UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

Shuo Cao*, Jiayang Li*, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin,
Yu Zhou, Xiangyu Chen, Wenlong Zhang, Bin Fu, Yu Qiao, Yihao Liu†

UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- UniPercept-Bench: Unified perceptual-level image understanding benchmark for MLLM
- UniPercept: Domain-adaptive pretraining + Task-aligned RL
- UniPercept as reward model and evaluation metric

UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

Why Perceptual-level Image Understanding?

- Perceptual-level image understanding focuses on how an image *looks and feels*—capturing aesthetics, quality degradations, structural regularity, and surface texture.
- Fundamentally different from semantic recognition, yet remain underexplored in MLLMs.
- **UniPercept:** a unified framework that defines, evaluates, and improves perceptual-level visual understanding across the IAA, IQA, and ISTA domains.

Semantic-level Understanding

Image Captioning
Q: Generate a natural-language description of the scene.
A: The image shows a grand cathedral with intricate spires and stone carvings under a clear blue sky, surrounded by people in the open square.

Object Visual Reasoning
Q: What are the people looking at near the shore?
A: The sunset over the ocean.

Interpreting Context & Meaning

Perceptual-level Understanding

Aesthetics Assessment
Q: Evaluate visual balance.
A: The central placement of the cat is a key factor in achieving visual balance in this image. The cat acts as the focal point, drawing immediate attention and anchoring ...

Structure Assessment
Q: Assess surface texture.
A: The image depicts a surface with a crumpled texture, characterized by irregular folds, creases, and a matte finish. The characteristics ...

Perceiving Appearance & Quality

IAA (Aesthetics): Composition / Design / Visual Elements / Structure / Creativity / Communication ...

IQA (Quality): Distortion Location / Distortion Severity / Distortion Type ...

ISTA (Structure & Texture): Scene Decomposition / Physical Structure / Material Representation ...

UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

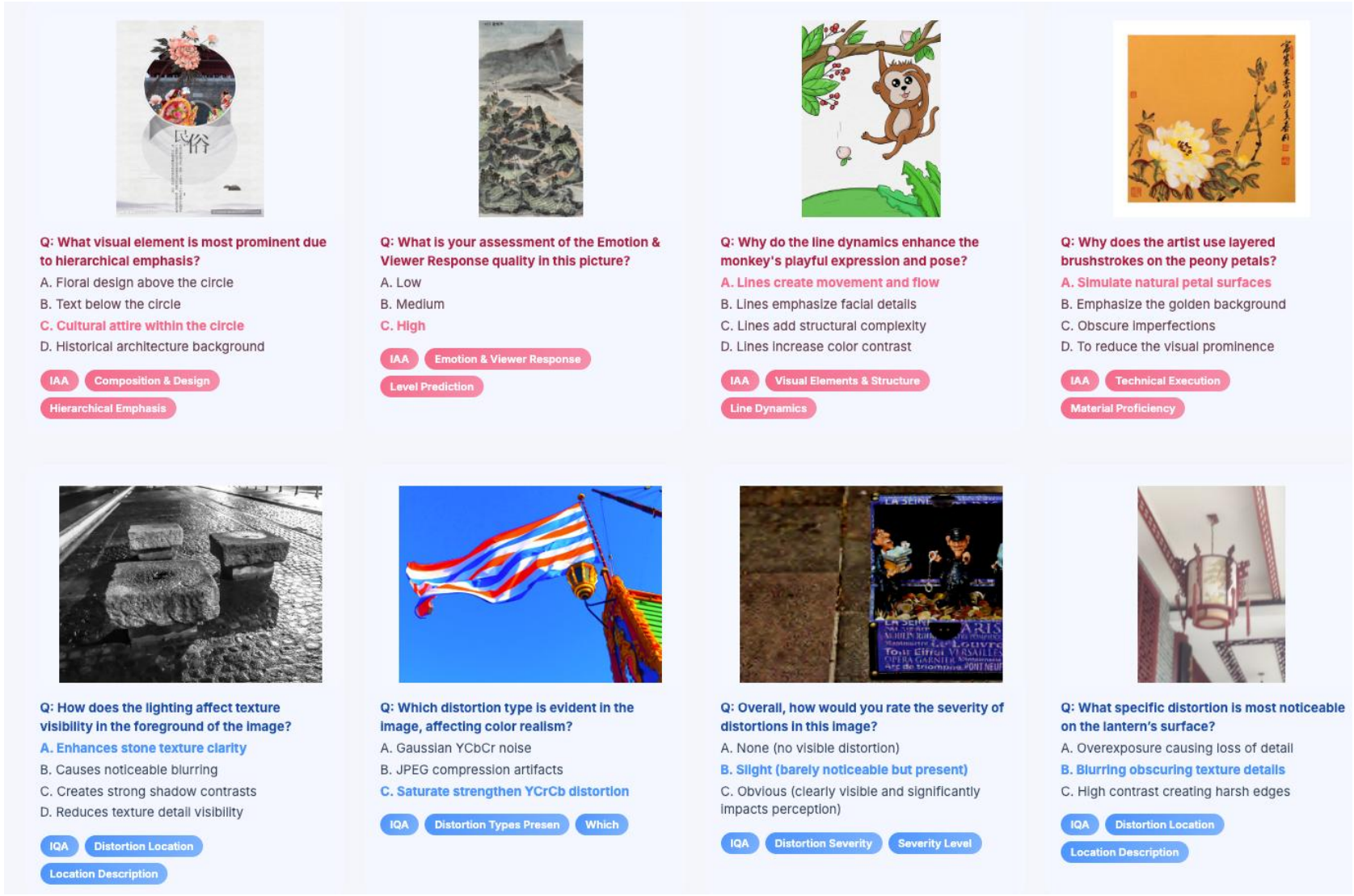
ISTA (Image Strcuture & Texture Assessment)

Table 10. **Definition details of ISTA domain.**

No.	Category	Criterion	Description
1	Scene Decomposition Principles (***Scene.***)	Scene Classification	Classify the scene as (A) Single-Object Scene with one primary subject, or (B) Composite Scene containing multiple distinguishable components. Describe the main object(s) clearly.
2	Physical Structure (***Phys.***)	Base Morphology	Describe surface texture using perceptual descriptors such as fibrous, grooved, marbled, veined, smooth, etc.
3	Physical Structure	Spatial Arrangement	Describe texture orientation, distribution pattern, and density variation across regions (e.g., horizontal, clustered, layered, radial, uniform).
4	Material Representation (***Mat.***)	Material Identification	Identify the perceived material category using a standardized taxonomy (natural, man-made, or environmental materials) present in the image.
5	Material Representation	Surface Behavior	Describe optical surface properties such as glossiness, translucency, or matte finish.
6	Geometric Composition (***Geo.***)	2D Contour	Classify the 2D outline shape using a standardized lexicon including basic shapes, polygons, special forms, or organic/curved forms.
7	Geometric Composition	3D Volume	Describe the implied 3D volumetric form using a taxonomy of basic solids, polyhedra, prisms, or complex mathematical shapes.
8	Semantic Perception (***Sem.***)	Functional Suggestion	Infer functional or symbolic implications of textures/motifs based on appearance, referencing standardized functional texture/style descriptors.
9	Semantic Perception	Stylistic Classification	Assign a stylistic category (e.g., Minimalist, Gothic, Art Deco, Futuristic, Cyberpunk, Chinese Cloud Pattern) based on visual elements and decorative cues.

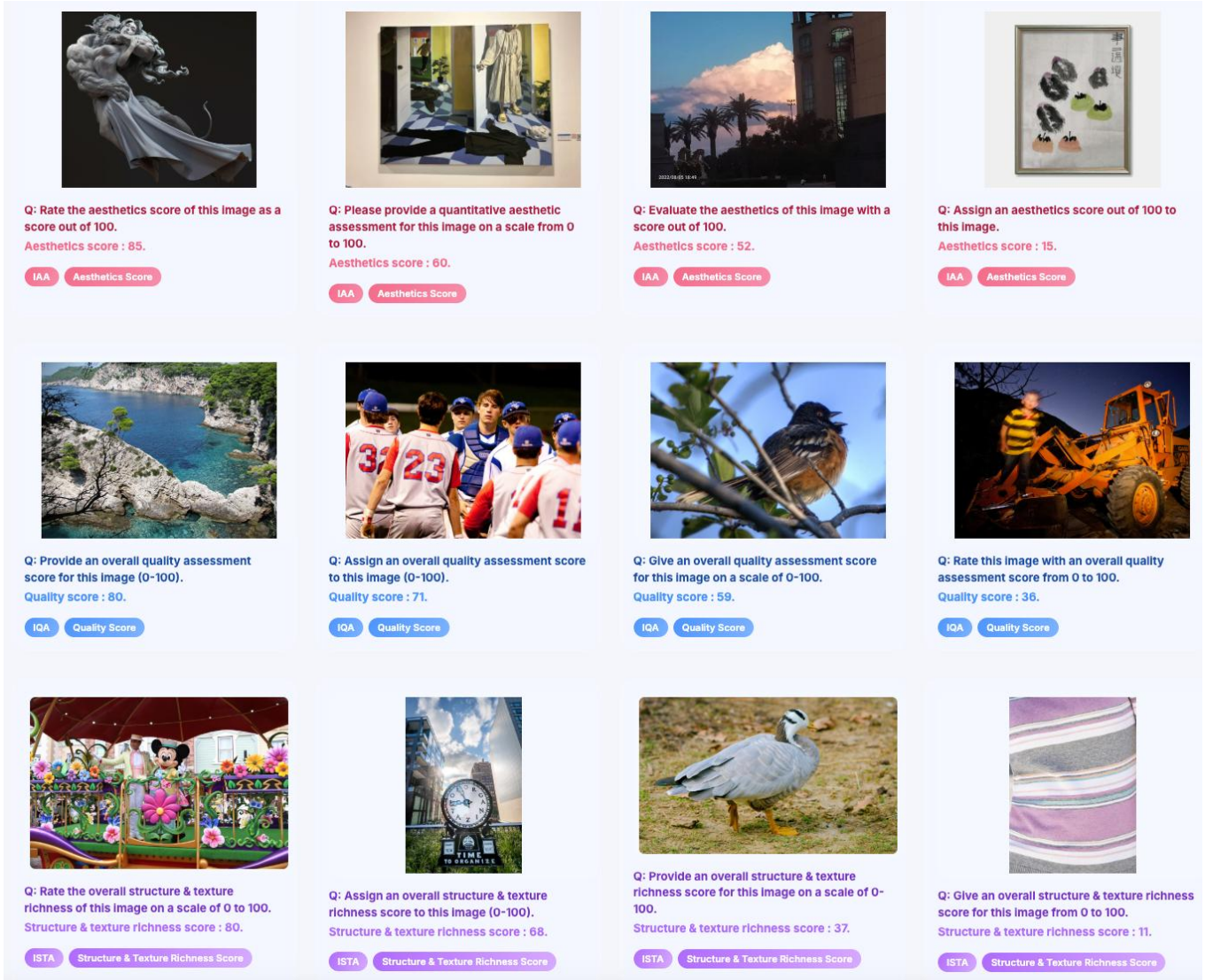
```
{
    "SceneType": "<SceneType>",
    "SceneName": "<SceneName>",
    "Components": [
        {
            "ComponentName": "<Component_1
                >",
            "DescriptionContent": {
                "PhysicalStructure": {
                    "BaseMorphology": ["<
                        Morphology_1>"],
                    "Arrangement": ["<
                        Arrangement_1>"]
                },
                "MaterialRepresentation": {
                    "MaterialClass": ["<
                        MaterialClass_1>"],
                    "SurfaceProperties":
                        ["<
                        SurfaceProperty_1
                        >"]
                },
                "GeometricComposition": {
                    "PlanarContour": ["<
                        PlanarContour_1>"],
                    "VolumetricForm": ["<
                        VolumetricForm_1>"]
                },
                "SemanticPerception": {
                    "FunctionalInference":
                        ["<
                        FunctionalInference_1
                        >"],
                    "StyleType": ["<
                        StyleType_1>"]
                }
            }
        },
        {
            "ComponentName": "<Component_2
                >",
                ...
            }
        }
    ]
}
```

UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture



UniPercept-Bench-VQA (Visual Question Answering)

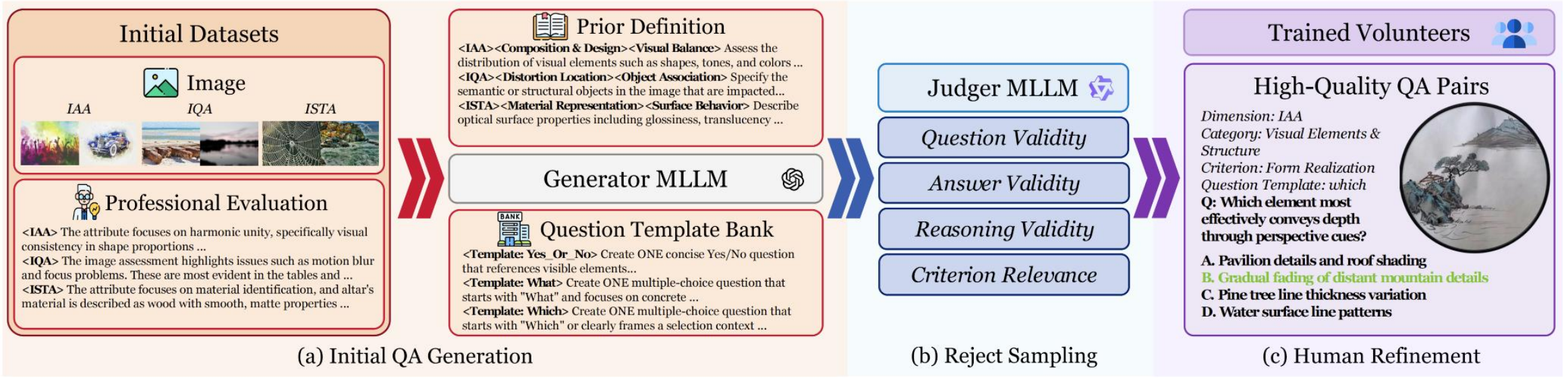
UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture



UniPercept-Bench-VR (Visual Rating)

UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

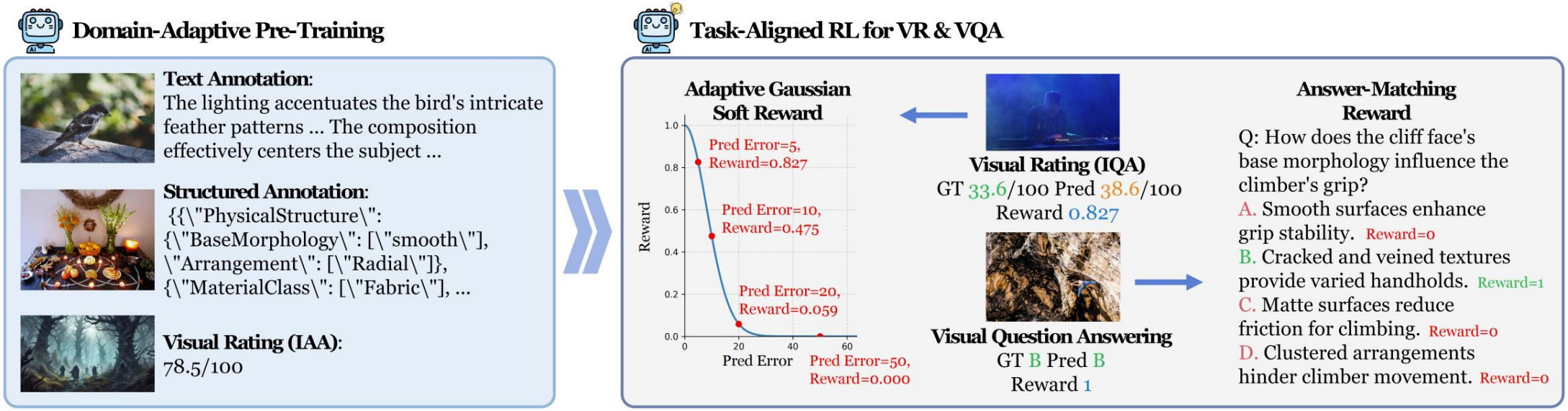
- Constuction pipeline of UniPercept-Bench



- MLLM performs poorly on professional captioning, while improve sinificantly with sufficient domain priors.
- Using heterogeneous MLLMs across different stages helps.
- Human refinement is essential.

UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- Training pipeline of UniPercept



- More diverse and large-scale pre-training data is better.
- RL brings better generalization. (For VR & VQA)

UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- Results & Comparison with SOTA models on UniPercept-Bench-VR (Visual Rating)

Models	VR - IAA					VR - IQA					VR - ISTA
	ArtiMuse-10K [4]	AVA [31]	TAD66K [11]	FLICKR-AES [32]	Avg.	KonIQ-10K [12]	SPAQ [9]	KADID [23]	PIPAL [10]	Avg.	ISTA-10K
Proprietary Models	*Large-scale Pretraining, Cross domain testing*										
GPT-4o [1]	0.333/0.276	0.509/0.485	0.278/0.282	0.605/0.597	0.431/0.410	0.695/0.744	0.874/0.881	0.677/0.646	0.325/0.349	0.643/0.655	-0.003/0.116
Llama-4-Scout [30]	0.204/0.147	0.345/0.329	0.236/0.210	0.548/0.506	0.333/0.298	0.503/0.653	-0.041/0.007	-0.099/-0.004	-0.007/0.023	0.089/0.170	-0.025/0.047
Gemini-2.5-pro [36]	0.187/0.035	0.248/0.100	0.143/0.037	0.357/0.206	0.234/0.095	0.582/0.316	0.087/0.212	0.436/0.274	0.225/-0.019	0.333/0.196	-0.230/-0.118
Claude-Sonnet-4.5 [24]	0.041/0.027	0.003/0.013	0.040/0.047	0.037/0.049	0.030/0.034	-0.037/-0.043	0.036/0.085	0.223/0.273	-0.131/-0.088	0.023/0.057	0.125/0.089
Claude-Sonnet-4.5-Think [24]	0.066/0.103	0.018/0.019	0.026/0.039	-/-	0.037/0.054	-/-	-/-	-/-	-/-	-/-	-/-
Open-Source Models	*Large-scale Pretraining, Cross domain testing*										
LLaVA-OneVision-1.5-Instruct-8B [2]	0.274/0.212	0.381/0.378	0.213/0.224	0.586/0.541	0.364/0.339	0.639/0.744	-/-	0.505/0.534	0.417/0.407	0.520/0.562	-0.094/0.027
GLM-4.5-V-106BA12B [37]	0.346/0.249	0.464/0.420	0.289/0.278	0.651/0.597	0.438/0.386	0.721/0.765	-0.040/-0.038	-0.142/-0.128	0.013/0.020	0.138/0.155	0.083/0.117
InternVL3-8B [50]	0.245/0.211	0.372/0.344	0.205/0.191	0.547/0.476	0.342/0.306	0.574/0.646	0.828/0.800	0.496/0.475	0.435/0.459	0.583/0.595	-0.127/0.046
InternVL3-78B [50]	0.223/0.206	0.385/0.344	0.221/0.220	0.518/0.433	0.337/0.301	0.635/0.676	0.849/0.852	0.579/0.553	0.415/0.457	0.619/0.634	-/-
InternVL3.5-8B [38]	0.135/0.104	0.308/0.295	0.180/0.182	0.519/0.448	0.286/0.257	0.663/0.660	0.783/0.777	0.541/0.478	0.351/0.386	0.585/0.575	-0.096/-0.025
InternVL3.5-38B [38]	0.219/0.175	0.359/0.357	0.201/0.208	0.559/0.529	0.334/0.317	0.578/0.652	0.840/0.831	0.568/0.537	0.448/0.457	0.608/0.619	0.262/0.345
QwenVL-2.5-Instruct-7B [3]	0.223/0.143	0.359/0.324	0.208/0.195	0.588/0.520	0.345/0.296	0.708/0.762	-/-	0.521/0.517	0.350/0.361	0.526/0.547	-0.046/0.076
QwenVL-2.5-Instruct-72B [3]	0.233/0.197	0.408/0.387	0.232/0.235	0.626/0.589	0.375/0.352	0.762/0.820	-/-	0.606/0.570	0.381/0.407	0.583/0.599	0.091/0.148
QwenVL-3-Instruct-8B [3]	0.156/0.094	0.280/0.170	0.191/0.121	0.507/0.388	0.283/0.193	0.761/0.822	0.612/0.604	0.723/0.696	0.434/0.427	0.633/0.637	0.033/0.044
QwenVL-3-Instruct-32B [3]	0.227/0.130	0.353/0.198	0.200/0.095	0.572/0.413	0.338/0.209	0.796/0.838	0.690/0.657	0.673/0.682	0.414/0.402	0.643/0.644	0.084/0.106
Specialized Models	*In domain*		*Cross domain*		*Avg.*	*In domain*		*Cross domain*		*Avg.*	*In domain*
ArtiMuse [4]	0.614/0.627	0.397/0.385	0.230/0.232	0.349/0.334	0.398/0.395	-/-	-/-	-/-	-/-	-/-	-/-
DeQA [44]	-/-	-/-	-/-	-/-	-/-	0.953/0.941	0.895/0.896	0.694/0.687	0.472/0.478	0.753/0.750	-/-
Q-Align* [41]	0.551/0.573	0.398/0.386	0.194/0.197	0.137/0.123	0.320/0.320	0.941/0.940	0.886/0.887	0.674/0.684	0.403/0.419	0.726/0.733	-/-
Q-Insight [20]	-/-	-/-	-/-	-/-	-/-	0.933/0.916	0.907/0.905	0.742/0.736	0.486/0.474	0.767/0.758	-/-
Q-Insight* [20]	0.228/0.175	0.405/0.376	0.212/0.217	0.617/0.537	0.366/0.326	0.733/0.750	0.800/0.938	0.580/0.548	0.369/0.368	0.621/0.651	0.060/0.152
UniPercept (Ours)	**0.746/0.738**	**0.589/0.577**	**0.336/0.346**	**0.688/0.681**	**0.590/0.586**	**0.940/0.949**	**0.904/0.895**	**0.872/0.870**	**0.581/0.594**	**0.824/0.827**	**0.778/0.767**

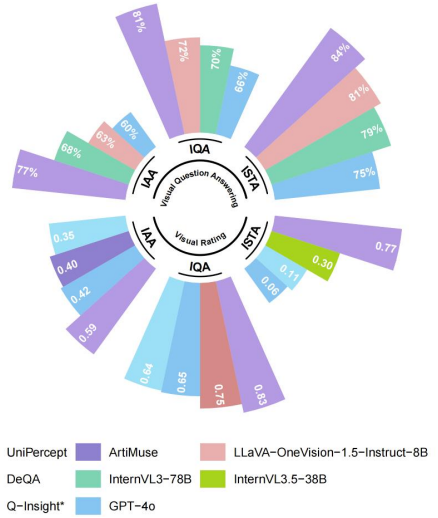
UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

➢ Results & Comparison with SOTA models on UniPercept-Bench-VQA (Visual Question Answering)

Models	IAA Categories								QA Templates						Overall
	Comp.	*VisStr.*	*Tech.*	*Creat.*	*Theme.*	*Emo.*	*Gest.*	*CompEv.*	*Lv.Pred*	*How*	*What*	*Which*	*Why*	*Yes-No*	
Random Guess	23.08%	27.27%	21.95%	29.63%	25.93%	22.86%	23.68%	32.56%	24.14%	21.28%	30.43%	25.32%	24.00%	29.49%	25.17%
Proprietary Models															
GPT-4o	64.62%	59.57%	57.58%	60.19%	65.19%	67.62%	51.95%	30.23%	38.86%	78.17%	72.46%	62.66%	72.67%	70.51%	60.04%
Llama-4-Scout	62.56%	68.45%	59.76%	61.11%	57.78%	70.48%	48.68%	32.56%	43.97%	70.92%	69.57%	61.39%	77.33%	70.51%	60.91%
Gemini-2.5-pro	71.79%	68.45%	61.59%	76.85%	67.41%	63.81%	61.84%	37.21%	45.98%	78.72%	73.91%	67.72%	84.67%	84.62%	66.44%
Claude-Sonnet-4.5	70.26%	70.05%	62.20%	71.30%	64.44%	67.62%	50.00%	46.51%	46.84%	77.30%	76.09%	65.19%	86.00%	69.23%	65.45%
Claude-Sonnet-4.5-Think	71.28%	69.52%	61.21%	68.52%	62.22%	66.67%	53.25%	41.86%	44.57%	75.89%	77.54%	67.09%	86.00%	66.67%	64.73%
Open-Source Models															
LLaVA-OneVision-1.5-Instruct-8B	67.18%	68.62%	61.21%	62.96%	67.41%	62.86%	53.25%	20.93%	34.86%	85.21%	79.71%	65.82%	83.33%	69.23%	62.60%
GLM-4.5-V-106BA12B	67.18%	65.78%	60.98%	75.00%	64.44%	68.57%	51.32%	46.51%	45.40%	71.63%	78.26%	65.82%	84.67%	70.51%	64.46%
InternVL3-8B	65.64%	67.55%	59.39%	67.59%	69.63%	62.86%	50.65%	25.58%	36.00%	81.69%	73.91%	67.72%	86.00%	71.79%	62.60%
InternVL3-78B	71.79%	73.26%	61.21%	73.15%	74.81%	74.29%	53.25%	37.21%	45.14%	85.82%	81.16%	72.15%	86.00%	75.64%	68.28%
InternVL3.5-8B	32.31%	29.41%	30.30%	26.85%	28.89%	26.67%	23.38%	9.30%	17.14%	41.13%	26.81%	19.62%	36.00%	58.97%	28.18%
InternVL3.5-38B	37.44%	40.11%	27.88%	39.81%	34.81%	38.10%	45.45%	6.98%	34.00%	47.52%	26.09%	28.48%	37.33%	50.00%	35.67%
QwenVL-2.5-Instruct-7B	67.18%	70.74%	56.36%	66.67%	68.89%	63.81%	48.05%	37.21%	38.86%	76.76%	75.36%	67.09%	87.33%	71.79%	63.19%
QwenVL-2.5-Instruct-72B	22.05%	24.60%	25.45%	29.63%	30.37%	18.10%	19.48%	6.98%	14.00%	19.86%	17.39%	24.05%	41.33%	51.28%	23.74%
QwenVL-3-Instruct-8B	31.28%	32.09%	32.12%	37.04%	34.07%	22.86%	37.66%	25.58%	35.43%	14.89%	17.39%	34.81%	28.67%	73.08%	31.92%
QwenVL-3-Instruct-32B	23.08%	26.74%	32.12%	26.85%	32.59%	20.95%	33.77%	20.93%	33.43%	9.22%	13.77%	31.01%	18.67%	66.67%	27.39%
Specialized Models															
ArtiMuse	67.69%	68.45%	64.85%	74.07%	71.85%	64.76%	61.04%	32.56%	39.14%	88.65%	76.81%	72.78%	85.33%	79.49%	66.31%
UniPercept (Ours)	**80.00%**	**77.54%**	**69.70%**	**80.56%**	**79.26%**	**80.95%**	**67.53%**	**69.77%**	**63.71%**	**92.20%**	**81.88%**	**75.32%**	**86.67%**	**84.62%**	**76.55%**

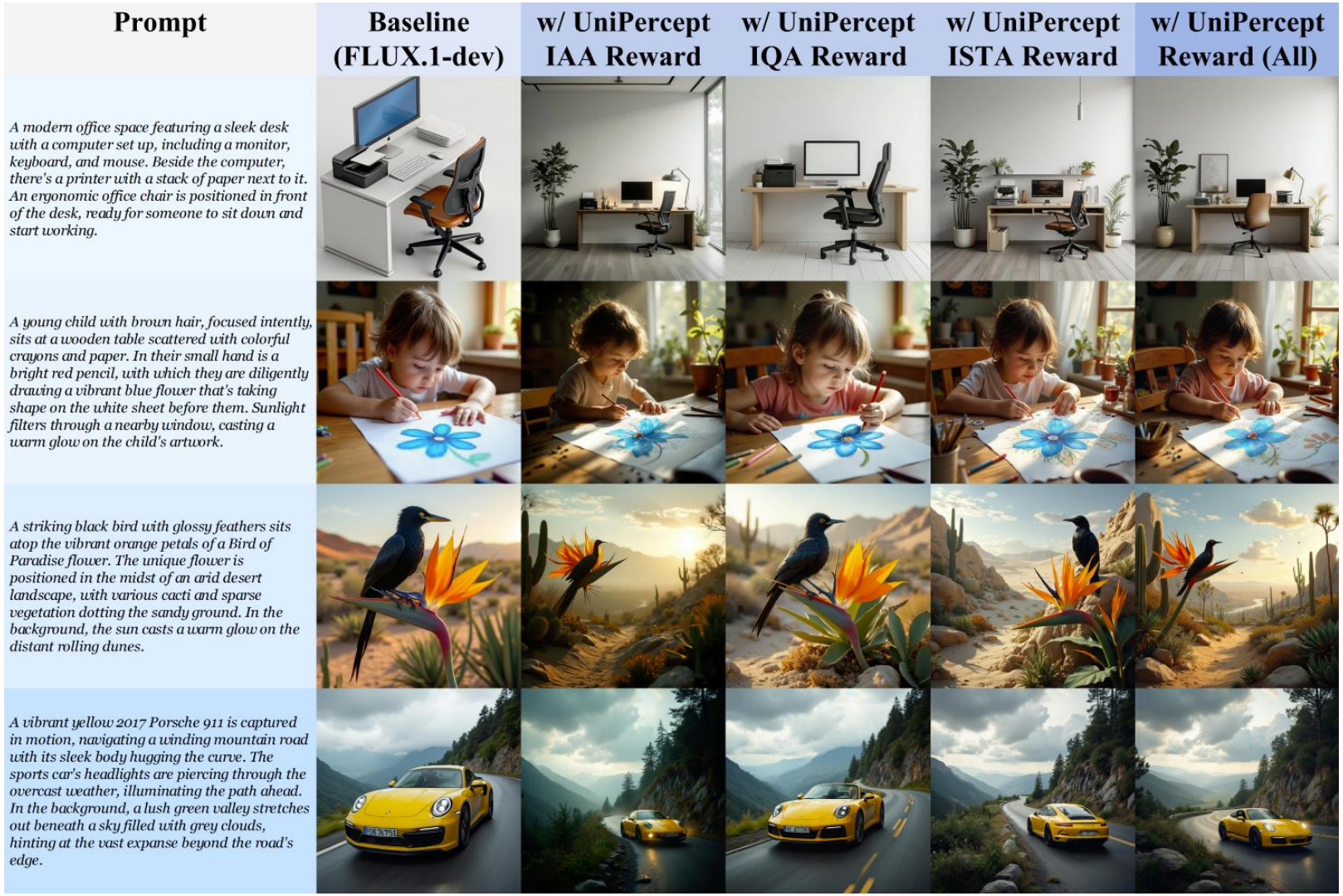
Models	IQA Categories			QA Templates						Overall
	Loc.	*Sev.*	*Type.*	*Lv.Pred*	*How*	*What*	*Which*	*Why*	*Yes-No*	
Random Guess	23.67%	24.75%	20.08%	24.75%	27.03%	16.05%	25.00%	21.39%	22.99%	23.16%
Proprietary Models										
GPT-4o	71.74%	53.18%	70.49%	53.18%	83.78%	59.26%	61.31%	80.21%	67.82%	66.36%
Llama-4-Scout	60.18%	58.19%	52.05%	58.19%	82.16%	37.04%	38.69%	66.31%	62.07%	57.81%
Gemini-2.5-pro	32.84%	52.84%	40.98%	52.84%	40.54%	32.72%	29.17%	41.18%	28.74%	40.17%
Claude-Sonnet-4.5	71.19%	51.51%	66.80%	51.51%	90.81%	50.00%	50.60%	82.89%	71.26%	64.80%
Claude-Sonnet-4.5-Think	71.19%	55.52%	66.80%	55.52%	89.19%	50.00%	51.79%	82.89%	72.41%	65.90%
Open-Source Models										
LLaVA-OneVision-1.5-Instruct-8B	76.51%	59.87%	77.46%	59.87%	91.35%	70.37%	61.31%	82.35%	75.86%	72.15%
GLM-4.5-V-106BA12B	70.09%	35.79%	54.51%	35.79%	88.11%	48.77%	44.05%	74.33%	68.97%	57.17%
InternVL3-8B	71.56%	52.84%	59.43%	52.84%	87.03%	59.88%	48.81%	71.12%	71.26%	63.69%
InternVL3-78B	75.41%	51.84%	81.56%	51.84%	93.51%	66.67%	63.10%	88.24%	66.67%	70.31%
InternVL3.5-8B	38.17%	44.82%	38.11%	44.82%	35.14%	41.98%	30.36%	36.36%	56.32%	39.98%
InternVL3.5-38B	38.90%	49.83%	45.08%	49.83%	46.49%	41.36%	31.55%	33.16%	62.07%	43.29%
QwenVL-2.5-Instruct-7B	74.13%	48.83%	66.39%	48.83%	88.65%	60.49%	53.57%	78.61%	77.01%	65.44%
QwenVL-2.5-Instruct-72B	31.01%	4.68%	16.39%	4.68%	35.14%	14.81%	11.31%	22.99%	66.67%	20.50%
QwenVL-3-Instruct-8B	34.68%	55.18%	16.39%	55.18%	20.54%	18.52%	27.38%	25.67%	77.01%	36.21%
QwenVL-3-Instruct-32B	29.54%	14.38%	16.80%	14.38%	11.89%	18.52%	25.60%	22.46%	74.71%	22.52%
UniPercept (Ours)	**77.43%**	**79.60%**	**90.98%**	**79.60%**	**87.03%**	**80.86%**	**75.60%**	**83.42%**	**79.31%**	**81.07%**

Models	ISTA Categories					QA Templates					Overall
	Scene.	*Phys.*	*Mat.*	*Geo.*	*Sem.*	*How*	*What*	*Which*	*Why*	*Yes-No*	
Random Guess	26.50%	23.63%	24.73%	30.30%	30.58%	26.28%	23.84%	24.29%	33.77%	33.33%	26.60%
Proprietary Models											
GPT-4o	75.64%	79.12%	73.48%	33.33%	77.27%	71.79%	78.78%	69.23%	77.92%	72.46%	74.64%
Llama-4-Scout	73.50%	75.27%	71.68%	72.73%	67.77%	75.64%	69.77%	69.64%	77.27%	69.57%	71.86%
Gemini-2.5-pro	76.50%	82.42%	77.06%	66.67%	77.69%	78.21%	78.20%	75.71%	82.47%	71.01%	77.73%
Claude-Sonnet-4.5	76.92%	78.57%	74.91%	90.91%	77.69%	76.92%	77.03%	74.49%	81.82%	79.71%	77.32%
Claude-Sonnet-4.5-Think	77.35%	78.02%	73.12%	87.88%	75.21%	76.28%	74.71%	74.09%	81.82%	76.81%	76.08%
Open-Source Models											
LLaVA-OneVision-1.5-Instruct-8B	78.63%	85.16%	82.44%	72.73%	80.17%	83.33%	81.40%	75.30%	84.42%	88.41%	81.13%
GLM-4.5-V-106BA12B	81.20%	79.67%	74.55%	72.73%	75.21%	80.77%	76.74%	73.68%	79.87%	78.26%	77.22%
InternVL3-8B	75.64%	79.12%	73.48%	33.33%	77.27%	71.79%	78.78%	69.23%	77.92%	72.46%	74.64%
InternVL3-78B	79.06%	85.16%	77.42%	69.70%	78.51%	81.41%	79.65%	73.68%	84.42%	81.16%	79.28%
InternVL3.5-8B	54.27%	50.55%	58.42%	39.39%	36.36%	46.79%	56.69%	48.58%	29.87%	71.01%	49.79%
InternVL3.5-38B	50.00%	55.49%	61.29%	30.30%	35.95%	50.64%	59.30%	42.91%	37.01%	57.97%	50.10%
QwenVL-2.5-Instruct-7B	74.79%	72.53%	74.91%	51.52%	73.55%	73.72%	77.33%	66.80%	74.03%	73.91%	73.30%
QwenVL-2.5-Instruct-72B	14.10%	29.12%	19.71%	12.12%	18.60%	20.51%	12.21%	14.57%	31.17%	46.38%	19.59%
QwenVL-3-Instruct-8B	27.78%	32.42%	25.45%	39.39%	24.79%	14.74%	23.26%	28.34%	25.32%	81.16%	27.63%
QwenVL-3-Instruct-32B	26.50%	24.73%	19.00%	15.15%	18.60%	11.54%	18.31%	22.67%	17.53%	66.67%	21.65%
UniPercept (Ours)	**89.74%**	**85.71%**	**82.44%**	**93.94%**	**78.51%**	**82.69%**	**89.24%**	**78.54%**	**83.12%**	**85.51%**	**84.23%**



UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- UniPercept can serve as a plug-and-play reward model for text-to-image generation.



UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- UniPercept can provide unified perceptual metrics for evaluation.

Models	Preference Score		Image Quality	Image Aesthetics		UniPercept Score		
	PickScore (Kirstain et al., 2023)	HPSv3 (Ma et al., 2025)	DeQA (You et al., 2025a)	LAION-Aes (Schuhmann & Beaumont, 2022)	ArtiMuse (Cao et al., 2025)	IAA	IQA	ISTA
Baseline	22.46	10.71	4.32	5.77	59.02	65.18	73.59	46.64
w/ UniPercept IAA Reward	22.47	10.09	4.09	6.19	67.02	76.20	76.39	54.83
w/ UniPercept IQA Reward	22.63	11.21	4.37	6.02	63.64	72.16	76.87	52.34
w/ UniPercept ISTA Reward	22.72	11.09	4.37	6.16	63.75	72.23	76.17	59.61
w/ UniPercept Reward (All)	**22.67**	**10.93**	**4.33**	**6.19**	**65.52**	**74.24**	**77.04**	**59.08**

Table 19. **Evaluation of T2I Models on DPG (Hu et al., 2024) Metrics and UniPercept Metrics.**

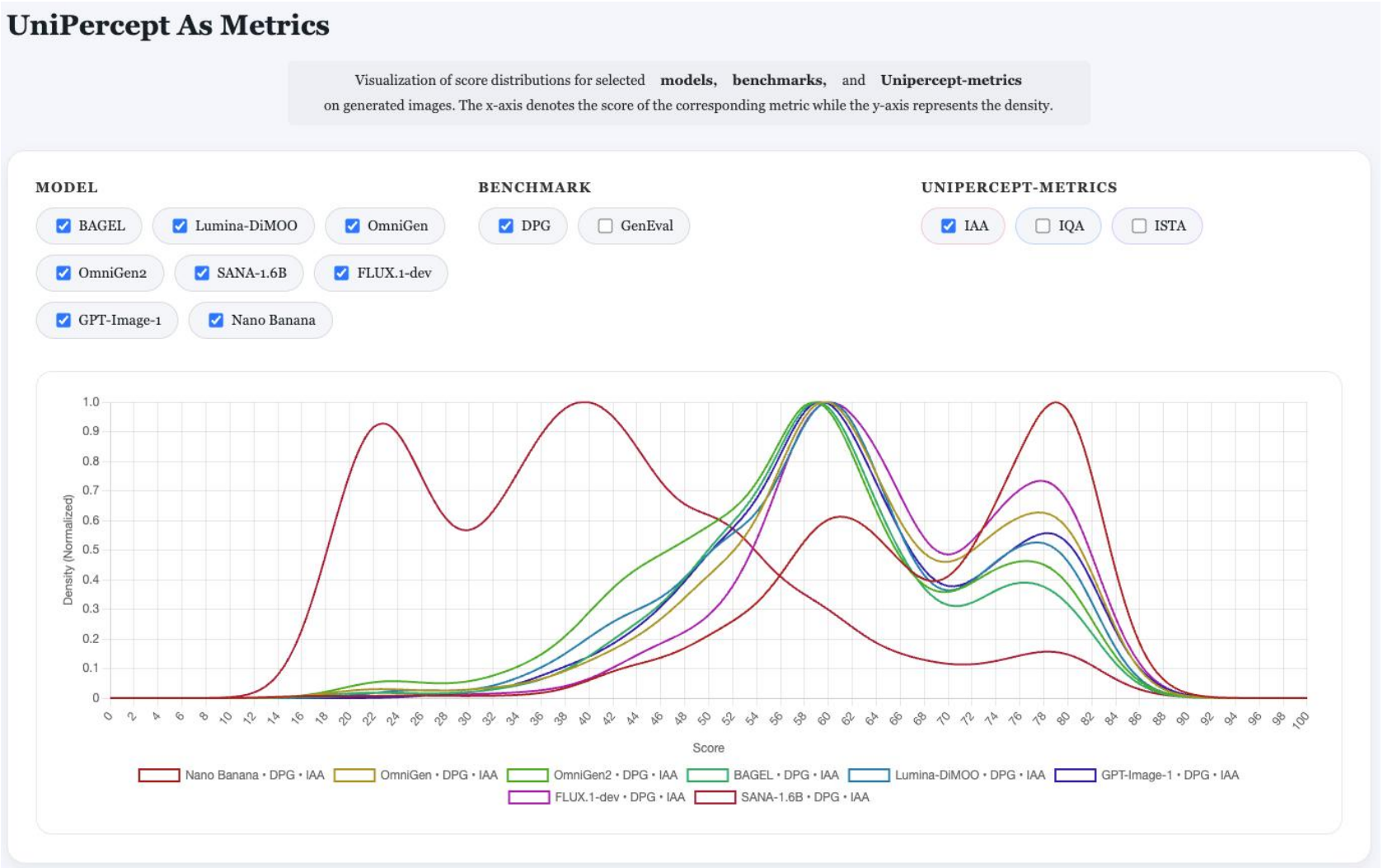
Models	DPG Metrics (Hu et al., 2024)						UniPercept Metrics			
	Global	Entity	Attribute	Relation	Other	Overall	IAA	IQA	ISTA	Avg.
OmniGen (Xiao et al., 2024)	–	–	–	–	–	–	62.83	72.22	45.09	60.04
OmniGen2 (Wu et al., 2025b)	88.81	88.83	90.18	89.37	90.27	83.57	58.51	71.89	43.31	57.90
BAGEL (Deng et al., 2025)	88.94	90.37	91.29	90.82	88.67	85.07	60.20	70.52	45.78	58.83
SANA-1.6B (Xie et al., 2024; 2025)	86.00	91.50	88.90	91.90	90.70	84.80	40.33	42.89	42.41	41.87
Lumina-DiMOO (Xin et al., 2025)	81.46	92.08	88.98	94.31	82.00	86.04	61.00	71.14	44.83	58.99
FLUX.1-dev (Labs et al., 2025)	74.35	90.00	88.96	90.87	88.33	83.84	65.18	73.59	46.64	61.80
GPT-Image-1 (Achiam et al., 2023)	88.89	88.94	89.84	92.63	90.96	85.15	62.27	72.87	44.88	60.00
Qwen-Image (Wu et al., 2025a)	91.32	91.56	92.02	94.31	92.73	88.32	62.89	72.15	47.23	60.76

Table 20. **Evaluation of T2I Models on GenEval (Ghosh et al., 2023) Metrics and UniPercept Metrics.**

Models	GenEval (Ghosh et al., 2023) Metrics							UniPercept Metrics			
	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Bind.	Overall	IAA	IQA	ISTA	Avg.
OmniGen (Xiao et al., 2024)	0.99	0.86	0.64	0.85	0.31	0.55	0.70	58.84	75.62	41.00	58.49
OmniGen2 (Wu et al., 2025b)	0.99	0.96	0.74	0.98	0.71	0.75	0.86	54.20	75.16	34.48	54.61
BAGEL (Deng et al., 2025)	0.99	0.94	0.81	0.88	0.64	0.63	0.82	58.68	71.24	38.35	56.09
SANA-1.6B (Xie et al., 2024; 2025)	0.99	0.77	0.62	0.88	0.21	0.47	0.66	34.34	35.11	31.22	33.56
Lumina-DiMOO (Xin et al., 2025)	1.00	0.94	0.85	0.89	0.85	0.76	0.88	51.93	71.98	30.86	51.59
FLUX.1-dev (Labs et al., 2025)	0.98	0.81	0.74	0.79	0.22	0.45	0.66	64.24	74.96	41.14	60.11
GPT-Image-1 (Achiam et al., 2023)	0.99	0.92	0.85	0.92	0.75	0.61	0.84	69.07	76.74	51.26	65.69
Qwen-Image (Wu et al., 2025a)	0.99	0.92	0.89	0.88	0.76	0.77	0.87	52.02	74.44	34.13	53.53

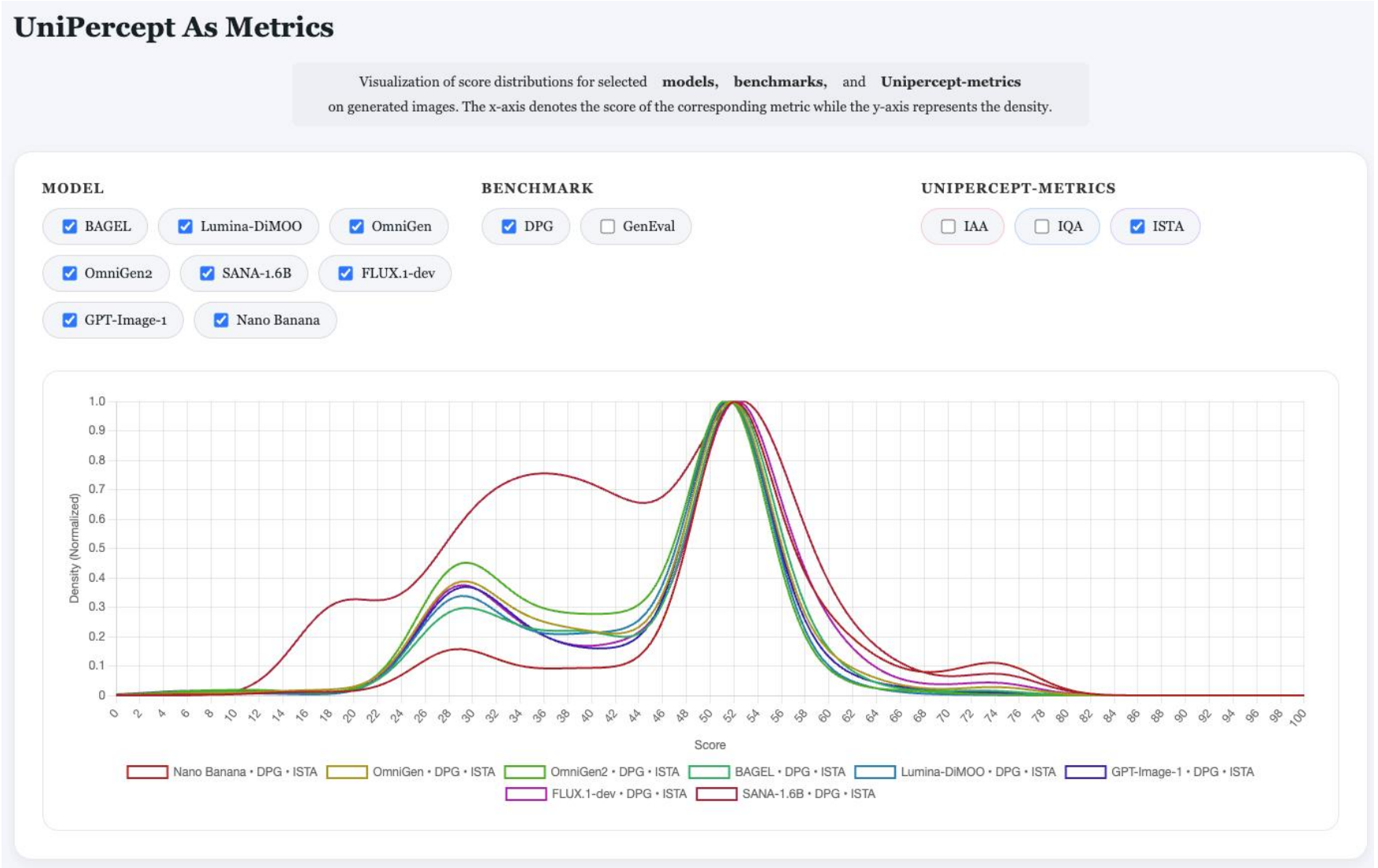
UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- UniPercept can provide unified perceptual metrics for evaluation.



UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- UniPercept can provide unified perceptual metrics for evaluation.



UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- UniPercept can provide unified perceptual metrics for evaluation.

Table 18. **UniPercept Metrics on Various Datasets.**

Dataset	UniPercept-IAA	UniPercept-IQA	UniPercept-ISTA	Avg.
Natural Images				
ImageNet (Russakovsky et al., 2015)	53.88	61.90	36.79	50.85
Unsplash (Unsplash, 2025)	62.49	69.19	43.32	58.33
DF2K (Timofte et al., 2017; 2018a;b; Ignatov et al., 2019)	45.99	52.92	34.78	44.56
LAION-5B (Schuhmann et al., 2022)	60.56	69.21	38.85	56.21
AIGC Images				
Blip3o-60K (Chen et al., 2025)	63.81	73.88	49.38	62.36
ImgEdit (Ye et al., 2025)	55.83	59.77	36.88	50.83

UniPercept: Towards Unified Perceptual-Level Image Understanding across Aesthetics, Quality, Structure, and Texture

- UniPercept constructs a comprehensive profile for each image, spanning the IAA, IQA, and ISTA domains and providing fine-grained, multi-dimensional outputs.

UniPercept-Constructed Image Profile

IAA Rating: 29/100

IQA Rating: 35/100

ISTA Rating: 46/100

Image Aesthetics Assessment (IAA)

Compostion & Design
- *The composition is nonexistent, with a completely blurred and indistinct subject ...*
- *There is no balance or rhythm to speak of. The elements are randomly scattered ...*

Technical Execution
- *The image exhibits catastrophic technical execution.*
- *Every possible mistake is present, with extreme blurriness rendering all details invisible.*
- *The focus is nonexistent, resulting in a complete lack of sharpnessness.*

...

Comprehensive Evaluation
- *The image suffers from extreme blurriness, rendering any aesthetic evaluation impossible.*
- *The lack of focus eliminates any potential visual impact, technical merit, or depth. Every element, from the pergola to the lanterns, is indistinct, failing to convey any meaningful ...*
- *In summary, the image's total absence of visual coherence results in a zero aesthetic merit.*

Image Quality Assessment (IQA)

Distortion Location
- *The entire scene, including the pergola, lanterns, and background foliage, is out of focus, resulting in a complete loss of sharpnessness. This blurring affects every element uniformly, making it impossible to identify ...*
- *The image fails to convey any meaningful content due to its total lack of clarity.*

Distortion Severity
- *The image suffers from extreme blurriness.*
- *This catastrophic blurring negates any possible evaluation of other distortion types, as nothing is visible.*
- *The final distortion severity is the maximum imaginable.*

...

Overall Quality
- *The image is severely out of focus, rendering all elements indistinct and lacking any discernible detail.*
- *The lack of sharpnessness, exposure, and any visible content results in a complete failure in image quality. This extreme blurriness negates any possible merit, making it impossible to evaluate any positive attributes.*
- *In summary, the image quality is catastrophic, with zero merit due to total blurriness.*

Image Structure & Texture Assessment (ISTA)

Scene Decomposition Principles
- **Scene Type**: *Blurred Composite Scene*
- **Scene Name**: *Garden with pergola and hanging lights*

Componet: Pergola
- **Physical Structure**: *[Base Morphology: Lined] [Arrangement: Horizontal]*
- **Material Representation**: *[Material Class: Glass] [Surface Properties: Glossy]*
- **Geometric Composition**: *[Volumetric Form: Sphere]*

Componet: Background Foliage
- **Physical Structure**: *[Base Morphology: Matted] [Arrangement: Layered]*
- **Material Representation**: *[Material Class: Foliage] [Surface Properties: Matte]*
- **Geometric Composition**: *[Volumetric Form: Naturalistic]*

Componet: Hanging Lights
- **Physical Structure**: *[Base Morphology: Smooth] [Arrangement: Clustered]*
- **Material Representation**: *[Material Class: Glass] [Surface Properties: Glossy]*
- **Geometric Composition**: *[Volumetric Form: Sphere]*

...

UniPercept-Constructed Image Profile

IAA Rating: 87/100

IQA Rating: 75/100

ISTA Rating: 56/100

Image Aesthetics Assessment (IAA)

Originality & Creativity
- *The image depicts a well-preserved historical house with a straightforward and classic presentation.*
- *While the composition and clarity are excellent, the photograph adheres closely to traditional documentation styles, lacking experimental or imaginative elements.*

Emotion & Viewer Response
- *The image evokes a strong sense of nostalgia and admiration for historical architecture, effectively capturing ...*
- *The well-preserved colonial house and lush green create a tranquil, inviting atmosphere, likely sparking curiosity and appreciation for historical preservation. While the composition is perfect and visually pleasing, it may ...*

...

Comprehensive Evaluation
- *The image showcases a well-preserved historical house with strong visual impact due to its symmetrical ...*
- *Overall, the image excels in historical representation but slightly lacks dynamic engagement. The aesthetic merit is strong, reflecting both historical and visual excellence.*

Image Quality Assessment (IQA)

Distortion Type / Location / Severity
- *There is no distortion in the image.*

...

Overall Quality
- *The image showcases a well-preserved historical building with excellent clarity and vibrant colors, capturing the architectural details and surrounding green effectively.*
- *The composition is balanced, with well-lit, sharp focus, and excellent exposure.*
- *The cobblestone path and lush trees enhance the scene's authenticity.*
- *There are no visible flaws, indicating perfect sharpnessness and color accuracy.*
- *This image excels in every aspect, achieving perfection.*
- *Overall, this is a flawless, high-quality photograph.*

Image Structure & Texture Assessment (ISTA)

Scene Decomposition Principles
- **Scene Type**: *Composite Scene*
- **Scene Name**: *Historic house with surrounding environment*

Componet: House
- **Physical Structure**: *[Base Morphology: Lined, Smooth] [Arrangement: Horizontal]*
- **Material Representation**: *[Material Class: Wood] [Surface Properties: Matte]*
- **Geometric Composition**: *[Planar Contour: Rectangle] [Volumetric Form: Cuboid]*
- **Semantic Perception**: *[Style Type: Colonial Architecture]*

Componet: Roof
- **Physical Structure**: *[Base Morphology: Tiled] [Arrangement: Layered]*
- **Material Representation**: *[Material Class: Tile] [Surface Properties: Matte]*
- **Geometric Composition**: *[Planar Contour: Rectangle]*
- **Semantic Perception**: *[Style Type: Traditional]*

Componet: Chimneys
- **Physical Structure**: *[Base Morphology: Brick] [Arrangement: Vertical]*
- **Material Representation**: *[Material Class: Brick] [Surface Properties: Matte]*
- **Geometric Composition**: *[Planar Contour: Rectangle] [Volumetric Form: Cuboid]*
- **Semantic Perception**: *[Style Type: Traditional]*

...

UniPercept-Constructed Image Profile

IAA Rating: 75/100

IQA Rating: 81/100

ISTA Rating: 82/100

Image Aesthetics Assessment (IAA)

Visual Elements & Structure
- *The image excels in visual elements and structure, showcasing a harmonious blend of vibrant colors ...*
- *The lush green of the jungle contrasts beautifully with the bright orange background, enhancing visual impact. The composition is meticulously balanced, with dynamic focal points like the pyramid and characters, creating a cohesive and engaging scene.*

Theme & Communication
- *The image masterfully communicates the theme of a jungle tomb raid through its vivid and intricate LEGO ...*
- *Every element, from the pyramid to the characters and animals, is meticulously crafted to evoke a sense of adventure and danger. The narrative is clear and engaging, effectively capturing the essence of Indiana Jones.*

Overall Gestalt
- *The image showcases a masterpiece of LEGO artistry, seamlessly blending every element to perfection.*
- *The overall gestalt is nothing short of perfection, with every detail meticulously crafted to create a cohesive, enchanting jungle scene. The vibrant colors, intricate details, and harmonious composition evoke a sense ...*

...

Image Quality Assessment (IQA)

Distortion Type / Location / Severity
- *There is no distortion in the image.*

...

Overall Quality
- *The image is sharp and well-composed, showcasing a detailed LEGO jungle scene without any distortion.*
- *The vibrant colors and intricate details are intact, with no signs of blurring, overexposure, or compression artifacts. The scene's perfection is evident in its clarity and balance. There are no visual flaws.*
- *The lighting and composition are flawless, enhancing the intricate details and depth.*
- *There are no visible flaws, making it a flawless representation.*
- *This image excels beyond perfection, meriting a perfect score. In summary, the image is flawless in every aspect, showcasing impeccable quality.*

Image Structure & Texture Assessment (ISTA)

Scene Decomposition Principles
- **Scene Type**: *Composite Scene*
- **Scene Name**: *Lego Jungle Temple*

Componet: Temple Structure
- **Physical Structure**: *[Base Morphology: Blocky, Grid] [Arrangement: Layered]*
- **Material Representation**: *[Material Class: Plastic] [Surface Properties: Matte]*
- **Geometric Composition**: *[Planar Contour: Rectangle] [Volumetric Form: Cuboid]*

Componet: Palm Trees
- **Physical Structure**: *[Base Morphology: Fibrous, Frilly] [Arrangement: Vertical]*
- **Material Representation**: *[Material Class: Plastic] [Surface Properties: Matte]*
- **Geometric Composition**: *[Volumetric Form: Cylinder]*

Componet: Flora
- **Physical Structure**: *[Base Morphology: Matted, Frilly] [Arrangement: Clustered]*
- **Material Representation**: *[Material Class: Plastic] [Surface Properties: Matte]*

Componet: Figures
- **Physical Structure**: *[Base Morphology: Blocky] [Arrangement: Clustered]*
- **Material Representation**: *[Material Class: Plastic] [Surface Properties: Matte]*

...