress cases help explain the remaining errors observed despite the quantitative gains reported in the main paper. They clarify the regime in which VISA is most useful: improving object-centric semantic occupancy supervision, especially for rare and visually confusable classes, while preserving the original inference architecture. The method does not fully resolve ambiguous crop-level grounding, contextual VLM bias, low-resolution evidence, viewpoint ambiguity, or local errors caused by incomplete geometric evidence. These limitations motivate future work on uncertainty-aware audit selection, stronger geometric consistency checks, and extending semantic auditing beyond visible object instances to stuff regions and free-space structure.