OASIS: An Open-dimension Assessment System for Image Scoring via Definition-Driven Visual Evidence Learning

Anonymous authors
Paper under double-blind review

Abstract

Image filtering and vision-system monitoring require application-specific criteria that may change after an assessment model has been trained. However, models trained for predefined assessment objectives offer limited flexibility when new criteria are introduced. To address this limitation, we study open-dimension image scoring, in which a model predicts a continuous score from an image and a natural-language evaluation definition. The focus is on criteria not used as scoring targets during task-specific training, with model parameters fixed at inference. To this end, we propose OASIS (Open-dimension Assessment System for Image Scoring), a two-stage framework that connects evaluation definitions, criterion-relevant visual evidence and continuous scores. Specifically, definition-conditioned descriptive supervision is first used to guide the model toward visual evidence relevant to the specified criterion. The model is then trained for scoring with semantics-preserving definition augmentation and ordinal distribution learning, which encourage reliance on definition content and account for the ordering and distance between score levels. To support training and evaluation, we construct MD61, comprising 40 base dimensions with scores and evidence descriptions curated through a Score-and-Rank Verification Process, and 21 held-out dimensions annotated by five computer vision experts. Experimental results on the held-out dimensions show that OASIS outperforms evaluated baselines, such as conventional and language-driven image quality assessment (IQA) methods, and achieves an overall PLCC/SRCC of 0.829/0.823. Further experiments demonstrate that its scores correlate with downstream vision performance under controlled image degradations and that multidimensional supervision improves cross-dataset transfer on conventional IQA benchmarks. These results support the feasibility of adapting image assessment to newly specified criteria through natural-language definitions without retraining.

1 Introduction

The suitability of an image depends on task-specific visual requirements. For example, an aerial image suitable for region segmentation may fail for road extraction, making a single overall quality score inadequate. Such differences are critical in training-data filtering and vision-system monitoring, where data selection directly affects downstream performance (Gadre et al., 2023). Furthermore, evaluation criteria may also change as applications evolve, even when the underlying images remain unchanged. Additional annotations or model adaptation may then be required, if the evaluator is tied to predefined objectives.

Visual scoring has progressed from single-criterion assessment to multi-attribute evaluation, with traditional IQA focusing on overall quality or predefined dimensions (Mittal et al., 2012; Su et al., 2020; Ke et al., 2021; Chen et al., 2024). Recent language-driven approaches and vision-language reward models have further extended this capability to broader visual tasks (Wang et al., 2023; Wu et al., 2024a; You et al., 2024; Wu et al., 2024b; You et al., 2025; Li et al., 2025b; Xu et al., 2023; Wang et al., 2025). These advances provide useful foundations for language-based assessment. Nevertheless, these models struggle when prac-

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.

tical applications introduce new requirements as natural-language definitions rather than predefined training tasks. This raises our central question: can a shared model produce meaningful continuous scores under a newly defined criterion without retraining?

To answer this question, we study open-dimension image scoring. Given an image x and a natural-language evaluation definition d, the model learns a conditional function $q_\theta(x, d) \in [0, 100]$ which produces a score according to d while keeping its parameters fixed. The model is specifically expected to generalize to unseen criteria during training. Thus, the evaluation target is determined by the input definition rather than restricted to predefined dimensions. Figure 1 illustrates how different dimensions specify distinct assessment targets for the same image.

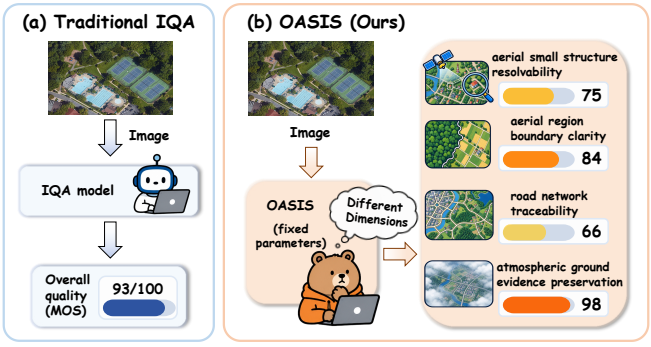


Figure 1: Open-dimension Assessment System for Image Scoring (OASIS) outputs customizable, continuous scores for multiple dimensions based on images and evaluation criteria, overcoming traditional IQA models' limitations.

The challenge of open-dimension scoring is to establish a mapping from the defined text to a continuous score. The model must identify the corresponding scoring direction and scale to complete the quantification. General multimodal large language models (MLLMs) remain limited in fine-grained visual judgment and numerical scoring (Wu et al., 2024c), making this challenge particularly demanding. Meanwhile, existing scoring datasets also tend to focus on fixed aspects, lacking annotations that simultaneously contain multi-dimensional definitions and continuous scores.

Our main contributions are summarized as follows:

1. We formulate open-dimension image scoring, in which a shared model produces continuous scores for criteria unseen during training based on natural-language definitions, without parameter updates.
2. We propose OASIS, a two-stage framework in which definition-conditioned descriptive supervision is followed by continuous scoring with definition augmentation and ordinal distribution learning. Its effectiveness is demonstrated through agreement with expert judgments, alignment with downstream-task performance, and transfer to conventional IQA benchmarks.
3. We construct MD61, which provides multidimensional scores and definition-relevant visual-evidence descriptions for 40 base dimensions, together with an expert-annotated benchmark for 21 held-out dimensions. SRVP is introduced to curate the synthetic training annotations.

2 Related Work

2.1 Perceptual Image Quality Assessment

Conventional perceptual IQA models human judgments using structural similarity, deep feature distances, and pairwise preferences (Wang et al., 2004; Zhang et al., 2018; Prashnani et al., 2018). No-reference methods combine statistical, distributional, and learned representations (Mittal et al., 2012; Talebi & Milanfar, 2018; Zhang et al., 2020; Madhusudana et al.,

2022), with advances in adaptive, multiscale, attention-based, and semantic modeling (Su et al., 2020; Ke et al., 2021; Yang et al., 2022; Golestaneh et al., 2022; Xu et al., 2024; Chen et al., 2024). SPAQ and KonIQ-10k provide annotations for authentic distortions (Fang et al., 2020; Hosu et al., 2020). However, these methods mainly predict predefined targets, such as the overall Mean Opinion Score (MOS) or a fixed set of attributes. In contrast, we use the evaluation definition as an explicit input, enabling the scoring objective to be dynamically specified by its content. Conventional IQA is therefore a special case of our framework when the evaluation definition d is fixed.

2.2 Language-Driven Image Quality Assessment

Vision-language models enable text-based quality characterization, explanation, and scoring. Building on CLIP representations (Radford et al., 2021), CLIP-IQA (Wang et al., 2023) uses prompts to describe image appearance, while LIQE (Zhang et al., 2023b) learns quality information through vision-language correspondence. Q-Instruct (Wu et al., 2024a) and DepictQA (You et al., 2024) introduce quality-related instruction and description supervision. Q-Align (Wu et al., 2024b) and DeQA-Score (You et al., 2025) enable continuous scoring through language-defined levels and distribution modeling. The limitations of general multimodal models in fine-grained quality assessment have also motivated specialized training and evaluation (Wu et al., 2024c; Cao et al., 2025; Wu et al., 2025; Jia et al., 2026). These studies demonstrate the value of language representations and distributed scoring for IQA. Unlike existing methods that score along predefined dimensions, our key is that OASIS can accept any natural-language definition at inference time and score images on unseen dimensions during training.

2.3 Downstream Task Applicability Assessment

Overall perceptual quality does not necessarily reflect an image's utility for downstream computer vision systems. Prior work has studied how degradations such as noise, blur, and compression affect recognition accuracy (Dodge & Karam, 2016; Hendrycks & Dietterich, 2019), while real-world quality benchmarks and machine-preference datasets have extended assessment toward operational task utility (Chiu et al., 2020; Li et al., 2025a). In contrast, our approach uses structured natural-language definitions to specify task-oriented visual information requirements, rather than training bespoke evaluators for individual systems. We select 25 frozen task models across diverse domains as external-validity benchmarks to test whether definition-conditioned scores can track downstream machine responses under controlled image degradations.

3 Problem Formulation

3.1 Open-dimension Image Scoring

Let $\mathcal{X}$ denote the image space and $\mathcal{D}$ specifies the space of natural-language evaluation definitions. Fixed-dimension assessment learns a criterion-specific mapping $q_j : \mathcal{X} \to [0, 100]$ for each predefined criterion. Open-dimension scoring takes an evaluation definition explicitly as an explicit input to a shared model:

$$q_\theta : \mathcal{X} \times \mathcal{D} \to [0, 100]. \tag{1}$$

Here, θ denotes model parameters, and $d \in \mathcal{D}$ specifies the evaluation target, relevant visual evidence, and scoring guidelines. The model learns the correspondence between visual concepts (e.g., boundaries, occlusions, surface depth) and definitions during training, and dynamically composes these acquired concepts to evaluate novel definitions at inference.

The training phase leverages scoring supervision (x_i, d_j, y_{ij}) over a base dimension set D_{B}, where i and j index images and evaluation definitions, respectively, and $y_{ij} \in [0, 100]$ is the reference score. Let Ω_{score} denote valid training records. The optimization objective is:

$$\theta^\star = \arg\min_\theta \frac{1}{|\Omega_{\mathrm{score}}|} \sum_{(i,j)\in\Omega_{\mathrm{score}}} \ell_{\mathrm{score}}(\theta; x_i, d_j, y_{ij}), \qquad d_j \in \mathcal{D}_{\mathrm{B}}. \tag{2}$$

Here, ℓ_{score} is the scoring loss for an individual record, as specified in Section 5. At inference, $\theta^\star$ is fixed, and the model outputs $\widehat{y}_{ij} = q_{\theta^\star}(x_i, d_j)$ for the image and supplied definition.

Evaluation follows a strict dimension hold-out protocol:$\mathcal{D}_{\text{O}}$ denotes the set of held-out dimensions that never appeared as independent scoring objectives during training, satisfying $\mathcal{D}_{\text{B}} \cap \mathcal{D}_{\text{O}} = \varnothing$. Held-out dimensions are synthesized from visual concepts acquired during training, thereby directly assessing the model's capacity to compose learned visual evidence under novel linguistic guidance.

3.2 Structure of Evaluation Definitions

An evaluation definition uses structured natural language to specify the evaluation target, scope of applicability, positive and negative visual evidence, and scoring guidance. A dimension name identifies the target, while its full definition specifies what to examine and which visible phenomena support higher or lower scores. All dimensions share the output range $[0, 100]$, with higher scores indicating better quality under the given criterion. We focus on definitions that can be adjudicated from directly observable visual evidence.

4 Open-dimension Supervision and Evaluation

4.1 The MD61 Dataset

We construct MD61 (Multi-Dimension, 61 evaluation dimensions), comprising MD61-Base and MD61-Open. MD61-Base contains 40 training dimensions covering perceptual quality, structure, text, scene content, and task-relevant evidence, with continuous scores and definition-related descriptions. MD61-Open contains 21 held-out dimensions across autonomous driving, depth estimation, medical imaging, and remote sensing, with reference scores from five computer vision experts for evaluating definition-based scoring. Dataset statistics are reported in Table 1, and image sources are listed in Appendix C. Dimension names and basic definitions were manually specified and expanded by GPT-5.6 into unified full definitions. Table 2 illustrates how definitions specify different targets through evaluation objects, visible evidence, and scoring anchors: boundary separability assesses whether adjacent contents can be distinguished, whereas occlusion-boundary depth clarity assesses whether visible contours support depth perception and foreground assignment. Although both involve boundary evidence, their evaluation requirements differ. Complete definitions and scoring anchors are provided in Appendix G, and semantic comparisons between the held-out and closest base dimensions are given in Table G.2.

Table 1: Composition of the MD61 dataset.

Subset	Dimensions	Image entries	Valid scores
MD61-Base	40	58,616	358,221
MD61-Open	21	6,492	33,670
Total	61	65,108	391,891

4.2 SRVP Annotation and Filtering

For MD61-Base, continuous scores and visual-evidence descriptions are constructed via the Score-and-Rank Verification Process (SRVP). Three vision-language teacher models (GPT-5.6, kimi-k2.7, Claude-Haiku-4.5) independently ingest the input image alongside its full definition, predicting numerical ratings together with detailed rationales. The mean of valid teacher ratings serves as candidate score $\widetilde{s}_{ij}$.

To ensure ordinal fidelity, a Rank-Gate module verifies the relative ordering of candidate scores. The ranking panel receives only the paired images and the evaluation definition—without access to candidate scores, teacher identities, or reasoning traces—to adjudicate the relative ordering of image pairs under identical dimensions. Records exhibiting insufficient

Table 2: Exemplary MD61 evaluation definitions: evaluation targets, visible evidence, and score anchors.

Field	MD61-Base: Boundary separability	MD61-Open: Occlusion-boundary depth clarity
Target	Whether the boundary between an entity/region and adjacent content is visible, continuous, and distinguishable.	Whether visible occlusion contours support clear relative depth and foreground assignment.
Visible evidence	Edge contrast and contour continuity; local merging, occlusion, or fragmentation lowers score.	Occlusion-edge continuity and front–back relationships; blurred contours lower score.
Score anchors	90–100: consistently separable; 0–39: largely unusable.	90–100: clear relative depth at occlusion boundaries; 0–24: cues largely absent.

ranking consensus or severe high-confidence contradictions are filtered out. Retained records adopt the candidate mean as continuous score supervision. Figure 2 illustrates the pipeline.

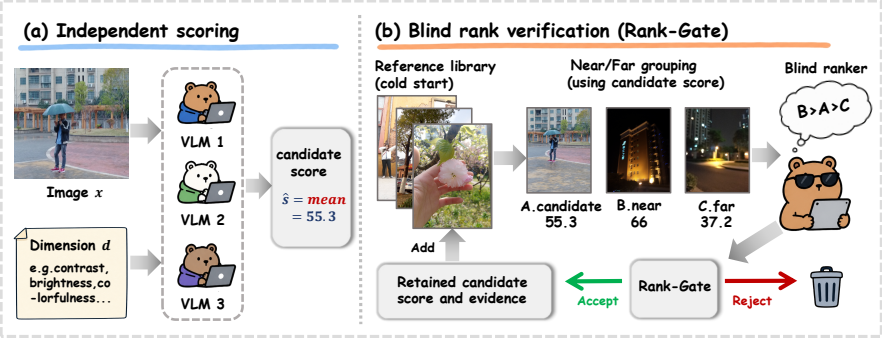


Figure 2: Overview of the SRVP annotation pipeline.

4.3 Evaluation Benchmark for Unseen Criteria

MD61-Open spans four scenarios: autonomous driving, depth estimation, medical imaging, and remote sensing. It comprises 1,082 public test/validation source images and their five-level directional degradations, totaling 6,492 images. Based on domain-specific task characteristics and quality definitions, we formulated an evaluation protocol specifying dimensions and scoring criteria. Five experts with research backgrounds in image classification, object detection, or semantic segmentation annotated the applicable image–dimension pairs. Each sample was independently evaluated by two mutually blinded experts, with a third adjudicating if inter-rater agreement fell below a prespecified threshold; the final ratings were averaged to yield the reference score y_{ij}^{H}. All images are strictly disjoint from those used for model training and validation.

5 Definition-Driven Evidence and Score Learning

OASIS establishes the mapping among evaluation definitions, visual evidence, and continuous scores via a two-stage training paradigm. In Stage 1, definition-conditioned descriptive supervision guides the model to identify visible merits, defects, and perceptual impacts under the specified criterion. In Stage 2, the acquired representations facilitate the selection of score tokens, followed by continuous score regression trained under ordinal distribution supervision. Figure 3 illustrates the architecture.

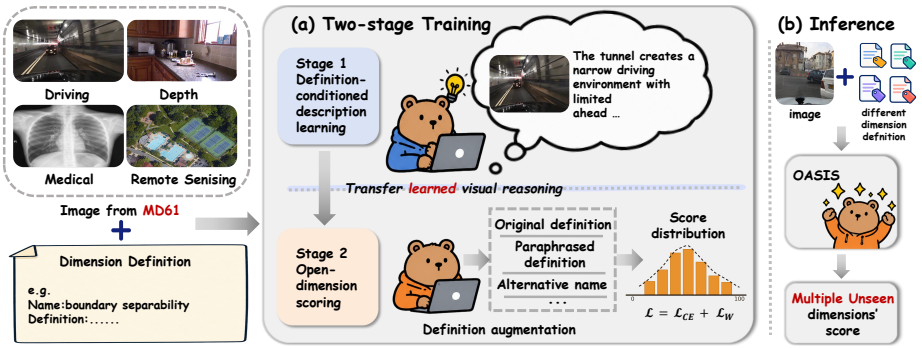


Figure 3: Two-stage training of OASIS: Stage 1 involves grounding visual evidence under descriptive supervision. Stage 2 combines definition augmentation and ordinal score distribution learning. Directly output scores based on image and definition during inference.

5.1 Definition-Conditioned Description Learning

Open-dimension scoring requires the model to modulate visual assessments according to semantic variations in evaluation definitions. Given identical images, different definitions prompt distinct target descriptions. This forces the model to adaptively attend to criterion-relevant visual features rather than memorizing a fixed image-to-score projection. The first-stage parameters ϕ are optimized by minimizing the description negative log-likelihood $\mathcal{L}_{\text{desc}}(\phi)$. Upon convergence, updated vision encoder weights are retained and language LoRA weights are merged, serving as parameter initialization θ for the scoring stage.

5.2 Definition Augmentation

To prevent the model from reducing definitions to categorical shortcuts or rigid dimension names, we apply two forms of definition augmentation during the scoring stage:

- Synonym Rewriting: Paraphrasing definition text while strictly preserving the evaluation target, applicable scope, positive/negative evidence polarity, and scoring anchors.
- Name Substitution: Masking or substituting canonical dimension names in the input prompt, retaining solely the definition body that details the criteria and scoring rubrics.

Both augmentations preserve the underlying image and target score. For training record (x_b, d_b, y_b), the augmented definition is denoted as $\widetilde{d}_b$.

5.3 Ordinal Score Modeling

Prior language-driven IQA methods demonstrate that continuous scoring can be achieved via discrete quality tokens and their probability distributions (Wu et al., 2024b; You et al., 2025; Cao et al., 2025). OASIS adopts this paradigm, bridging evidence learning and continuous regression through semantic-driven token selection and an ordinal distribution loss.

Score Token Selection and Ordinal Distribution Learning. Candidate quality tokens are ranked by their next-token logit differences across high- and low-quality contexts. 11 single-token candidates are sampled at uniform intervals:

$$\mathcal{V}_q = \{v_0, \ldots, v_{10}\} \subset \mathcal{V}, \qquad r_k = 10k, \quad k = 0, \ldots, 10. \tag{3}$$

Given continuous supervision $y_b \in [0, 100]$, adjacent-tier linear interpolation constructs soft targets:

$$k_b = \min\left(\left\lfloor \frac{y_b}{10} \right\rfloor, 9\right), \quad \alpha_b = \frac{y_b}{10} - k_b, \quad \pi_{bk} = \begin{cases} 1 - \alpha_b, & k = k_b, \\ \alpha_b, & k = k_b + 1, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

satisfying $\sum_{k=0}^{10} \pi_{bk} = 1$ and $\sum_{k=0}^{10} r_k \pi_{bk} = y_b$. We combine cross-entropy $\mathcal{L}_{\text{CE}}$ with normalized first-order Wasserstein distance $\mathcal{L}_{\text{W}}$ to enforce ordinal distance preservation:

$$\mathcal{L}_{\text{score}}(\theta) = \mathcal{L}_{\text{CE}}(\theta) + \lambda_{\text{W}} \mathcal{L}_{\text{W}}(\theta), \quad \mathcal{L}_{\text{W}} = \frac{1}{10B} \sum_{b=1}^{B} \sum_{k=0}^{9} \left| \sum_{h=0}^{k} \bar{p}_{bh} - \sum_{h=0}^{k} \pi_{bh} \right|, \tag{5}$$

where $\bar{p}_{bh} = p_b(v_h) / \sum_{m=0}^{10} p_b(v_m)$ and $\lambda_{\text{W}} = 0.5$.

6 Experiments

6.1 Experimental Setup

Open-dimension scoring models are trained on MD61-Base and evaluated on MD61-Open, with conventional IQA transfer experiments additionally incorporating KonIQ or SPAQ supervision. All setups adopt Qwen3-VL-8B-Instruct as the backbone. In Stage 1, the vision encoder and language LoRA adapters are jointly updated; the vision weights are retained, LoRA is merged, and the resulting checkpoint initializes Stage 2. Stage 2 introduces fresh LoRA adapters while continuing vision tuning. Performance is evaluated via macro-averaged Pearson (PLCC) and Spearman (SRCC) correlation coefficients across all dimensions.

6.2 Validation of Training Label Quality

We validate the SRVP pipeline on SPAQ full images across five human-annotated attributes. As shown in Table 3, aggregating three teachers improves correlation over a single teacher (SRCC 0.7964 vs. 0.7833), while Rank-Gate retains high fidelity (SRCC 0.7963, PLCC 0.8132) after filtering out anomalous records.

Table 3: Validation of SRVP labels against human attributes in SPAQ.

Method	Label retention	SRCC	PLCC
GPT-5.6, single teacher	100.00%	0.7833	0.7893
Three-teacher mean	100.00%	0.7964	0.8110
Three-teacher mean + Rank-Gate	95.99%	0.7963	0.8132

6.3 Main Results on MD61-Open

On 21 held-out dimensions (Table 4), Full Prompt (0.829/0.823) and Definition Only (0.811/0.814) confirm that explicit criteria effectively guide model evaluation.

Table 4: Correlation with expert ratings on MD61-Open (PLCC/SRCC).

Method / Input setting	Autonomous driving (6)	Depth estimation (5)	Medical imaging (5)	Remote sensing (5)	Overall (21 dims.)
Qwen3-VL-8B (direct)	0.664/0.659	0.714/0.716	0.596/0.617	0.703/0.769	0.669/0.689
DeQA (You et al., 2025)	0.613/0.631	0.754/0.746	0.588/0.587	0.640/0.694	0.647/0.663
Q-Align (Wu et al., 2024b)	0.621/0.633	0.751/0.747	0.535/0.551	0.668/0.727	0.643/0.663
RALI (Zhao et al., 2026)	0.604/0.626	0.580/0.609	0.497/0.500	0.617/0.688	0.576/0.607
Q-Insight (Li et al., 2025b)	0.651/0.656	0.746/0.689	0.614/0.613	0.720/0.752	0.681/0.677
VisualQuality-R1 (Wu et al., 2025)	0.623/0.642	0.748/0.714	0.595/0.617	0.715/0.746	0.668/0.678
OASIS: full prompt	0.832/0.835	0.849/0.813	0.810/0.799	0.826/0.844	0.829/0.823
OASIS: definition only	0.814/0.822	0.818/0.798	0.799/0.796	0.811/0.838	0.811/0.814
OASIS: name only	0.701/0.718	0.779/0.746	0.777/0.772	0.783/0.825	0.757/0.763
OASIS: shared MOS	0.624/0.660	0.781/0.755	0.719/0.721	0.726/0.786	0.708/0.727
OASIS: image only	0.462/0.498	0.607/0.636	0.462/0.250	0.555/0.560	0.518/0.487

6.4 Downstream Task Alignment

We test whether predicted scores track downstream task performance under controlled degradations. Following MPD (Li et al., 2025a), we use five open-source models per domain and their per-image metrics as task references. Table 5 reports cross-image correlations; model and metric details are provided in Appendix F.

Table 5: Correlation with downstream task performance (PLCC/SRCC). Best and second-best model results are highlighted in bold red and underlined blue, respectively.

Method / Input setting	Autonomous driving	Depth estimation	Medical imaging	Remote sensing
Qwen3-VL-8B (direct)	0.563/0.527	0.562/0.551	0.370/0.410	0.363/0.335
DeQA (You et al., 2025)	0.512/0.531	0.570/0.553	0.260/0.281	0.284/0.260
Q-Align (Wu et al., 2024b)	0.518/0.528	0.563/0.546	0.211/0.248	0.372/0.342
RALI (Zhao et al., 2026)	0.512/0.544	0.474/0.467	0.320/0.305	0.288/0.303
Q-Insight (Li et al., 2025b)	0.608/0.598	0.527/0.479	0.379/0.422	0.285/0.277
VisualQuality-R1 (Wu et al., 2025)	0.551/0.574	0.552/0.506	0.392/0.420	0.360/0.333
OASIS: full prompt	0.615/0.608	0.657/0.616	0.483/0.517	0.407/0.392
OASIS: definition only	0.601/0.599	0.634/0.613	0.473/0.507	0.387/0.370
OASIS: name only	0.581/0.581	0.605/0.566	0.455/0.493	0.359/0.344
OASIS: shared MOS	0.507/0.546	0.573/0.552	0.379/0.406	0.315/0.302
OASIS: image only	0.319/0.340	0.479/0.472	0.289/0.145	0.156/0.146
Expert reference ratings	0.600/0.586	0.567/0.510	0.433/0.459	0.393/0.396

6.5 IQA Transfer Performance

Table 6 reports cross-dataset IQA transfer performance under different training-data configurations. K, S, and M denote KonIQ-10k, SPAQ, and MD61-Base.

Table 6: Cross-dataset IQA transfer performance (PLCC/SRCC). Best and second-best results highlighted in bold red and underlined blue.

Method	Data	KonIQ	SPAQ	KADID	PIPAL	LIVEW	AGIQA	CSIQ
CLIP-IQA+ (Wang et al., 2023)	K	0.909 0.834	0.866 0.758	0.653 0.465	0.427 0.452	0.832 0.832	0.736 0.636	0.772 0.627
Q-Align (Wu et al., 2024b)	K	0.941 0.940	0.886 0.887	0.674 0.684	0.403 0.419	0.853 0.860	0.772 0.735	0.671 0.737
DeQA (You et al., 2025)	K	0.953 0.941	0.895 0.896	0.694 0.687	0.472 0.478	0.892 0.879	0.809 0.729	0.787 0.744
RALI (Zhao et al., 2026)	K	0.939 0.922	0.897 0.897	0.723 0.723	0.527 0.528	0.896 0.876	0.779 0.715	0.828 0.788
Q-Insight (Li et al., 2025b)	K	0.933 0.916	0.907 0.905	0.742 0.736	0.472 0.478	0.893 0.865	0.811 0.764	0.870 0.824
VisualQuality-R1 (Wu et al., 2025)	K	0.923 0.908	0.891 0.892	0.712 0.711	0.441 0.438	0.874 0.849	0.822 0.767	0.712 0.662
OASIS (Ours)	K	0.954 0.944	0.900 0.894	0.699 0.704	0.502 0.541	0.885 0.892	0.816 0.743	0.831 0.826
Q-Align (Wu et al., 2024b)	K+S	0.943 0.940	0.933 0.931	0.692 0.708	0.401 0.411	0.883 0.879	0.795 0.727	0.795 0.767
DeQA (You et al., 2025)	K+S	0.953 0.943	0.936 0.933	0.724 0.719	0.468 0.474	0.902 0.888	0.810 0.738	0.836 0.785
RACT (Zhao et al., 2026)	K+S	0.929 0.909	0.922 0.918	0.752 0.742	0.477 0.467	0.878 0.847	0.814 0.757	0.890 0.842
Q-Insight (Li et al., 2025b)	K+S	0.928 0.913	0.928 0.924	0.717 0.711	0.496 0.482	0.886 0.861	0.816 0.747	0.816 0.754
Refine-IQA-S2(think) (Jia et al., 2026)	K+S	0.916 0.920	0.918 0.930	0.715 0.703	– –	0.892 0.870	0.835 0.789	0.739 0.711
Refine-IQA-S2(no-think) (Jia et al., 2026)	K+S	0.924 0.938	0.921 0.927	0.702 0.698	– –	0.885 0.860	0.841 0.798	0.758 0.724
OASIS (Ours)	K+S	0.943 0.933	0.935 0.931	0.700 0.702	0.602 0.618	0.871 0.872	0.822 0.732	0.825 0.801
OASIS (Ours)	K+S+M	0.916 0.904	0.934 0.929	0.763 0.766	0.625 0.640	0.898 0.884	0.851 0.766	0.894 0.855

6.6 Ablation Study

Table 7 systematically evaluates OASIS components across descriptive pretraining (D), semantic score tokens (T), definition augmentation (A), and ordinal Wasserstein loss (L). Progressing from Setting 1 (score-only baseline) to full OASIS, overall PLCC/SRCC across 21 held-out dimensions increases from 0.812/0.802 to 0.829/0.823. Descriptive supervision establishes evidence grounding (+0.008/+0.012), semantic score tokens contribute +0.006/+0.005, and definition augmentation yields an additional +0.002/+0.003 by mitigating lexical memorization.

Table 7: Component ablation on MD61-Open (PLCC/SRCC). D: Stage-1 descriptive supervision; T: Semantic score tokens; A: Definition augmentation; L: Ordinal loss $\mathcal{L}_{\mathrm{W}}$.

ID	D	T	A	L	Autonomous driving	Depth estimation	Medical imaging	Remote sensing	Overall (21 dims.)
1	–	–	–	–	0.797/0.801	0.841/0.800	0.788/0.770	0.825/0.839	0.812/0.802
2	✓	–	–	–	0.822/0.825	0.848/0.812	0.797/0.786	0.815/0.832	0.820/0.814
3	✓	✓	–	–	0.826/0.830	0.847/0.808	0.800/0.787	0.830/0.848	0.826/0.819
4	✓	✓	✓	–	0.828/0.831	0.853/0.820	0.806/0.793	0.824/0.844	0.828/0.822
5	✓	✓	✓	✓	0.832/0.835	0.849/0.813	0.810/0.799	0.826/0.844	0.829/0.823

7 Discussion and Limitations

Concept Coverage. Open-dimension image scoring in this work operates over a compositional vocabulary of learned visual concepts. The 40 base dimensions cover foundational visual elements such as boundaries, occlusions, depth, textures, and structures, while the 21 held-out dimensions evaluate the composition and generalization over this established vocabulary. When presented with evaluation definitions involving visual concepts entirely absent from training, the model may struggle to reliably parse and ground the requisite evidence. Harnessing zero-shot generalization of larger foundation models or lifelong learning represents a promising future avenue.

Annotation Quality Control. Labels synthesized via the SRVP pipeline remain susceptible to shared inductive biases among teacher models. While Rank-Gate validates relative orderings rather than absolute calibration, incorporating independent expert audits alongside inter-rater reliability diagnostics would further refine teacher selection.

Evaluation Domain Diversity. Although MD61-Open spans 21 held-out dimensions across diverse domains, real-world visual pipelines present distortion profiles and evaluative requirements that exceed our benchmark coverage. Validating robustness under extreme in-the-wild conditions remains essential for production workflows.

8 Conclusion

We study open-dimension image scoring, in which a model scores images on user-specified definitions. To support this setting, we construct MD61 and introduce SRVP for base-dimension annotation and reference ratings from a five-expert annotation team for held-out evaluation. OASIS connects definition-conditioned descriptive supervision, definition augmentation, and ordinal distribution learning to produce continuous scores. Experiments show that definition-driven scoring significantly outperforms unified baselines in both expert correlation and downstream task alignment, and that multidimensional supervision improves transfer to conventional IQA benchmarks. However, direct validation on real-world decision-making workflows and robustness under extreme in-the-wild conditions remain open challenges. Future work includes extending concept coverage via larger foundation models, strengthening annotation reliability through expert audits, and validating OASIS in practical image selection and visual pipeline deployment.

Reproducibility statement

To ensure the reproducibility of all experimental results, we detail the complete dataset statistics, evaluation prompts, and dimension catalogues in the Appendices. Training configurations and hardware specifications are reported in Appendix D. Downstream task model configurations and evaluation metrics are provided in Appendix F.

Ethics statement

The research presented in this paper complies with standard ethical guidelines. The image data utilized in this study are drawn from established public academic benchmarks (BDD100K, NYU Depth V2, MedMNIST+, iSAID, SpaceNet3, KonIQ-10k, and SPAQ) in strict accordance with their original licenses and research usage terms. Expert annotations were conducted by professional researchers under informed consent and standard institutional protocols.

AI use statement

In this work, we explicitly used generative AI tools for two specific tasks:

1. Generating synthetic datasets: Generative AI models (specifically GPT-5.6, kimi-k2.7, and Claude-Haiku-4.5) were used as teacher models to generate candidate quality scores and descriptive rationales within the SRVP annotation pipeline for curating the MD61-Base training dataset.
2. Aiding and polishing writing: Large language models were used to assist in language polishing, grammar correction, and improving the overall readability of the English manuscript.

References

Kate Baumli et al. Vision-language models as a source of rewards. arXiv preprint arXiv:2312.09187, 2023.

Shuo Cao et al. ArtiMuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. arXiv preprint arXiv:2507.14533, 2025.

Liang-Chieh Chen et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, pp. 801–818, 2018.

Chaofeng Chen et al. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. IEEE Transactions on Image Processing, 33:1234–1248, 2024.

Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In CVPR, 2020.

Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In QoMEX, pp. 1–6, 2016.

Yuming Fang et al. Perceptual quality assessment of smartphone photography. In CVPR, pp. 3677–3686, 2020.

Samir Yitzhak Gadre et al. DataComp: In search of the next generation of multimodal datasets. In NeurIPS Datasets and Benchmarks Track, 2023.

S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In WACV, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In ICLR, 2019.

Congrui Hetang et al. Segment anything model for road network graph extraction. In CVPR Workshops, pp. 2556–2566, 2024.

Vlad Hosu et al. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing, 29:4041–4056, 2020.

Ziheng Jia, Jiaying Qian, Zicheng Zhang, Zijian Chen, and Xiongkuo Min. Refine-IQA: Multi-stage reinforcement finetuning for perceptual image quality assessment. In AAAI, 40(27):22301–22309, 2026.

Junjie Ke et al. MUSIQ: Multi-scale image quality transformer. In ICCV, pp. 5148–5157, 2021.

Chunyi Li et al. Image quality assessment: From human to machine preference. In CVPR, pp. 7570–7581, 2025a.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-Insight: Understanding image quality via visual reinforcement learning. arXiv preprint arXiv:2503.22679, 2025b.

Ziyu Liu et al. Visual-RFT: Visual reinforcement fine-tuning. In ICCV, 2025.

Pavan C. Madhusudana et al. Image quality assessment using contrastive learning. IEEE Transactions on Image Processing, 31:4149–4161, 2022.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21 (12):4695–4708, 2012.

Ekta Prashnani et al. PieAPP: Perceptual image-error assessment through pairwise preference. In CVPR, pp. 1808–1817, 2018.

Alec Radford et al. Learning transferable visual models from natural language supervision. In ICML, pp. 8748–8763, 2021.

Nathan Silberman et al. Indoor segmentation and support inference from RGBD images. In ECCV, pp. 746–760, 2012.

Shaolin Su et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In CVPR, 2020.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. IEEE Transactions on Image Processing, 27(8):3998–4011, 2018.

Zhou Wang et al. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In AAAI, 37:2555–2563, 2023.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236, 2025.

Haoning Wu et al. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In CVPR, pp. 25490–25500, 2024a.

Haoning Wu et al. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In ICML, pp. 54015–54029, 2024b.

Tianhe Wu et al. A comprehensive study of multimodal large language models for image quality assessment. In ECCV, 2024c.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. In NeurIPS, 38, 2025.

Jiazheng Xu et al. ImageReward: Learning and evaluating human preferences for text-to-image generation. arXiv preprint arXiv:2304.05977, 2023.

Kangmin Xu et al. Boosting image quality assessment through efficient transformer adaptation with local feature enhancement. In CVPR, 2024.

Sidi Yang et al. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In CVPR Workshops, 2022.

Jiancheng Yang et al. MedMNIST v2 – a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, 10(1):41, 2023.

Lihe Yang et al. Depth Anything V2. In NeurIPS, 37, 2024.

Zhiyuan You et al. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In ECCV, 2024.

Zhiyuan You et al. Teaching large language models to regress accurate image quality scores using score distribution. In CVPR, pp. 14483–14494, 2025.

Fisher Yu et al. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In CVPR, pp. 2636–2645, 2020.

Syed Waqas Zamir et al. iSAID: A large-scale dataset for instance segmentation in aerial images. In CVPR Workshops, pp. 28–37, 2019.

Richard Zhang et al. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, pp. 586–595, 2018.

Weixia Zhang et al. Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology, 30 (1):36–47, 2020.

Sheng Zhang et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915, 2023a.

Weixia Zhang et al. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In CVPR, 2023b.

Jipeng Zhang, Kehao Miao, Renjie Pi, Zhaowei Wang, Runtao Liu, Rui Pan, and Tong Zhang. VL-GenRM: Enhancing vision-language verification via vision experts and iterative training. arXiv preprint arXiv:2506.13888, 2025.

Shijie Zhao, Xuanyu Zhang, Weiqi Li, Junlin Li, Li Zhang, Tianfan Xue, and Jian Zhang. Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. In ICLR, 2026.

A Notations and Statistical Conventions

The mathematical notations defined in Table A.1 are consistently adopted across the main paper and appendices. Score rating tiers, teacher annotators, and downstream task configurations are indexed by k, e, and f, respectively.

Table A.1: Comprehensive summary of primary mathematical notations.

Symbol	Definition and Semantic Operationalization
x_i, d_j	The i-th image entry and j-th natural-language evaluation definition.
$q_\theta(x, d)$	OASIS continuous predicted score for image x conditioned on definition d.
$\mathcal{D}_{\mathrm{B}}, \mathcal{D}_{\mathrm{O}}$	Sets of 40 base training dimensions and 21 held-out evaluation dimensions ($\mathcal{D}_{\mathrm{B}} \cap \mathcal{D}_{\mathrm{O}} = \varnothing$).
$\mathcal{D}_s, M_s$	Evaluation dimension set for scenario s and its cardinality $M_s = \lvert\mathcal{D}_s\rvert$.
c_{ij}, n_{ij}	Dimension-relevant description rationale and count of valid answer tokens; t indexes token position.
$y_{ij}, y^{\mathrm{H}}_{ij}, \widehat{y}_{ij}$	Base training reference score, expert consensus reference score, and model prediction.
ϕ, θ	Model parameters during Stage-1 description training and Stage-2 scoring training.
v_k, r_k	The k-th score token and its associated scalar value $r_k = 10k$, $k = 0, \ldots, 10$.
π_{bk}	Soft target score distribution constructed via adjacent-tier linear interpolation.
$p_b(u), \bar{p}_{bk}$	Full-vocabulary output probability and normalized probability within scoring subspace.
$F_b(k), G_b(k)$	Cumulative probability distributions of predicted and target score distributions at tier k.
$\mathcal{L}_{\mathrm{desc}}, \mathcal{L}_{\mathrm{CE}}, \mathcal{L}_{\mathrm{W}}$	Description log-likelihood loss, full-vocabulary cross-entropy, and normalized Wasserstein loss.
λ_{W}	Ordinal Wasserstein loss weight, fixed to 0.5 across all scoring experiments.
a_{ij}, A_{is}	Applicability indicator and count of applicable dimensions for image i in scenario s.
$\widehat{Q}_s(x_i), Q^{\mathrm{H}}_s(x_i)$	Model predicted and expert consensus composite scores for image i in scenario s.
t_{if}, T_i	Per-image downstream task score under configuration f and unweighted task mean across 5 models.
$E_{ij}, s^{(e)}_{ij}, \widetilde{s}_{ij}$	Count of valid teachers, individual teacher rating, and candidate score mean for pair (x_i, d_j).
$\widehat{\sigma}_j, \Delta^{(j)}_{ii'}, \rho_{ij}$	Robust candidate score scale estimate, standardized difference, and blind ranking concordance rate.

B Dimension Catalogue

Table B.1 lists the 40 base dimensions and 21 held-out dimensions. Identifiers follow the released definitions. Detailed definitions are provided in Appendix G.

Table B.1: MD61 dimension catalogue. Valid ratings include all records in each part; MD61-Base includes its internal training, validation, and test splits.

ID	Dimension	Valid ratings
MD61-Base: 40 base dimensions		
B01	acquisition_artifact_cleanliness	24,121
B02	layered_structure_observability	5,806
B03	surface_clarity	3,397
B04	attribute_observability	3,782

Continued on next page

Table B.1 (continued)

ID	Dimension	Valid ratings
B05	background_noninterference	9,867
B06	boundary_separability	15,688
B07	brightness_suitability	14,405
B08	color_harmony	5,397
B09	color_richness	9,678
B10	composition_balance	900
B11	contrast_suitability	15,704
B12	dense_instance_isolation	3,572
B13	depth_of_field_control	5,364
B14	depth_structure_observability	2,523
B15	detail_fidelity	25,460
B16	layout_coherence	6,642
B17	entity_recognizability	12,223
B18	environmental_occlusion_visibility	17,480
B19	symbol_form_integrity	8,433
B20	lighting_harmony	5,393
B21	lighting_Reasonableness	23,266
B22	material_cue_observability	5,394
B23	microstructure_resolvability	7,603
B24	multiscale_resolvability	5,100
B25	noise_cleanliness	5,906
B26	oriented_instance_shape_integrity	3,081
B27	sequential_layout_traceability	2,739
B28	region_separability	28,460
B29	surface_region_observability	2,494
B30	scene_relation_clarity	5,720
B31	semantic_content_appeal	5,392
B32	spatial_topology_continuity	1,715
B33	structural_coherence	17,275
B34	structured_pattern_parseability	6,103
B35	subject_emphasis_clarity	5,375
B36	foreground_background_separability	8,915
B37	symbol_geometric_regularity	8,591
B38	symbol_legibility	8,591
B39	thin_linear_structure_traceability	8,768
B40	visual_symbol_recognizability	1,898
MD61-Base total		358,221
MD61-Open: Autonomous driving		
O01	driving_perception_evidence_mos	2,742
O02	road_user_instance_resolvability	2,656
O03	drivable_surface_boundary_clarity	2,740
O04	lane_marking_traceability	1,606
O05	road_scene_exposure_evidence_preservation	2,742
O06	localized_road_glare_nonobstruction	2,742
MD61-Open: Remote sensing		
O07	aerial_segmentation_evidence_mos	1,614
O08	aerial_small_structure_resolvability	1,614
O09	aerial_region_boundary_clarity	1,614
O10	road_network_traceability	1,374
O11	atmospheric_ground_evidence_preservation	1,614
MD61-Open: Medical imaging		
O12	medical_structure_evidence_mos	1,379
O13	anatomical_boundary_visibility	1,377
O14	tissue_texture_discriminability	1,327
O15	medical_microstructure_preservation	1,376
O16	medical_intensity_structure_separation	1,380
MD61-Open: Depth estimation		
O17	monocular_depth_evidence_mos	755

Continued on next page

Table B.1 (continued)

ID	Dimension	Valid ratings
O18	occlusion_boundary_depth_clarity	754
O19	perspective_scale_cue_consistency	754
O20	surface_gradient_depth_observability	755
O21	localized_highlight_depth_evidence_preservation	755
MD61-Open total		33,670

C Image Data Sources and Benchmark Construction

C.1 MD61-Base Training Suite

The 19 image sources comprising MD61-Base are drawn from established public benchmarks across natural scenes, fine art, text and document layouts, road scenes, depth maps, medical imaging, and remote sensing: SPAQ, KonIQ-10k, TID2013, COCO2017, Visual Genome, OpenImages V7, VisDrone, TextOCR, NYU Depth V2, SpaceNet3 Roads, BDD100K, MedMNIST+, iSAID, DocLayNet, DocBank, OmniDocBench, AVA, ArtiMuse-10K, and PARA. Internal training, validation, and test splits contain 300,992, 18,155, and 39,074 annotations, respectively, all retaining structured descriptive rationales paired with score labels.

C.2 MD61-Open Held-Out Benchmark

MD61-Open comprises 1,082 clean source images selected from public validation and test sets, strictly isolated from the training distribution:

- Autonomous Driving: 457 source scenes selected from the BDD100K validation split (Yu et al., 2020).
- Depth Estimation: 126 indoor RGB-D scenes from the NYU Depth V2 test partition (Silberman et al., 2012).
- Medical Imaging: 230 biomedical images across 5 modalities from MedMNIST+ (Yang et al., 2023) (ChestMNIST, DermaMNIST, OCTMNIST, PathMNIST, PneumoniaMNIST).
- Remote Sensing: 269 aerial and satellite images, including 215 high-resolution aerial scenes from iSAID (Zamir et al., 2019) and 54 road network extraction tiles from SpaceNet3 (Hetang et al., 2024).

For each pristine source image, five progressive levels of directional physical degradation are synthesized (underexposure, high-frequency blur/attenuation, localized glare, radiographic contrast loss, and atmospheric obscuration), yielding a total evaluation benchmark of $1{,}082 \times 6 = 6{,}492$ image variants.

D Two-Stage Training and Ordinal Scoring

The backbone is Qwen3-VL-8B-Instruct. RGB inputs preserve their original aspect ratios. The long edge is resized to 720 for remote-sensing and document images and to 448 for all other domains. Stage 1 updates the visual modules and applies LoRA to the language modules. After training, the visual updates are retained and the language LoRA is merged; the resulting model initializes Stage 2. We denote the parameters during description learning by ϕ and the scoring parameters optimized after initialization by θ.

Table D.1: Two-stage training configuration on MD61-Base.

Configuration	Description learning (Stage 1)	Score learning (Stage 2)
Learning rate	4×10^{-5}	2×10^{-5}
Weight decay	0.05	0
LoRA rank / alpha	16 / 32	128 / 256
LoRA dropout	0.05	0.05
GPUs / batch size per GPU	4 / 64	4 / 64
Gradient accumulation / global batch size	1 / 256	1 / 256
Schedule / warmup ratio	cosine / 0.03	cosine / 0.03

Training uses four NVIDIA RTX PRO 6000 GPUs with BF16, SDPA, and gradient checkpointing. The software environment consists of PyTorch 2.10.0+cu128, Transformers 4.57.0, PEFT 0.17.1, and Accelerate 1.7.0.

D.1 Description Supervision

Stage-1 description supervision is constructed according to the training-data configuration. In IQA transfer experiments that exclude MD61, description pretraining uses the training images of the corresponding IQA datasets together with image-quality descriptions generated by GPT-5.4. The K setting uses KonIQ, whereas the K+S setting uses KonIQ and SPAQ; neither setting introduces MD61 data.

The complete description dataset and the current mini-batch are defined as

$$\mathcal{T}_{\text{desc}} = \{(x_i, d_j, c_{ij}) \mid (i,j) \in \Omega_{\text{desc}}\}, \qquad \mathcal{B}_{\text{desc}} \subseteq \Omega_{\text{desc}}. \tag{D.1}$$

Here, $c_{ij} = (c_{ij}^1, \ldots, c_{ij}^{n_{ij}})$ is the valid response sequence, including the end-of-sequence token, and $c_{ij}^{<t}$ denotes the response prefix before position t. The loss is

$$\mathcal{L}_{\text{desc}}(\phi) = -\frac{1}{N_{\text{desc}}} \sum_{(i,j)\in\mathcal{B}_{\text{desc}}} \sum_{t=1}^{n_{ij}} \log p_\phi\left(c_{ij}^t \mid x_i, d_j, c_{ij}^{<t}\right), \qquad N_{\text{desc}} = \sum_{(i,j)\in\mathcal{B}_{\text{desc}}} n_{ij}. \tag{D.2}$$

Both summation and normalization use valid response tokens in the current mini-batch, while prompt positions are masked. Different definitions applied to the same image form separate training examples. The target description explains the strengths, defects, and effects of the relevant regions.

D.2 Score Data and Definition Augmentation

The score dataset and definition augmentation are defined as

$$\mathcal{T}_{\text{score}} = \{(x_i, d_j, y_{ij}) \mid (i,j) \in \Omega_{\text{score}}\}, \qquad \widetilde{d}_b = \mathcal{A}(d_b; \xi_b). \tag{D.3}$$

The score mini-batch has size B, where $b = 1, \ldots, B$ indexes one (i,j) pair in the batch; thus, (x_b, d_b, y_b) denotes the corresponding record. The variable ξ_b specifies the random augmentation applied at that access, while an unaugmented record uses the identity transformation. Synonym rewriting and dimension-name replacement preserve the evaluation object, scope of applicability, evidence direction, and score guidance; the image and label remain unchanged.

D.3 Score Vocabulary

After freezing the model trained in Stage 1, quality-related tokens are selected from its full vocabulary $\mathcal{V}$. Each candidate must be independently encoded as a single token by the

backbone tokenizer. Given high- and low-quality textual contexts, the quality tendency of a candidate token is defined using its next-token logits:

$$\eta(v) = \ell_{\text{high}}(v) - \ell_{\text{low}}(v). \tag{D.4}$$

The candidates are sorted by $\eta(v)$, and 11 tokens are selected at equally spaced ranking positions:

$$\mathcal{V}_q = \{v_0, \ldots, v_{10}\} \subset \mathcal{V}, \qquad r_k = 10k, \quad k = 0, \ldots, 10. \tag{D.5}$$

The token semantics provide an initial ordering, while the scoring stage learns the output probability at each level under reference-score supervision. The selection procedure uses textual contexts only.

D.4 Continuous Labels and Soft Targets

For a reference score $y_b \in [0, 100]$, define

$$k_b = \min\left(\left\lfloor \frac{y_b}{10} \right\rfloor, 9\right), \qquad \alpha_b = \frac{y_b}{10} - k_b. \tag{D.6}$$

Here, k_b is the lower index of the two adjacent interpolation levels. The target probabilities are

$$\pi_{bk} = \begin{cases} 1 - \alpha_b, & k = k_b, \\ \alpha_b, & k = k_b + 1, \\ 0, & \text{otherwise.} \end{cases} \tag{D.7}$$

They satisfy

$$\sum_{k=0}^{10} \pi_{bk} = 1, \qquad \sum_{k=0}^{10} r_k \pi_{bk} = y_b. \tag{D.8}$$

When $y_b = 0$, all probability mass is assigned to v_0; when $y_b = 100$, $k_b = 9$, $\alpha_b = 1$, and all mass is assigned to v_{10}. Thus, the distribution encodes one continuous target through linear interpolation.

D.5 Full-Vocabulary Cross-Entropy and Ordinal Distance

At the designated scoring position, let $z_{\theta,b}(u)$ denote the logit of token $u \in \mathcal{V}$. The full-vocabulary probability and cross-entropy loss are

$$p_b(u) \equiv p_\theta(u \mid x_b, \widetilde{d}_b) = \frac{\exp z_{\theta,b}(u)}{\sum_{u' \in \mathcal{V}} \exp z_{\theta,b}(u')}, \tag{D.9}$$

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{b=1}^{B} \sum_{k=0}^{10} \pi_{bk} \log p_b(v_k). \tag{D.10}$$

Full-vocabulary normalization makes the score tokens compete with all other tokens. Restricting the distribution to the scoring subspace gives

$$\overline{p}_{bk} = \frac{p_b(v_k)}{\sum_{h=0}^{10} p_b(v_h)} = \overline{p}_\theta(v_k \mid x_b, \widetilde{d}_b), \qquad F_b(k) = \sum_{h=0}^{k} \overline{p}_{bh}, \qquad G_b(k) = \sum_{h=0}^{k} \pi_{bh}. \tag{D.11}$$

The normalized first-order Wasserstein distance is

$$\begin{aligned} \mathcal{L}_{\text{W}} &= \frac{1}{B} \sum_{b=1}^{B} \frac{1}{100} \sum_{k=0}^{9} (r_{k+1} - r_k) \left| F_b(k) - G_b(k) \right| \\ &= \frac{1}{10B} \sum_{b=1}^{B} \sum_{k=0}^{9} \left| F_b(k) - G_b(k) \right|. \end{aligned} \tag{D.12}$$

The two forms are equivalent, and $\mathcal{L}_{\mathrm{W}} \in [0,1]$. The final objective follows Eq. (5) in the main text, with $\lambda_{\mathrm{W}} = 0.5$. The scoring stage supervises only the designated scoring position and does not include an EOS loss. The description stage retains the autoregressive objective with the end-of-sequence token.During inference, continuous scores are decoded as:

$$q_\theta(x,d) = \sum_{k=0}^{10} r_k \bar{p}_\theta(v_k \mid x, d). \tag{D.13}$$

E SRVP Annotation and Blind Rank Verification

SRVP is used to acquire additional training annotations for the base dimensions.

E.1 Candidate Scores

Let $s_{ij}^{(e)}$ denote the score assigned to the image–definition pair (x_i, d_j) by the e-th valid teacher. Given $E_{ij} \geq 1$ valid teachers, the candidate mean is

$$\widetilde{s}_{ij} = \frac{1}{E_{ij}} \sum_{e=1}^{E_{ij}} s_{ij}^{(e)}. \tag{E.1}$$

For newly added records accepted by the gate, we use $y_{ij} = \widetilde{s}_{ij}$ as score supervision. Reused human labels retain their matched scores. The evidence description c_{ij} is stored separately for each dimension, while the sources of teacher scores and descriptions are recorded independently.

E.2 Blind Ranking Groups

Let $\mathcal{I}_j$ be the set of image indices with candidate scores for dimension j. We estimate the scale using the median absolute deviation:

$$\widehat{\sigma}_j = 1.4826\,\mathrm{median}_{i\in\mathcal{I}_j} \left|\widetilde{s}_{ij} - \mathrm{median}_{h\in\mathcal{I}_j}\, \widetilde{s}_{hj}\right|, \qquad \Delta_{ii'}^{(j)} = \frac{|\widetilde{s}_{ij} - \widetilde{s}_{i'j}|}{\max(\widehat{\sigma}_j, \epsilon)}. \tag{E.2}$$

Here, i' is a reference image for the same dimension, and $\epsilon > 0$ ensures a valid denominator. Near/Far triplets are constructed from the standardized score differences to cover small and large quality gaps. Candidate scores are used only for group selection; ranking reviewers see only the images and the definition and provide a relative ordering and confidence level.

E.3 Acceptance Rule

Let $N_{\mathrm{agree}}^{(ij)}$ and $N_{\mathrm{conflict}}^{(ij)}$ count valid comparisons whose blind ranking agrees or conflicts with the candidate ordering, respectively. When their sum is positive, define

$$\rho_{ij} = \frac{N_{\mathrm{agree}}^{(ij)}}{N_{\mathrm{agree}}^{(ij)} + N_{\mathrm{conflict}}^{(ij)}}. \tag{E.3}$$

A record is accepted if $\rho_{ij} \geq \rho_{\min}$ and no high-confidence severe contradiction is observed. The accepted score is the candidate mean in Eq. (E.1). Filtered records are added to the reference pool, from which Near/Far references are selected for subsequent incremental candidates.

F Task Scores and Correlation Computation

F.1 Scenario Scores and Statistical Units

For a scenario dimension set $\mathcal{D}_s = \{d_{j_1}, \ldots, d_{j_{M_s}}\}$, where $M_s = |\mathcal{D}_s|$, the shared model outputs

$$\mathbf{q}_{\mathcal{D}_s}(x) = \left[q_\theta(x, d_{j_1}), \ldots, q_\theta(x, d_{j_{M_s}})\right]^\top. \tag{F.1}$$

For scenario s, define

$$\mathcal{J}_{is} = \{j : d_j \in \mathcal{D}_s,\ a_{ij} = 1\}, \qquad A_{is} = |\mathcal{J}_{is}| > 0. \tag{F.2}$$

The model and expert composite scores are computed using the same applicability set on both sides.

Each image is evaluated with five fixed task configurations indexed by $f = 1, \ldots, 5$. Let t_{if} denote the per-image task score of configuration f. The equally weighted task score is

$$T_i = \frac{1}{5} \sum_{f=1}^{5} t_{if}. \tag{F.3}$$

All task scores are oriented so that higher values indicate better performance and use the corresponding task-specific ground-truth metric. The configurations are listed in Table F.1.

Table F.1: Fixed model configurations for downstream tasks. Each group uses five configurations with equal task-score weight of 0.2.

Task group	Model configurations
BDD100K	YOLOP; YOLOPv2; HybridNets; A-YOLOM-S; A-YOLOM-N
NYU Depth V2	Depth Anything V2 Metric Indoor Small; ZoeDepth-N; Metric3D v2 ViT-Small; UniDepth V2 ViT-L/14; Depth Pro
MedMNIST+	BiomedCLIP; MedCLIP-ViT; PMC-CLIP; UniMed-CLIP ViT-B/16; PubMedCLIP ViT-B/32
iSAID	DeepLabV3+ R50-D8; PSPNet R50-D8; HRNet-W48 FCN; DeepLabV3+ R18-D8; PSPNet R18-D8
SpaceNet3	SAM-Road; PFR-DPN92 model01; RNGDet; RNGDet++; Sat2Graph model1

F.2 Correlation Computation

Downstream-task validation directly computes SRCC and PLCC between the raw composite score of the model or experts and the corresponding task-score average. For remote sensing, correlations are computed separately for iSAID and SpaceNet3 and then averaged with equal weights. Medical records from the five subsets are pooled. For dimension-wise model–expert evaluation, correlations are first computed over all applicable records for each dimension and then averaged equally across the 21 dimensions. Remote-sensing dimensions are not further split by source in this dimension-wise evaluation.

F.3 Autonomous Driving

For BDD100K (Yu et al., 2020), we measure the agreement between task-model predictions under different image conditions and the corresponding ground-truth annotations. For object detection, all predicted boxes belonging to the same class are merged, and the union is compared with the union of all ground-truth boxes of that class. Let $\mathcal{B}^{\text{pred}}_{if,c}$ and $\mathcal{B}^{\text{gt}}_{if,c}$ denote the predicted and ground-truth box unions for class c. The detection score and overall driving score are

$$S^{\text{det}}_{if} = \frac{1}{|\mathcal{C}_{if}|} \sum_{c \in \mathcal{C}_{if}} \text{IoU}\left(\mathcal{B}^{\text{pred}}_{if,c}, \mathcal{B}^{\text{gt}}_{if,c}\right), \tag{F.4}$$

$$t^{\text{drive}}_{if} = \frac{1}{3} S^{\text{det}}_{if} + \frac{1}{3} \text{IoU}(D_{if}, D^{\text{gt}}_{if}) + \frac{1}{3} \text{IoU}(L_{if}, L^{\text{gt}}_{if}). \tag{F.5}$$

Here, $\mathcal{C}_{if}$ is the set of object classes included in the detection evaluation. D_{if} and L_{if} denote the predicted drivable-area and lane masks, while D^{gt}_{if} and L^{gt}_{if} are the corresponding ground-truth masks. The metric weights object detection, drivable-area segmentation, and lane segmentation equally and uses IoU to quantify task performance under different image conditions.

F.4 Depth Estimation

For NYU Depth V2 (Silberman et al., 2012), the per-image threshold accuracy is

$$t_{if}^{\text{depth}} = \delta_{1,if} = \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \mathbf{1}\left[\max\left(\frac{\widehat{Z}_{if}(p)}{Z_i(p)}, \frac{Z_i(p)}{\widehat{Z}_{if}(p)}\right) < 1.25\right]. \tag{F.6}$$

Here, $Z_i(p)$ and $\widehat{Z}_{if}(p)$ are the ground-truth and predicted depths at pixel p, and $\mathcal{P}_i$ is the set of valid pixels within the Eigen crop satisfying $10^{-3} < Z_i(p) < 10$ meters. The five configurations, including Depth Anything V2 (Yang et al., 2024), are evaluated separately and averaged according to Eq. (F.3).

F.5 Medical Ground-Truth State Probabilities

MedMNIST (Yang et al., 2023) uses five vision–language models, including BiomedCLIP (Zhang et al., 2023a), for zero-shot class matching. Let y_i^{cls} denote the single-label ground-truth class, $y_{ic}^{\text{bin}} \in \{0, 1\}$ the ground-truth state for class c in ChestMNIST, and p_{ifc}^{task} the predicted probability of the positive state. We define

$$t_{if}^{\text{med}} = \begin{cases} p_{if}^{\text{task}}(y_i^{\text{cls}} \mid x_i), & \text{single-label subsets,} \\ \dfrac{1}{14} \displaystyle\sum_{c=1}^{14} \left[y_{ic}^{\text{bin}} p_{ifc}^{\text{task}} + (1 - y_{ic}^{\text{bin}})(1 - p_{ifc}^{\text{task}})\right], & \text{ChestMNIST.} \end{cases} \tag{F.7}$$

The class or state ground truth here is distinct from the image-quality label y_{ij}. Each configuration uses its corresponding similarity scaling and class prompts; the probabilities serve only as task references for the quality evaluation.

F.6 Remote-Sensing Pixel Metrics

iSAID (Zamir et al., 2019) uses five segmentation configurations, including DeepLabV3+ (Chen et al., 2018). Let $C_i(p)$ and $\widehat{C}_{if}(p)$ denote the ground-truth and predicted pixel classes, with background class 0. For the foreground pixel set

$$\mathcal{F}_i = \{p : C_i(p) > 0\},$$

the score is

$$t_{if}^{\text{aerial}} = \frac{\sum_{p \in \mathcal{F}_i} \mathbf{1}\left[\widehat{C}_{if}(p) = C_i(p)\right]}{|\mathcal{F}_i|}. \tag{F.8}$$

For SpaceNet3, let $\mathcal{R}_i$ and $\widehat{\mathcal{R}}_{if}$ denote the ground-truth and predicted road-pixel sets. We use intersection over union:

$$t_{if}^{\text{road}} = \frac{|\widehat{\mathcal{R}}_{if} \cap \mathcal{R}_i|}{|\widehat{\mathcal{R}}_{if} \cup \mathcal{R}_i|} = \frac{\text{TP}_{if}}{\text{TP}_{if} + \text{FP}_{if} + \text{FN}_{if}}. \tag{F.9}$$

Here, TP_{if}, FP_{if}, and FN_{if} denote the numbers of true-positive, false-positive, and false-negative road-pixel predictions, respectively, and the ground-truth road-pixel set is nonempty. The outputs of SAM-Road (Hetang et al., 2024) and other models are aligned to the ground-truth grid. RNGDet and RNGDet++ use their segmentation branches, while Sat2Graph outputs are rasterized. The iSAID metric evaluates foreground-pixel classification accuracy, whereas the SpaceNet3 metric uses road-pixel IoU to account for both missed and false detections.

The main-text Table 5 reports raw task correlations using multidimensional composite scores over the same applicable dimensions. Scorer training and checkpoint selection use only base-dimension supervision; task scores are used solely as external evaluation references.

G Dimension Definitions

Table G.1 reproduces the original English definitions. Dimension identifiers appear above each entry; Definition, Positive, Negative, and Scoring correspond to description, positive prompt, negative prompt, and score guidance, respectively.

Table G.1: Original English definitions of the 61 MD61 dimensions.

ID / Field	Original definition
MD61-Base: 40 base dimensions	
B01 acquisition_artifact_cleanliness	
B01 / Definition	Evaluate freedom from directly visible structured artifacts introduced during image formation, acquisition, processing, or reconstruction. A higher score means the image is free of motion streaks, striping or banding, ringing or halos, defective pixels, acquisition marks, clipped regions, ghosting, duplicated structures, and reconstruction defects that affect visible evidence, while a lower score means one or more of these artifacts materially obscure, distort, duplicate, or remove visible evidence. Judge primarily from the visible pattern, severity, spatial extent, and affected image area of the named artifacts. Weight the evidence by artifact severity and affected image area. Do not penalize ordinary defocus or softness, unstructured noise or grain, darkness, low contrast, low resolution, or missing fine detail unless the visible pattern itself is one of the named artifacts.
B01 / Positive	the image is free of motion streaks, striping or banding, ringing or halos, defective pixels, acquisition marks, clipped regions, ghosting, duplicated structures, and reconstruction defects that affect visible evidence. Visible evidence: the presence, severity, spatial extent, and affected image area of these named patterns.
B01 / Negative	one or more named structured artifacts materially obscure, distort, duplicate, or remove visible evidence. Apply the penalty in proportion to artifact severity and affected image area.
B01 / Scoring	90-100: no meaningful named artifact; 70-89: minor local named artifact; 40-69: obvious evidence-affecting named artifact; 0-39: severe or widespread named artifact.
B02 layered_structure_observability	
B02 / Definition	Evaluate observability of relevant layered or nested visual structures. A higher score means boundaries, layers, spatial ordering, and internal organization are visibly interpretable, while a lower score means they are concealed, merged, distorted, or too poorly resolved to interpret. Judge primarily from layer separation, structural arrangement, local contrast, continuity, and artifact overlap. Weight the evidence by severity, affected relevant area, and task relevance. Do not assume a semantic identity for the layers; assess only structure supported by visible pixels.
B02 / Positive	relevant boundaries, layers, spatial ordering, and internal structural cues are visibly interpretable. Visible evidence: layer separation, structural arrangement, local contrast, continuity, and artifact overlap.
B02 / Negative	relevant layered structures are concealed, merged, distorted, or too poorly resolved to interpret. Apply the penalty in proportion to severity and affected relevant area.
B02 / Scoring	90-100: relevant layered structure is clearly observable; 70-89: minor uncertainty; 40-69: notable local ambiguity; 0-39: structure is broadly uninterpretable.
B03 surface_clarity	
B03 / Definition	Evaluate the visible clarity of relevant surfaces and their local patterns. A higher score means surface boundaries, textures, markings, and fine variations remain clear enough to inspect, while a lower score means haze, blur, shadow masking, noise, reflections, or other visible degradation makes those surface cues indistinct or difficult to separate. Judge primarily from the sharpness and continuity of surface patterns, local contrast, obscuring coverage, and the extent to which degradation affects relevant surface area. Weight the evidence by severity, affected relevant area, and task relevance. Do not penalize a naturally smooth, low-texture, or uniformly colored surface when its visible boundaries and relevant cues remain clear.
B03 / Positive	relevant surface boundaries, textures, markings, and local patterns are clear and inspectable. Visible evidence: pattern sharpness, contour continuity, local contrast, unobscured surface area, and separable fine variation.
B03 / Negative	haze, blur, shadow masking, noise, reflections, or other degradation makes relevant surface patterns indistinct, merged, or difficult to inspect. Apply the penalty in proportion to severity and affected relevant area.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B03 / Scoring	90-100: relevant surfaces and local patterns are consistently clear; 70-89: minor local softness or obstruction; 40-69: substantial relevant areas have indistinct surface evidence; 0-39: most relevant surface cues are unclear or unusable.
B04 attribute_observability	
B04 / Definition	Evaluate whether visible properties of clearly depicted entities can be determined directly from the image, separately from identifying the entity category itself. A higher score means color, material, state, pose, markings, and part properties of visually salient or inspectable entities are unambiguous. A lower score means those visibly presented properties are hidden, clipped, too small, color-distorted, saturated, blurred, or mutually confusable. Judge only properties supported by visible pixels, and do not assume any property that is not visibly presented. A recognizable silhouette or category alone is not positive evidence for this dimension.
B04 / Positive	visibly presented colors, materials, states, poses, markings, and parts of clear entities are directly observable
B04 / Negative	visibly presented entity properties are hidden, clipped, too small, color-distorted, saturated, blurred, or confusable
B04 / Scoring	90-100 nearly all visibly presented properties of clear entities are direct and unambiguous; 70-89 most are observable with minor local uncertainty; 40-69 several visible properties remain ambiguous; 0-39 most visibly presented properties cannot be determined. Do not award points merely because an entity category is recognizable.
B05 background_noninterference	
B05 / Definition	Evaluate whether surrounding content interferes with salient entities. A higher score means background does not conceal, mimic, or distract from important entities, while a lower score means clutter, overlap, or look-alike texture obscures or confuses important entities. Judge primarily from occlusion, clutter density, color and texture similarity, and distracting overlap. Weight the evidence by severity, affected relevant area, and task relevance. Do not treat ordinary subject matter, artistic style, or an unusual but still observable property as a failure unless it directly removes the stated visual evidence.
B05 / Positive	background does not conceal, mimic, or distract from important entities. Visible evidence: occlusion, clutter density, color and texture similarity, and distracting overlap.
B05 / Negative	clutter, overlap, or look-alike texture obscures or confuses important entities. Apply the penalty in proportion to severity and affected relevant area.
B05 / Scoring	90-100 minimal interference; 70-89 modest clutter; 40-69 noticeable confusion; 0-39 background dominates evidence.
B06 boundary_separability	
B06 / Definition	Evaluate separation of entities or regions from adjacent content. A higher score means important boundaries are visible, continuous, and distinguishable, while a lower score means important boundaries merge with adjacent regions or are weak, occluded, or fragmented. Judge primarily from edge contrast, contour continuity, adjacency, occlusion, and clutter. Weight the evidence by severity, affected relevant area, and task relevance. Do not treat ordinary subject matter, artistic style, or an unusual but still observable property as a failure unless it directly removes the stated visual evidence.
B06 / Positive	important boundaries are visible, continuous, and distinguishable. Visible evidence: edge contrast, contour continuity, adjacency, occlusion, and clutter.
B06 / Negative	important boundaries merge with adjacent regions or are weak, occluded, or fragmented. Apply the penalty in proportion to severity and affected relevant area.
B06 / Scoring	90-100 consistently separable; 70-89 small local merges; 40-69 frequent ambiguity; 0-39 broadly unusable.
B07 brightness_suitability	
B07 / Definition	Whether the image brightness is appropriate and well balanced for viewing, so the scene is neither underexposed nor overexposed. For this dimension, a higher score means the brightness is more appropriate, more comfortable, and better balanced for viewing.
B07 / Positive	an image with appropriate and balanced brightness, clear visibility, comfortable overall luminance, and no obvious underexposure or overexposure
B07 / Negative	an image with inappropriate brightness, obvious underexposure or overexposure, uncomfortable luminance, or lost visibility

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B07 / Scoring	Dimension-specific score meaning: For this dimension, the 0-100 score measures how good, appropriate, and visually successful the image is on the named dimension. A higher score means the image is better on this dimension according to the dimension explanation and reference prompts. A lower score means the image shows stronger visible weaknesses on this dimension.
B08 color_harmony	
B08 / Definition	Evaluate whether the visible dominant, supporting, and accent colors form a coherent palette. A higher score means hue relations, saturation differences, and color accents establish a clear visual hierarchy and consistent atmosphere. A lower score means visible colors clash, compete without hierarchy, or form an incoherent palette. Judge visible color relationships rather than color richness, white balance, exposure, or subject appeal.
B08 / Positive	coherent visible dominant supporting and accent colors with stable hue and saturation relationships
B08 / Negative	clashing or incoherent color relationships with competing accents and weak palette hierarchy
B08 / Scoring	90-100 the visible palette is exceptionally coherent; 70-89 minor local mismatch; 40-69 several competing or discordant color relations; 0-39 the visible palette is broadly incoherent.
B09 color_richness	
B09 / Definition	Whether the image colorfulness is appropriate, natural, and visually suitable, rather than too dull, too weak, distorted, or unnaturally oversaturated. For this dimension, a higher score means the colors are more natural, more appropriate, and more visually pleasing for the image content.
B09 / Positive	an image with appropriate, natural, and visually pleasing colorfulness, with neither dull weak color nor unnatural oversaturation
B09 / Negative	an image with inappropriate colorfulness, such as dull faded color, weak color, distorted color, or unnaturally oversaturated color
B09 / Scoring	Dimension-specific score meaning: For this dimension, the 0-100 score measures how good, appropriate, and visually successful the image is on the named dimension. A higher score means the image is better on this dimension according to the dimension explanation and reference prompts. A lower score means the image shows stronger visible weaknesses on this dimension.
B10 composition_balance	
B10 / Definition	Evaluate whether the visible arrangement of subjects, visual masses, empty space, and directional lines forms a stable and coherent composition. A higher score means placement, spacing, scale, and visual weight create clear balance and visual flow. A lower score means the frame appears lopsided, cramped, fragmented, or weakly organized. Judge only visible spatial arrangement rather than subject attractiveness, technical sharpness, or exposure.
B10 / Positive	balanced visual weight, coherent subject placement, effective spacing, and stable directional flow
B10 / Negative	lopsided, cramped, fragmented, or weakly organized visible arrangement
B10 / Scoring	90-100 visible spatial arrangement is exceptionally balanced and coherent; 70-89 minor imbalance or weak use of space; 40-69 several placement or visual-weight conflicts weaken the composition; 0-39 the visible arrangement lacks stable compositional structure.
B11 contrast_suitability	
B11 / Definition	Whether the image contrast is appropriate and well balanced, providing clear tonal separation without looking either flat or overly harsh. For this dimension, a higher score means the tonal separation is more appropriate, clearer, and better balanced without excessive harshness or flatness.
B11 / Positive	an image with appropriate and balanced contrast, clear tonal separation, good visual depth, and well controlled light and dark regions
B11 / Negative	an image with inappropriate contrast, such as muddy weak separation, flat tones, or overly harsh imbalanced light and dark regions

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B11 / Scoring	Dimension-specific score meaning: For this dimension, the 0-100 score measures how good, appropriate, and visually successful the image is on the named dimension. A higher score means the image is better on this dimension according to the dimension explanation and reference prompts. A lower score means the image shows stronger visible weaknesses on this dimension.
B12 dense_instance_isolation	
B12 / Definition	Evaluate separability of individual instances in crowded scenes. A higher score means nearby instances retain individually traceable extents and cues, while a lower score means crowding, overlap, or small scale merges instances into indistinguishable groups. Judge primarily from instance gaps, overlapping contours, visible parts, and countable cues. Weight the evidence by severity, affected relevant area, and task relevance. Do not treat ordinary subject matter, artistic style, or an unusual but still observable property as a failure unless it directly removes the stated visual evidence.
B12 / Positive	nearby instances retain individually traceable extents and cues. Visible evidence: instance gaps, overlapping contours, visible parts, and countable cues.
B12 / Negative	crowding, overlap, or small scale merges instances into indistinguishable groups. Apply the penalty in proportion to severity and affected relevant area.
B12 / Scoring	90-100 dense instances separable; 70-89 limited merges; 40-69 frequent ambiguity; 0-39 most collapse into groups.
B13 depth_of_field_control	
B13 / Definition	Evaluate whether the visible focus plane, resolved depth range, and foreground-to-background blur pattern form a coherent depth-of-field treatment. A higher score means visually dominant subjects or planes are clearly resolved and blur transitions are smooth and consistent. A lower score means visibly dominant content is soft while less prominent planes receive focus, or depth blur changes abruptly and distractingly. Do not reward shallow depth of field by default; a consistently all-in-focus image can score highly.
B13 / Positive	clearly resolved visually dominant planes with smooth and coherent depth-blur transitions
B13 / Negative	visually dominant planes remain soft or depth blur changes abruptly and inconsistently
B13 / Scoring	90-100 focus placement, resolved depth range, and blur transitions are exceptionally coherent; 70-89 minor local focus or transition mismatch; 40-69 several visible planes are inconsistently resolved; 0-39 the visible focus pattern lacks a coherent depth structure.
B14 depth_structure_observability	
B14 / Definition	Evaluate visibility of relative depth and 3D scene layout cues. A higher score means near-far ordering, support surfaces, occlusion, and geometry cues make layout observable, while a lower score means depth ordering and scene geometry are ambiguous, collapsed, or contradicted. Judge primarily from occlusion ordering, perspective, support contact, scale gradients, shading, and layout continuity. Weight the evidence by severity, affected relevant area, and task relevance. Do not treat ordinary subject matter, artistic style, or an unusual but still observable property as a failure unless it directly removes the stated visual evidence.
B14 / Positive	near-far ordering, support surfaces, occlusion, and geometry cues make layout observable. Visible evidence: occlusion ordering, perspective, support contact, scale gradients, shading, and layout continuity.
B14 / Negative	depth ordering and scene geometry are ambiguous, collapsed, or contradicted. Apply the penalty in proportion to severity and affected relevant area.
B14 / Scoring	90-100 depth strongly observable; 70-89 minor ambiguity; 40-69 several uncertain regions; 0-39 layout broadly ambiguous.
B15 detail_fidelity	
B15 / Definition	Whether the image sharpness is appropriate and visually suitable, with details clearly resolved rather than blurred, smeared, or unnaturally oversharpened. For this dimension, a higher score means the image details are more clearly resolved with more appropriate and natural sharpness.
B15 / Positive	an image with appropriate sharpness, crisp but natural edges, clear textures, and well resolved fine structures without obvious oversharpening
B15 / Negative	an image with inappropriate sharpness, such as blur, softness, smeared textures, weak detail definition, or obvious oversharpening artifacts

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B15 / Scoring	Dimension-specific score meaning: For this dimension, the 0-100 score measures how good, appropriate, and visually successful the image is on the named dimension. A higher score means the image is better on this dimension according to the dimension explanation and reference prompts. A lower score means the image shows stronger visible weaknesses on this dimension.
B16 layout_coherence	
B16 / Definition	Evaluate whether visible blocks, regions, and repeated elements form a coherent spatial layout. A higher score means alignment, spacing, grouping, hierarchy, and boundaries follow a stable organizing system. A lower score means visible elements collide, drift, overlap, fragment, or use inconsistent alignment and spacing. Judge visible spatial organization rather than symbol recognition or the meaning of the depicted content.
B16 / Positive	coherent block and region organization with stable alignment, spacing, grouping, and hierarchy
B16 / Negative	fragmented or colliding visible elements with inconsistent alignment, grouping, boundaries, or spacing
B16 / Scoring	90-100 spatial organization is stable and internally consistent; 70-89 minor local irregularity; 40-69 repeated layout conflicts; 0-39 the visible layout is broadly unusable.
B17 entity_recognizability	
B17 / Definition	Evaluate how readily salient entities can be identified. A higher score means entities show distinctive shapes, parts, and cues for confident identification, while a lower score means entities are obscured, tiny, blurred, truncated, or ambiguous. Judge primarily from silhouette, discriminative parts, texture cues, occlusion, blur, and scale. Weight the evidence by severity, affected relevant area, and task relevance. Do not treat ordinary subject matter, artistic style, or an unusual but still observable property as a failure unless it directly removes the stated visual evidence.
B17 / Positive	entities show distinctive shapes, parts, and cues for confident identification. Visible evidence: silhouette, discriminative parts, texture cues, occlusion, blur, and scale.
B17 / Negative	entities are obscured, tiny, blurred, truncated, or ambiguous. Apply the penalty in proportion to severity and affected relevant area.
B17 / Scoring	90-100 immediately identifiable; 70-89 minor ambiguity; 40-69 several uncertain entities; 0-39 most relevant entities unidentifiable.
B18 environmental_occlusion_visibility	
B18 / Definition	Evaluate visibility of relevant evidence through intervening media or surface obstruction. A higher score means veiling media, suspended particles, droplets, diffuse attenuation, or surface obstruction do not materially obscure relevant evidence, while a lower score means they substantially veil, mask, or distort it. Judge primarily from transmission loss, veil density, coverage, local contrast loss, and affected relevant regions. Weight the evidence by severity, affected relevant area, and task relevance. Do not infer the physical cause of an obstruction beyond what is visibly supported.
B18 / Positive	intervening media or surface obstruction do not materially obscure relevant evidence. Visible evidence: transmission loss, veil density, coverage, local contrast loss, and affected relevant regions.
B18 / Negative	intervening media or surface obstruction substantially veils, masks, or distorts relevant evidence. Apply the penalty in proportion to severity and affected relevant area.
B18 / Scoring	90-100: environmental obstruction is negligible; 70-89: small affected areas; 40-69: significant local obstruction; 0-39: most relevant evidence is obscured.
B19 symbol_form_integrity	
B19 / Definition	Evaluate the geometric integrity of directly inspectable symbols, marks, or compact forms. A form is inspectable only when its strokes, contours, and interior gaps can be resolved from image pixels without relying on semantic context. A higher score means visible parts, joins, terminals, enclosed spaces, and separations retain complete and plausible geometry; a lower score means they are broken, fused, missing, duplicated, or malformed. Do not treat occlusion, fading, perspective, blur, or small size as form damage unless the geometry of an inspectable form is itself visibly deformed.
B19 / Positive	directly inspectable visible symbols or compact forms with complete plausible strokes, contours, joins, enclosed spaces, and separations
B19 / Negative	directly inspectable visible forms with broken, fused, missing, duplicated, or malformed geometry

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B19 / Scoring	90-100: nearly all relevant inspectable forms are intact; 70-89: minor local defects; 40-69: repeated visible geometry defects; 0-39: relevant form geometry is broadly unusable. Mark not applicable when no relevant inspectable form is present.
B20 lighting_harmony	
B20 / Definition	Evaluate whether the visible direction, distribution, softness, and color of light form a coherent aesthetic treatment across the image. A higher score means highlights and shadows jointly reveal form and establish a consistent visual atmosphere without discordant light patches. A lower score means visible lighting direction, color, or local intensity conflicts across subjects and regions. Do not duplicate brightness suitability, clipping, localized glare, or simple visibility: a dark or bright image can still have harmonious lighting.
B20 / Positive	coherent direction, distribution, softness, and color of light supporting shape, mood, and attention
B20 / Negative	discordant visible lighting direction, color, or local intensity across image regions
B20 / Scoring	90-100 visible lighting is exceptionally coherent across the image; 70-89 minor local mismatch; 40-69 multiple discordant lighting regions; 0-39 the visible lighting treatment is broadly incoherent.
B21 lighting_Reasonableness	
B21 / Definition	Evaluate whether the visible lighting is reasonable for viewing the scene and interpreting its relevant content. A higher score means illumination is appropriately distributed and controlled: important regions are sufficiently lit, shadows retain useful evidence, highlights are not distracting or excessively clipped, and the lighting direction and intensity support natural, stable interpretation. A lower score means lighting is poorly placed, uneven, overly harsh, implausible, or otherwise makes relevant content difficult or uncomfortable to view. Judge primarily from illumination distribution, shadow detail, highlight control, lighting direction, and the effect on relevant evidence. Weight the evidence by severity, affected relevant area, and task relevance. Do not penalize a deliberately dark, bright, or dramatic scene when its lighting remains coherent and its relevant visual evidence is still interpretable.
B21 / Positive	illumination is appropriately distributed and controlled, with readable important regions, retained shadow detail, controlled highlights, and lighting that supports natural scene interpretation.
B21 / Negative	lighting is poorly placed, uneven, overly harsh, implausible, or obscures relevant evidence through unreadable shadows, distracting highlights, or unsuitable illumination.
B21 / Scoring	90-100: lighting is highly appropriate, coherent, and supportive of relevant evidence; 70-89: minor local lighting imbalance or highlight/shadow issue; 40-69: noticeable lighting problems reduce interpretation; 0-39: lighting broadly prevents reliable viewing of relevant content.
B22 material_cue_observability	
B22 / Definition	Evaluate whether visible surfaces provide recoverable material cues through texture, granularity, reflectance, translucency, gloss, or repeated pattern. A higher score means surface appearance cues are resolved, coherent within surfaces, and differentiated across relevant surfaces. A lower score means visible surface evidence is smeared, clipped, aliased, homogenized, or too ambiguous to support a stable material appearance. Judge visible surface evidence rather than object shape, color richness, generic edge sharpness, or semantic identity.
B22 / Positive	recoverable visible material cues through coherent texture granularity reflectance translucency gloss and pattern
B22 / Negative	smeared clipped aliased homogenized or ambiguous visible surface appearance cues
B22 / Scoring	90-100 visible material cues are exceptionally complete and coherent; 70-89 minor local ambiguity; 40-69 several surfaces lose distinctive appearance cues; 0-39 visible material evidence is broadly unrecoverable.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B23 microstructure_resolvability	
B23 / Definition	Evaluate resolvability of relevant fine-scale repeated, layered, granular, or textured structures. A higher score means fine structures retain distinguishable local organization, while a lower score means they are blurred, aliased, compressed, merged, or undersampled so local organization cannot be resolved. Judge primarily from small-scale texture, repeated-element separation, fine boundaries, sampling density, and local blur. Weight the evidence by severity, affected relevant area, and task relevance. Do not infer the semantic identity or physical scale of fine structures.
B23 / Positive	fine repeated, layered, granular, or textured structures retain distinguishable local organization. Visible evidence: small-scale texture, repeated-element separation, fine boundaries, sampling density, and local blur.
B23 / Negative	fine structures are blurred, aliased, compressed, merged, or undersampled so local organization cannot be resolved. Apply the penalty in proportion to severity and affected relevant area.
B23 / Scoring	90-100: relevant microstructure is clearly resolved; 70-89: minor local loss; 40-69: widespread uncertainty; 0-39: microstructure is largely unresolved. Use only where such structure is task-relevant.
B24 multiscale_resolvability	
B24 / Definition	Evaluate whether visible content remains resolvable across at least two clearly present scale bands in the same image. A higher score means global layout or large structures and the smallest visibly meaningful instances or structures both retain sufficient pixel support. A lower score means one visible scale band, usually fine or distant content, collapses even though another scale band remains usable. Judge scale bands evident from the pixels, including smallest-instance detail, scale-dependent crowding, and preservation of global context. Do not use this dimension as a second generic sharpness score.
B24 / Positive	visible large-scale context and the smallest meaningful instances or structures are both resolved across multiple scale bands
B24 / Negative	one clearly visible scale band collapses especially fine or distant content while another scale band remains usable
B24 / Scoring	90-100 global context and the smallest clearly visible evidence are both highly usable; 70-89 the smallest visible scale band has minor loss; 40-69 one clearly present scale band is substantially unresolved; 0-39 the smallest visibly meaningful content is effectively absent despite usable coarse context. Use only scale bands visible in the image and do not substitute a generic sharpness score.
B25 noise_cleanliness	
B25 / Definition	Whether the image has an appropriate cleanliness level, meaning visible random noise, grain, and sensor artifacts are well controlled and not distracting. For this dimension, a higher score means lower visible noise and cleaner image quality.
B25 / Positive	a clean image with well controlled visible noise, low distracting grain, minimal sensor or compression artifacts, and broadly readable image structure. Smooth regions should look naturally clean rather than artificially smeared, and the absence of visible noise should not come from severe blur or loss of detail
B25 / Negative	an image with inappropriate visible noise, obvious distracting grain, random speckles, sensor noise, or visible compression artifacts. Large-area defocus blur, motion blur, smeared detail, or image-wide loss of readable structure should be considered a negative factor for this dimension when it prevents viewers from clearly perceiving the image content. The score penalty should depend on the severity, spatial extent, and impact on overall readability; minor local blur or small isolated soft regions should not by itself imply a low noisiness score
B25 / Scoring	Dimension-specific score meaning: For this dimension, the 0-100 score measures how good, appropriate, and visually successful the image is on the named dimension. A higher score means the image is better on this dimension according to the dimension explanation and reference prompts. A lower score means the image shows stronger visible weaknesses on this dimension.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B26 oriented_instance_shape_integrity	
B26 / Definition	Evaluate whether the principal axis and, when visible asymmetry permits, the front-to-back heading of each visually discrete, spatially bounded instance can be inferred from its own shape structure. Count a region only when it has a closed, finite extent and appears as one object-like instance. Do not use continuous scene-layout lines, extended surfaces, open boundaries, or general repeated alignment as substitutes for bounded-instance evidence. A higher score means visible endpoints, axial parts, asymmetry, and repeated within-instance directional cues agree on a stable angle or heading. A lower score means foreshortening, near-symmetry, tiny scale, blur, or missing directional parts makes the angle or heading ambiguous. Judge orientation-specific geometry only. Do not score generic edge contrast, separation from the background, semantic identity, instance density, or overall sharpness; a bounded instance may have a clearly separated outer boundary yet receive a low score when its orientation remains ambiguous.
B26 / Positive	stable visible principal axes of discrete bounded instances and, where asymmetry is visible, recoverable front-to-back headings supported by endpoints and within-instance directional parts
B26 / Negative	ambiguous axes or headings of discrete bounded instances caused by near-symmetry, foreshortening, tiny scale, blur, or missing directional parts, without substituting continuous layouts or open structures for instance evidence
B26 / Scoring	90-100 the angles and visible headings of nearly all qualifying bounded instances are exceptionally stable; 70-89 most principal axes are recoverable with minor front-back ambiguity; 40-69 a substantial portion of bounded instances has uncertain axes or headings; 0-39 orientation-specific geometry of the bounded instances is broadly unusable.
B27 sequential_layout_traceability	
B27 / Definition	Evaluate whether a coherent sequence can be traced across visible spatially ordered elements. A higher score means transitions among aligned blocks, repeated groups, directional cues, and continuation links are visually unambiguous. A lower score means competing paths, detached elements, or broken continuation cues make the next step uncertain. Judge visible sequence cues rather than symbol recognition or general layout neatness.
B27 / Positive	an unambiguous visible sequence across ordered regions, groups, or directional elements
B27 / Negative	ambiguous visible sequence caused by competing paths, detached elements, or broken continuation cues
B27 / Scoring	90-100 the complete visible sequence is unambiguous; 70-89 one or two minor transitions require care; 40-69 several order transitions are uncertain; 0-39 no reliable sequence can be reconstructed.
B28 region_separability	
B28 / Definition	Evaluate whether semantically distinct regions are visually separable. A higher score means a relevant region has a distinguishable extent, coherent interior evidence, and separable adjacency, while a lower score means region membership is visually ambiguous because interiors blend with neighbors or relevant boundaries collapse. Judge primarily from region-to-region contrast, texture or appearance consistency, boundary continuity, occlusion, and local clutter. Weight the evidence by severity, affected relevant area, and task relevance. Do not treat ordinary subject matter, a valid style, or a property that remains visually observable as a failure unless it directly removes the stated evidence.
B28 / Positive	a relevant region has a distinguishable extent, coherent interior evidence, and separable adjacency. Visible evidence: region-to-region contrast, texture or appearance consistency, boundary continuity, occlusion, and local clutter.
B28 / Negative	region membership is visually ambiguous because interiors blend with neighbors or relevant boundaries collapse. Apply the penalty in proportion to severity and affected relevant area.
B28 / Scoring	90-100: relevant regions are consistently distinguishable; 70-89: small ambiguous areas; 40-69: frequent local mixing; 0-39: relevant regions are broadly inseparable.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B29 surface_region_observability	
B29 / Definition	Evaluate observability of relevant surface regions and their local state. A higher score means surface extent, boundaries, condition cues, and continuity are visually clear, while a lower score means surface extent or state is concealed by darkness, weather, reflection, occlusion, shadow, or blending with surroundings. Judge primarily from edge visibility, surface texture, region continuity, reflections, and obstruction. Weight the evidence by severity, affected relevant area, and task relevance. Do not infer an unshown function or semantic category for a surface.
B29 / Positive	relevant surface extent, boundaries, condition cues, and continuity are visually clear. Visible evidence: edge visibility, surface texture, region continuity, reflections, and obstruction.
B29 / Negative	relevant surface extent or state is concealed by darkness, weather, reflection, occlusion, shadow, or blending with surroundings. Apply the penalty in proportion to severity and affected relevant area.
B29 / Scoring	90-100: relevant surface regions are readily interpretable; 70-89: minor local ambiguity; 40-69: substantial ambiguous sections; 0-39: relevant surface regions are broadly unobservable.
B30 scene_relation_clarity	
B30 / Definition	Evaluate clarity of spatial, action, and interaction relations. A higher score means relative positions, contacts, containment, orientation, and interactions are unambiguous, while a lower score means relations are ambiguous due to occlusion, cropping, overlap, weak context, or unclear orientation. Judge primarily from relative position, contact, overlap, directed configuration, relative alignment, depth ordering, and context. Weight the evidence by severity, affected relevant area, and task relevance. Do not infer an interaction that is not supported by visible evidence.
B30 / Positive	relative positions, contacts, containment, orientation, and interactions are unambiguous. Visible evidence: relative position, contact, overlap, directed configuration, relative alignment, depth ordering, and context.
B30 / Negative	relations are ambiguous due to occlusion, cropping, overlap, weak context, or unclear orientation. Apply the penalty in proportion to severity and affected relevant area.
B30 / Scoring	90-100 relations directly observable; 70-89 small ambiguity; 40-69 several uncertain; 0-39 broadly ambiguous.
B31 semantic_content_appeal	
B31 / Definition	Evaluate the visual and semantic interest expressed by the depicted subjects, actions, expressions, events, or visible ideas, independently of rendering quality and composition. A higher score means the visible content contains a distinctive moment, expressive interaction, unusual subject, or clearly engaging visual idea. A lower score means the visible content is generic, repetitive, emotionally flat, or lacks a discernible point of interest. Judge only what is depicted in the image rather than sharpness, color treatment, or composition.
B31 / Positive	engaging subject matter, moment, expression, event, or visual idea with sustained semantic interest
B31 / Negative	generic, repetitive, emotionally flat content with little discernible semantic interest
B31 / Scoring	90-100 the depicted content contains exceptionally distinctive and compelling visible interest; 70-89 the visible content is clearly engaging; 40-69 visible interest is limited or inconsistent; 0-39 little discernible semantic point of interest is visible.
B32 spatial_topology_continuity	
B32 / Definition	Evaluate continuity and connectivity evidence for relevant linear or branching networks. A higher score means network segments and junctions can be traced through their visible extent, while a lower score means they are broken, occluded, merged, or ambiguous. Judge primarily from line continuity, junction visibility, crossing geometry, occlusion, and gaps. Weight the evidence by severity, affected relevant area, and task relevance. Do not infer a network's semantic type from its appearance.
B32 / Positive	relevant linear or branching networks can be traced through continuous segments and junctions. Visible evidence: line continuity, junction visibility, crossing geometry, occlusion, and gaps.
B32 / Negative	network structures are broken, occluded, merged, or ambiguous. Apply the penalty in proportion to severity and affected relevant area.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B32 / Scoring	90-100 topology easy to trace; 70-89 small gaps; 40-69 repeated ambiguous connections; 0-39 connectivity unusable.
B33 structural_coherence	
B33 / Definition	Evaluate coherence of spatial structures and object shapes. A higher score means lines, surfaces, and object parts form geometrically consistent structures, while a lower score means relevant structures are warped, broken, duplicated, or internally inconsistent. Judge primarily from line continuity, shape consistency, perspective, repeated patterns, and region adjacency. Weight the evidence by severity, affected relevant area, and task relevance. Do not treat ordinary subject matter, artistic style, or an unusual but still observable property as a failure unless it directly removes the stated visual evidence.
B33 / Positive	lines, surfaces, and object parts form geometrically consistent structures. Visible evidence: line continuity, shape consistency, perspective, repeated patterns, and region adjacency.
B33 / Negative	relevant structures are warped, broken, duplicated, or internally inconsistent. Apply the penalty in proportion to severity and affected relevant area.
B33 / Scoring	90-100 coherent; 70-89 small irregularities; 40-69 clear disruption; 0-39 pervasive incoherence.
B34 structured_pattern_parseability	
B34 / Definition	Evaluate whether the internal organization of visible structured patterns can be parsed from the image. A higher score means repeated units, ordered groups, bounded subregions, branches, separators, and associations are visually recoverable. A lower score means alignment, grouping, separators, or associations collapse. Judge internal organization rather than semantic interpretation or overall composition.
B34 / Positive	internally parseable structured patterns with recoverable repeated units, ordered groups, bounded subregions, separators, and associations
B34 / Negative	structured patterns whose separators, alignment, grouping, or associations collapse
B34 / Scoring	90-100 all relevant internal relations are recoverable; 70-89 minor local ambiguity; 40-69 several visible units, groups, or associations are uncertain; 0-39 the structured pattern cannot be reliably parsed.
B35 subject_emphasis_clarity	
B35 / Definition	Evaluate whether the visible image establishes an unambiguous dominant subject or focal hierarchy. A higher score means framing, scale, contrast, isolation, pose, or directional cues consistently concentrate attention on one subject or an ordered set of subjects. A lower score means competing elements, weak hierarchy, awkward cropping, or background competition leave the focal structure unclear. Judge visible emphasis rather than category recognizability, semantic appeal, or generic background cleanliness.
B35 / Positive	an unambiguous visually dominant subject supported by coherent framing, scale, isolation, and attention cues
B35 / Negative	weak or conflicting visible focal hierarchy with competing elements and unclear emphasis
B35 / Scoring	90-100 the visible focal hierarchy is immediate and coherent; 70-89 minor competition; 40-69 several elements compete without a stable order; 0-39 no reliable dominant subject or focal hierarchy is visible.
B36 foreground_background_separability	
B36 / Definition	Evaluate visual separation of relevant foreground marks, objects, or regions from their immediate background. A higher score means foreground evidence remains distinct from the surface behind it, while a lower score means it blends into similarly colored, textured, reflective, or cluttered surroundings. Judge primarily from local luminance and color contrast, texture competition, reflections, and boundary visibility. Weight the evidence by severity, affected relevant area, and task relevance.
B36 / Positive	relevant foreground evidence remains distinct from the surface behind it. Visible evidence: local luminance and color contrast, texture competition, reflections, and boundary visibility.
B36 / Negative	relevant foreground evidence blends into similarly colored, textured, reflective, or cluttered background. Apply the penalty in proportion to severity and affected relevant area.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B36 / Scoring	90-100 strongly separable; 70-89 modest competition; 40-69 frequent blending; 0-39 foreground and background are broadly merged.
B37 symbol_geometric_regularity	
B37 / Definition	Evaluate geometric stability of visible symbols, marks, or repeated compact forms. A higher score means alignment, proportions, spacing, and projection are stable enough for visual interpretation, while a lower score means warp, curvature, skew, stretch, or irregular spacing impairs interpretation. Judge primarily from alignment continuity, orientation, aspect ratio, spacing, and perspective distortion. Weight the evidence by severity, affected relevant area, and task relevance.
B37 / Positive	alignment, proportions, spacing, and perspective are stable enough for visual interpretation. Visible evidence: alignment continuity, orientation, aspect ratio, spacing, and perspective distortion.
B37 / Negative	severe warp, curvature, skew, stretch, or irregular spacing impairs visual interpretation. Apply the penalty in proportion to severity and affected relevant area.
B37 / Scoring	90-100 stable geometry; 70-89 mild effects; 40-69 noticeable distortion; 0-39 severe geometry.
B38 symbol_legibility	
B38 / Definition	Evaluate interpretability of visible symbols, marks, and compact graphical forms. A higher score means their strokes, contours, or distinctive parts are sharp, sufficiently sampled, and discriminable, while a lower score means they are too small, blurred, distorted, occluded, or low contrast to distinguish. Judge primarily from stroke or contour clarity, separation, resolution, perspective, blur, and occlusion. Weight the evidence by severity, affected relevant area, and task relevance.
B38 / Positive	relevant symbols, marks, or compact forms are sharp, sufficiently sampled, and distinguishable. Visible evidence: stroke or contour clarity, separation, resolution, perspective, blur, and occlusion.
B38 / Negative	relevant symbols or compact forms are too small, blurred, distorted, occluded, or low contrast to distinguish. Apply the penalty in proportion to severity and affected relevant area.
B38 / Scoring	90-100 most relevant forms are readily distinguishable; 70-89 minor uncertainty; 40-69 only some are distinguishable; 0-39 relevant forms are largely uninterpretable.
B39 thin_linear_structure_traceability	
B39 / Definition	Evaluate traceability of relevant thin and elongated structures. A higher score means relevant narrow lines, paths, markings, or strokes can be followed continuously through their extent, while a lower score means they are broken, merged, occluded, blurred, or too weak to trace reliably. Judge primarily from line continuity, local contrast, width support, junction visibility, crossings, and gap severity. Weight the evidence by severity, affected relevant area, and task relevance. Do not infer the semantic type of a thin structure from its appearance.
B39 / Positive	relevant narrow lines, paths, markings, or strokes can be followed continuously through their extent. Visible evidence: line continuity, local contrast, width support, junction visibility, crossings, and gap severity.
B39 / Negative	relevant thin structures are broken, merged, occluded, blurred, or too weak to trace reliably. Apply the penalty in proportion to severity and affected relevant area.
B39 / Scoring	90-100: relevant linear structures are readily traceable; 70-89: minor short gaps; 40-69: repeated ambiguous segments; 0-39: paths cannot be followed reliably.
B40 visual_symbol_recognizability	
B40 / Definition	Evaluate recognizability of relevant visual symbols, icons, signs, signals, and non-text graphical indicators. A higher score means relevant symbols have sufficient scale, contrast, shape integrity, and visibility for confident visual interpretation, while a lower score means they are too small, blurred, occluded, overexposed, distorted, or ambiguous to interpret. Judge primarily from symbol shape, distinctive internal cues, color state where relevant, scale, occlusion, blur, and local contrast. Weight the evidence by severity, affected relevant area, and task relevance.
B40 / Positive	relevant visual symbols have sufficient scale, contrast, shape integrity, and visibility for confident visual interpretation. Visible evidence: symbol shape, distinctive internal cues, color state where relevant, scale, occlusion, blur, and local contrast.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
B40 / Negative	relevant visual symbols are too small, blurred, occluded, overexposed, distorted, or ambiguous to interpret. Apply the penalty in proportion to severity and affected relevant area.
B40 / Scoring	90-100: relevant symbols are immediately interpretable; 70-89: minor uncertainty; 40-69: several relevant symbols are uncertain; 0-39: most relevant symbols are unusable.
MD61-Open: Autonomous driving	
O01 driving_perception_evidence_mos	
O01 / Definition	Evaluate the joint visibility of evidence needed to interpret a road scene: visible road users, drivable surfaces, lane markings, and their spatial boundaries. A higher score means that these elements are simultaneously resolvable, separable, and spatially coherent across the image, while a lower score means that blur, darkness, glare, noise, weather, occlusion, or weak contrast materially removes one or more kinds of evidence. Judge the joint clarity of visible road-scene elements, weight a severe bottleneck more than a small cosmetic defect, and do not reward an empty or simple road merely for containing fewer elements.
O01 / Positive	Visible road users, drivable surfaces, lane markings, and spatial boundaries are jointly clear, separable, and coherent.
O01 / Negative	One or more major kinds of road-scene evidence are unresolved, merged, concealed, or spatially ambiguous over consequential regions.
O01 / Scoring	90-100: all major visible road-scene evidence is strong. 75-89: broadly clear with minor local limitations. 50-74: usable but at least one evidence type is materially weak. 25-49: several evidence types are difficult to recover. 0-24: the road scene is broadly uninterpretable.
O02 road_user_instance_resolvability	
O02 / Definition	Evaluate whether every visible road user, including vehicles, riders, cyclists, and pedestrians, retains enough shape, extent, local contrast, and separation to be individually resolved. A higher score means that both large and small road users have visible instance structure and recoverable extents, while a lower score means that they collapse into indistinct marks, merge with nearby objects, or lose defining parts through blur, occlusion, darkness, glare, or background blending. Judge all visible road users and weight the fraction and severity of affected instances. If no road user is visible, mark the dimension not applicable.
O02 / Positive	Visible road users retain individual shape, extent, local contrast, and separation, including small or distant instances.
O02 / Negative	Visible road users are blurred, merged, hidden, undersampled, or blended so that individual instances and extents are difficult to resolve.
O02 / Scoring	90-100: nearly every visible road user is readily resolved. 75-89: most are clear with limited difficult instances. 50-74: several instances are marginal. 25-49: most instances are poorly resolved. 0-24: visible road users cannot be individually recovered. Use 0 only as the score placeholder when not applicable.
O03 drivable_surface_boundary_clarity	
O03 / Definition	Evaluate whether visible road-like traversable surfaces form coherent regions with traceable boundaries against sidewalks, curbs, shoulders, medians, obstacles, and non-road surroundings. A higher score means that surface interiors and outer limits remain visually distinct and continuous, while a lower score means that shadow, reflection, glare, weather, occlusion, weak contrast, or texture loss makes the surface extent ambiguous. Judge all visible road-surface regions from their pixels alone. Do not assume a hidden route or vehicle-specific driving policy.
O03 / Positive	Visible road-surface interiors are coherent and their boundaries against surrounding regions are continuous and distinct.
O03 / Negative	Road-surface extent is visually ambiguous because interiors fragment or boundaries disappear into adjacent regions.
O03 / Scoring	90-100: surface extent and boundaries are unambiguous. 75-89: minor localized uncertainty. 50-74: repeated ambiguous sections remain. 25-49: large surface regions are difficult to separate. 0-24: road-surface extent is broadly unrecoverable.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
O04 lane_marking_traceability	
O04 / Definition	Evaluate whether every visible lane marking can be followed as a thin linear structure through its visible extent. A higher score means that lane strokes retain width, contrast, continuity, orientation, and separation from road texture, while a lower score means that they fade, blur, break, merge with glare or shadows, or become confused with surface patterns. Judge only visible painted or physically marked lane evidence. If no lane marking is visible, mark the dimension not applicable.
O04 / Positive	Visible lane markings retain continuous strokes, clear orientation, stable width, and separation from the road surface.
O04 / Negative	Visible lane markings fade, fragment, blur, merge with surface texture, or lose a traceable path.
O04 / Scoring	90-100: nearly all visible markings are continuously traceable. 75-89: minor gaps or weak segments. 50-74: several markings are intermittently traceable. 25-49: most markings are fragmented or ambiguous. 0-24: visible markings cannot be followed. Use 0 only as the score placeholder when not applicable.
O05 road_scene_exposure_evidence_preservation	
O05 / Definition	Evaluate whether the visible road scene preserves structural evidence across shadows, midtones, and highlights. A higher score means that road users, surface boundaries, and lane details remain visible without consequential black crush, highlight clipping, flare, or local glare, while a lower score means that exposure loss erases these structures. Weight the affected visible area and the number of evidence types lost rather than the global mean brightness. Do not penalize a naturally dark or bright scene when its structures remain observable.
O05 / Positive	Shadows, midtones, and highlights preserve visible road users, surface boundaries, and lane structure without consequential clipping or glare.
O05 / Negative	Blocked shadows, blown highlights, flare, or glare erase visible road users, surface boundaries, or lane structure.
O05 / Scoring	90-100: exposure preserves nearly all visible road-scene structure. 75-89: minor localized loss. 50-74: several regions lose useful evidence. 25-49: major structures are clipped or blocked. 0-24: exposure failure removes most road-scene evidence.
O06 localized_road_glare_nonobstruction	
O06 / Definition	Evaluate how fully visible road users, road-surface boundaries, and lane markings remain readable inside and around localized bright veils, flare, reflections, or glare patches. A higher score means that bright interference is absent or preserves underlying shape, edge, and texture evidence, while a lower score means that one or more bright regions wash out consequential road-scene structure. Judge the visible overlap between bright interference and road-scene evidence, weighting both opacity and affected area. Do not penalize ordinary bright objects, lamps, sky, or highlights when their pixels do not conceal surrounding structure.
O06 / Positive	Localized bright regions are absent or preserve the underlying shapes, boundaries, textures, and lane strokes of the visible road scene.
O06 / Negative	Glare, flare, reflection, or bright veil washes out visible road users, road-surface boundaries, or lane markings over consequential areas.
O06 / Scoring	90-100: no consequential glare obstruction and underlying road evidence is fully preserved. 75-89: small or translucent bright interference with minor evidence loss. 50-74: one or more localized regions partially erase road evidence. 25-49: large or opaque glare removes major structures. 0-24: bright interference broadly prevents interpretation of the road scene.
MD61-Open: Remote sensing	
O07 aerial_segmentation_evidence_mos	
O07 / Definition	Evaluate the joint visibility of region interiors, object footprints, and boundaries in overhead or near-orthographic imagery. A higher score means that visible ground regions and structures retain coherent appearance, distinct adjacency, traceable outlines, and usable spatial sampling, while a lower score means that haze, cloud, shadow, compression, blur, mixed pixels, or sensor artifacts cause regions to merge, fragment, or disappear. Judge the complete visible footprint and protect small structures from being ignored by large uniform areas. Base the judgment entirely on visible region interiors, footprints, adjacency, outlines, and spatial detail.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
O07 / Positive	Ground regions and visible structures have coherent interiors, separable appearance, traceable outlines, and sufficient spatial detail.
O07 / Negative	Ground regions or structures merge, fragment, disappear, or become poorly bounded because visible evidence is degraded.
O07 / Scoring	90-100: nearly all visible regions and structures are clearly separable. 75-89: reliable with minor local ambiguity. 50-74: usable but repeated region or boundary difficulty. 25-49: major areas or many small structures are ambiguous. 0-24: spatial partition evidence is broadly absent.
O08 aerial_small_structure_resolvability	
O08 / Definition	Evaluate whether the smallest visible man-made or natural structures in overhead imagery retain coherent footprints, edges, local contrast, and separation from surrounding texture. A higher score means that small structures remain distinct spatial entities, while a lower score means that they collapse into mixed pixels, aliasing, blur, compression blocks, or background texture. Judge the smallest recurring visible structure band rather than overall sharpness. Do not require knowing a structure's semantic class.
O08 / Positive	Small visible overhead structures retain coherent footprints, local contrast, and separable spatial extent.
O08 / Negative	Small overhead structures collapse into mixed pixels, blur, aliasing, compression blocks, or surrounding texture.
O08 / Scoring	90-100: the smallest visible structures are consistently resolved. 75-89: minor uncertainty. 50-74: several are marginal. 25-49: most small structures are unresolved. 0-24: small-structure evidence is effectively absent.
O09 aerial_region_boundary_clarity	
O09 / Definition	Evaluate the visibility and continuity of boundaries between adjacent ground regions and visible object footprints in overhead imagery. A higher score means that outlines remain spatially stable, locally contrasted, and traceable around corners and narrow sections, while a lower score means that boundaries soften, staircase, break, or merge because of blur, haze, shadow, compression, or mixed pixels. Judge boundary evidence only, not region identity or aesthetic neatness.
O09 / Positive	Adjacent aerial regions and object footprints have stable, locally distinct, and continuously traceable boundaries.
O09 / Negative	Aerial boundaries soften, break, staircase, or merge so that region extents become uncertain.
O09 / Scoring	90-100: nearly all consequential boundaries are precise and continuous. 75-89: minor local uncertainty. 50-74: repeated weak or broken sections. 25-49: major outlines are difficult to follow. 0-24: region boundaries are broadly unrecoverable.
O10 road_network_traceability	
O10 / Definition	Evaluate whether every visible road-like linear surface in overhead imagery can be followed through segments, bends, intersections, crossings, and partial obstructions. A higher score means that road width, local contrast, continuity, and junction connections remain visually recoverable, while a lower score means that segments fade, merge with surroundings, break, or form ambiguous connections. Weight topology-critical gaps and junction ambiguity more than small edge roughness. If no road-like network is visible, mark the dimension not applicable.
O10 / Positive	Visible road-like segments remain distinct and continuous, and their bends, branches, crossings, and junction connections are traceable.
O10 / Negative	Road-like segments fade, merge, break, or form ambiguous junctions that prevent continuous visual tracing.
O10 / Scoring	90-100: the visible network and junctions are almost completely traceable. 75-89: minor gaps or local ambiguity. 50-74: repeated gaps or uncertain connections. 25-49: major branches are disconnected or confused. 0-24: visible network structure is unrecoverable. Use 0 only as the score placeholder when not applicable.
O11 atmospheric_ground_evidence_preservation	
O11 / Definition	Evaluate how well visible ground texture, region boundaries, and small structures survive atmospheric veil, haze, cloud, smoke, or broad low-frequency obscuration in overhead imagery. A higher score means that surface contrast and spatial evidence remain recoverable across the footprint, while a lower score means that the obscuring layer erases or homogenizes ground evidence. Weight severity and affected area. Do not penalize naturally low-contrast terrain when its local structures remain visible.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
O11 / Positive	Ground textures, region boundaries, and small structures remain visible through minimal atmospheric or broad veiling interference.
O11 / Negative	Haze, cloud, smoke, or broad veil suppresses ground contrast and erases visible spatial structure over consequential areas.
O11 / Scoring	90-100: ground evidence is almost entirely preserved. 75-89: mild localized veil. 50-74: repeated loss across visible regions. 25-49: large portions are obscured. 0-24: ground evidence is almost entirely concealed.
MD61-Open: Medical imaging	
O12 medical_structure_evidence_mos	
O12 / Definition	Evaluate the joint visibility of anatomical shape, tissue organization, local texture, boundaries, and intensity structure in a medical image. A higher score means that these visible structures are simultaneously coherent, resolved, and minimally obscured, while a lower score means that blur, noise, clipping, weak contrast, artifacts, or spatial distortion materially removes structural evidence. Judge only visible pixels and weight losses within the depicted anatomy more than empty borders or presentation background. Base the judgment entirely on visible anatomy, tissue organization, local texture, boundaries, and intensity structure.
O12 / Positive	Visible anatomy, tissue organization, local texture, boundaries, and intensity structure are jointly coherent and well preserved.
O12 / Negative	Visible medical structure is materially obscured, unresolved, clipped, distorted, or contaminated across consequential anatomical regions.
O12 / Scoring	90-100: visible medical structure is exceptionally well preserved. 75-89: strong evidence with minor local limitations. 50-74: usable but visibly degraded. 25-49: substantial structural evidence is lost. 0-24: the depicted anatomy is broadly uninterpretable.
O13 anatomical_boundary_visibility	
O13 / Definition	Evaluate whether visible boundaries between anatomical structures, tissue layers, lesions-like regions, cavities, or specimen components remain locally distinct and spatially traceable. A higher score means that boundaries retain contrast, continuity, and stable position, while a lower score means that they blur, merge, fragment, or become hidden by noise, clipping, glare, or artifacts. Judge boundaries that are visibly present without deciding what they represent medically.
O13 / Positive	Visible anatomical and tissue boundaries retain local contrast, continuity, and traceable spatial extent.
O13 / Negative	Visible anatomical or tissue boundaries blur, merge, fragment, or disappear into adjacent structure.
O13 / Scoring	90-100: nearly all visible boundaries are distinct and continuous. 75-89: minor local uncertainty. 50-74: several boundaries are weak or intermittent. 25-49: major boundaries are difficult to trace. 0-24: boundary evidence is broadly absent.
O14 tissue_texture_discriminability	
O14 / Definition	Evaluate whether visibly different tissue or specimen textures remain distinguishable from one another by local pattern, granularity, orientation, color, or intensity variation. A higher score means that texture families preserve separable internal statistics, while a lower score means that blur, noise, compression, oversmoothing, or contrast collapse makes different regions look alike. Judge texture distinction only where multiple visible textures are present; otherwise mark the dimension not applicable.
O14 / Positive	Visible tissue or specimen textures retain distinct local patterns, granularity, orientation, color, or intensity variation.
O14 / Negative	Different visible textures become confounded because their local patterns are blurred, noisy, oversmoothed, or contrast-collapsed.
O14 / Scoring	90-100: visible texture families are readily distinguishable. 75-89: mostly distinct with minor overlap. 50-74: several regions are difficult to distinguish. 25-49: most textures are confounded. 0-24: visible texture distinction is lost. Use 0 only as the score placeholder when not applicable.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
O15 medical_microstructure_preservation	
O15 / Definition	Evaluate whether the finest visible structures in the medical image retain separable edges, internal pattern, and spatial continuity at the displayed resolution. A higher score means that fine layers, vessels, fibers, cells, pores, nodules, or similarly small visible structures remain resolved, while a lower score means that they collapse through undersampling, blur, noise, compression, or oversmoothing. Judge the finest actually visible structure band and do not assume structures that cannot be seen.
O15 / Positive	The finest visible medical structures retain separable edges, internal pattern, and spatial continuity.
O15 / Negative	Fine visible medical structures collapse into blur, noise, compression texture, or oversmoothed regions.
O15 / Scoring	90-100: fine visible structures are consistently resolved. 75-89: minor loss. 50-74: several fine structures are marginal. 25-49: most fine structure is unresolved. 0-24: visible microstructural evidence is absent.
O16 medical_intensity_structure_separation	
O16 / Definition	Evaluate whether visible structures remain separated across the image's intensity or color range. A higher score means that dark, midtone, and bright regions retain internal variation and do not clip or compress into indistinguishable masses, while a lower score means that black crush, highlight saturation, weak contrast, color collapse, or uneven illumination merges visible structures. Judge structural separation rather than aesthetic contrast and ignore empty presentation borders.
O16 / Positive	Dark, midtone, bright, or differently colored medical structures retain internal variation and visible separation.
O16 / Negative	Clipping, compressed range, color collapse, or uneven illumination merges visible medical structures into indistinguishable regions.
O16 / Scoring	90-100: intensity and color range preserve nearly all visible structure. 75-89: minor local compression. 50-74: several regions lose separation. 25-49: major structures merge or clip. 0-24: intensity structure is broadly unusable.
MD61-Open: Depth estimation	
O17 monocular_depth_evidence_mos	
O17 / Definition	Evaluate the joint visibility and consistency of depth-ordering evidence in a single image, including occlusion boundaries, perspective convergence, relative scale, surface gradients, texture changes, and contact relationships. A higher score means that near-far ordering and surface layout are visually coherent across the scene, while a lower score means that blur, clipping, darkness, glare, texture loss, reflection, transparency, or structural ambiguity removes or contradicts these cues. Judge depth ordering entirely from the listed geometric and appearance cues visible in the image.
O17 / Positive	Occlusion, perspective, relative scale, surface gradients, texture changes, and contact relationships provide coherent visible depth ordering.
O17 / Negative	Visible depth cues are erased, contradictory, or spatially ambiguous across consequential surfaces and boundaries.
O17 / Scoring	90-100: depth ordering and surface layout are strongly supported. 75-89: mostly coherent with minor ambiguous regions. 50-74: usable but several surfaces lack stable cues. 25-49: major surfaces or orderings are ambiguous. 0-24: visible depth structure is broadly unrecoverable.
O18 occlusion_boundary_depth_clarity	
O18 / Definition	Evaluate whether visible contours where one surface or object interrupts another provide clear near-far ordering. A higher score means that occluding edges are locally distinct, continuous, and consistently attached to the foreground surface, while a lower score means that blur, low contrast, transparency, glare, or texture confusion makes ownership and ordering ambiguous. Judge only visible overlap boundaries; do not use semantic assumptions about which object should be nearer.
O18 / Positive	Visible overlap boundaries are distinct, continuous, and provide consistent foreground-background ordering.
O18 / Negative	Overlap boundaries are blurred, broken, transparent, or locally confused so that edge ownership and depth ordering are ambiguous.

Continued on next page

Table G.1 (continued)

ID / Field	Original definition
O18 / Scoring	90-100: nearly all consequential occlusion boundaries give clear ordering. 75-89: minor local uncertainty. 50-74: several ambiguous boundaries. 25-49: major ordering edges are unclear. 0-24: occlusion-based ordering is broadly unavailable.
O19 perspective_scale_cue_consistency	
O19 / Definition	Evaluate whether visible perspective convergence, repeated-pattern scaling, foreshortening, and relative-size changes form a consistent spatial progression. A higher score means that these geometric image cues agree and remain observable across the frame, while a lower score means that distortion, blur, crop, low detail, or contradictory structure weakens the progression. Judge only the relative geometric progression directly visible among structures in the image.
O19 / Positive	Perspective convergence, repeated-pattern scaling, foreshortening, and relative-size changes form a clear and consistent spatial progression.
O19 / Negative	Perspective or scale progression is weak, erased, distorted, or internally contradictory across visible structures.
O19 / Scoring	90-100: geometric progression is strong and consistent. 75-89: mostly clear with minor local weakness. 50-74: partial or intermittent cues. 25-49: cues are weak or contradictory. 0-24: perspective and scale progression is absent or unusable.
O20 surface_gradient_depth_observability	
O20 / Definition	Evaluate whether visible surfaces retain gradual shading, texture-density, focus, and orientation changes that reveal their shape and depth variation. A higher score means that gradients are smooth enough to follow yet locally detailed enough to distinguish plane changes, while a lower score means that clipping, flat illumination, blur, noise, or texturelessness erases surface variation. Judge visible surface evidence without assuming material identity or a light-source model.
O20 / Positive	Visible surfaces retain coherent shading, texture-density, focus, and orientation gradients that reveal shape variation.
O20 / Negative	Surface gradients are clipped, flattened, blurred, noisy, or textureless so that shape variation becomes ambiguous.
O20 / Scoring	90-100: surface shape gradients are consistently observable. 75-89: minor local ambiguity. 50-74: several surfaces have weak gradients. 25-49: major surfaces appear spatially indeterminate. 0-24: surface-gradient evidence is broadly absent.
O21 localized_highlight_depth_evidence_preservation	
O21 / Definition	Evaluate whether localized bright, veiled, reflective, or clipped regions preserve visible evidence of surface shape and near-far ordering. A higher score means that contours, occlusion edges, texture gradients, and contact relationships remain readable within and around bright regions, while a lower score means that highlight saturation or veiling removes these cues and creates spatially ambiguous patches. Weight the opacity, area, and geometric importance of the affected visible regions. Judge only contours, gradients, contacts, and near-far relationships directly visible within and around the bright regions.
O21 / Positive	Localized bright or veiled regions preserve visible contours, surface gradients, contacts, and near-far ordering.
O21 / Negative	Highlight clipping or bright veil erases contours, surface gradients, contacts, or depth ordering over consequential regions.
O21 / Scoring	90-100: bright regions do not remove consequential depth evidence. 75-89: minor localized cue loss. 50-74: several bright regions make surface layout partly ambiguous. 25-49: major contours or surfaces are washed out. 0-24: highlight interference broadly destroys visible depth structure.

G.1 Similarity to Base Dimensions

Table G.2 relates the 21 held-out dimensions to the most similar base dimensions by evaluation scope, visible evidence, and scoring direction. Composite criteria may correspond to multiple base dimensions. IDs follow Table B.1.

Table G.2: Semantic correspondence between held-out and similar base dimensions.

Held-out dimension	Similar base IDs	Evaluation focus
O01 driving_perception_evidence_mos	B29, B17, B39	Joint road-scene evidence.
O02 road_user_instance_resolvability	B17, B12, B24	Road-user scope and instance separation.
O03 drivable_surface_boundary_clarity	B29, B28, B06	Road-surface extent and boundaries.
O04 lane_marking_traceability	B39, B29	Lane-specific thin-line tracing.
O05 road_scene_exposure_evidence_preservation	B07, B11, B21	Exposure-related structural evidence.
O06 localized_road_glare_nonobstruction	B21, B29	Localized glare over road evidence.
O07 aerial_segmentation_evidence_mos	B28, B06, B24, B03	Joint aerial region and boundary evidence.
O08 aerial_small_structure_resolvability	B24, B15, B26	Smallest visible aerial structures.
O09 aerial_region_boundary_clarity	B06, B28	Aerial boundaries and footprint outlines.
O10 road_network_traceability	B32, B39	Road-network connectivity and junctions.
O11 atmospheric_ground_evidence_preservation	B03, B18	Ground evidence under atmospheric obscuration.
O12 medical_structure_evidence_mos	B02, B23, B11	Joint anatomical and tissue evidence.
O13 anatomical_boundary_visibility	B02, B06	Visible anatomical and tissue boundaries.
O14 tissue_texture_discriminability	B22, B23, B28	Separation between tissue textures.
O15 medical_microstructure_preservation	B23, B15, B39	Finest visible medical structures.
O16 medical_intensity_structure_separation	B11, B07, B28	Medical structure across intensity and color ranges.
O17 monocular_depth_evidence_mos	B14, B33, B30	Joint monocular depth cues.
O18 occlusion_boundary_depth_clarity	B14, B06, B30	Occlusion ordering and edge ownership.
O19 perspective_scale_cue_consistency	B14, B33	Perspective and scale progression.
O20 surface_gradient_depth_observability	B14, B22, B15	Surface gradients supporting depth.
O21 localized_highlight_depth_evidence_preservation	B21, B14	Depth cues within and around bright regions.