Stanford University

NEURAL INFORMATION PROCESSING SYSTEMS

SALART-VQA: Diagnosing Whether VLMs Understand Salient Artifacts in Generated Images

Ruotian Zhang*, Xiaoxiao Sun*, Junzhe Huang, James Burgess, Serena Yeung-Levy

Salient Artifacts in Generated Images

Salient artifact: human-visible defect that is visually prominent, localizable, and describable in natural language; the benchmark focuses on such target defect



Imagen 4 [Google DeepMind, 2025],
Midjourney [Midjourney, 2026],
FLUX.2 [klein] [Black Forest Labs, 2026b]



GPT Image 2.5

Can VLMs Understand Salient Artifacts in Generated Images?

(a) Image and question example

Does this image contain a salient artifact: a localized, visually identifiable structural defect? If yes, describe where it is and what is wrong.

Yes. The man's left hand has an extra finger… ✓



Gemini 3.1 Pro

Yes. The most salient artifacts are located on the left side of the image (the man's right hand): Mangled Hand Anatomy: The hand holding the stem of the banana lacks coherent physical structure ...

Claude 4.7 Opus

Yes, there's a clear artifact in the hands holding the banana, particularly on the right side of the image (the man's left hand). The fingers there look anatomically off — the fingertips appear oddly shaped …

GPT 5.5

Yes. There is a localized artifact at the left end of the banana: the peel/stem area looks structurally malformed, with a small jagged protruding piece that does not look naturally connected to the banana ...

All three say "Yes" — two point to the wrong region, none names the true defect.

Image-level accuracy can't tell whether a claim is grounded in visual evidence.

Ruotian Zhang*, Xiaoxiao Sun*, Junzhe Huang, James Burgess, Serena Yeung-Levy

SALART-VQA: A Hierarchy of Artifact Understanding Tasks



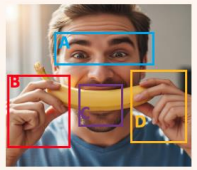
Q1: Is there is a clearly visible artifact in this AI generated image?
A: yes
B: no

Perception

Q2: If there is a clearly visible artifact, which **region** best contains it?
A: the man's eyes
B: the man's right hand
C: the banana
D: the man's left hand
E: none

semantic grounding

Q3: If there is a clearly visible artifact, which **box best localizes** it?



A: box A
B: box B
C: box C
D: box D
E: None

spatial grounding

Q4: If there is a clearly visible artifact, which option **best describes** it?
A: the man's pupils are misshapen
B: right hand … unnatural …
C: … banana…unnaturally smooth
D: extra fingers on his left hand…
E: none of these descriptions

evidence reasoning

Why a hierarchy?

A single yes/no hides where a claim breaks. Asking each step separately shows which one fails — e.g. 1101: only the box is wrong.

Why aligned options?

Every option A–D, not just the answer, refers to the same region in Q2, Q3 and Q4. If B were a hand in Q2 but a shadow in Q4, the results couldn't be compared.

How are the options designed?

Plausible distractors make the options hard; the closed set keeps secondary flaws out, so the answer is never ambiguous.

How the benchmark is built

DATA PIPELINE

Generate ~13k photorealistic images
Imagen 4, Midjourney, FLUX.2 · COCO-caption prompts that never ask for defects

Selecting **one** salient artifact

Selecting **no** salient artifact



Art-DG · 356
one clearly most-obvious artifact per image

Ref-PG · 119
clean generated scenes



Pair: retrieve a matched real photo (caption / CLIP)

Pair: inpaint one body part (MediaPipe → SAM 2 → FLUX.1-Fill)



Ref-Real · 356
real COCO photo, similar content

Art-PG · 119
same scene + one inserted artifact



Art-DG ↔ Ref-Real: catch false alarms
Does the model flag “artifacts” in real photos with similar content?

Art-PG ↔ Ref-PG: catch hallucination
Same scene, same options — Ref-PG is the original before we inserted the defect.
If the model still picks the defect option, it is guessing, not seeing.

Overall evaluation results

Question / Data type / Model	Artifact images: Artifact: Q1 detection recall	Artifact: All$_{Q1-4}$ complete accuracy	Artifact: All$_{Q2-4}$ grounding accuracy	Reference images: Ref-Real: Q1 no artifact specificity	Ref-Real: All$_{Q1-4}$ complete accuracy	Ref-Real: All$_{Q2-4}$ grounding accuracy	Ref-PG: All$_{Q2-4}$ grounding accuracy	Overall: All splits: Q1 detection accuracy	Overall: All splits: All$_{Q1-4}$ complete accuracy	Overall: All splits: All$_{Q2-4}$ grounding accuracy
Random	50.00	0.40	0.80	50.00	0.40	0.80	0.80	50.00	0.40	0.80
Human	100.00	100.00	100.00	99.53	99.53	99.53	99.16	99.80	99.80	99.72
Closed-source										
Gemini 3.1 Pro	**99.37**	**53.26**	**53.89**	50.28	7.87	8.71	0.00	78.34	33.81	30.21
Gemini 3.1 Flash Lite	83.37	32.63	37.47	93.26	16.01	16.01	0.00	**87.61**	25.51	24.73
Claude 4.6 Opus	47.58	11.79	17.47	**100.00**	96.35	96.35	0.84	70.04	**48.01**	44.95
GPT-5.4	4.63	1.89	6.32	**100.00**	95.79	95.79	20.17	45.49	42.12	41.58
Claude 4.5 Haiku	0.42	0.00	0.84	99.44	49.16	49.16	7.56	42.84	21.06	19.79
GPT-5.4-nano	0.21	0.00	0.42	99.72	78.65	78.93	19.33	42.84	33.69	32.21
Open-weight (Thinking / Hybrid-thinking by default)										
Gemma-4-31B-it	86.32	33.26	37.68	68.26	3.93	3.93	0.00	78.58	20.70	20.31
Gemma-4-26B-it-A4B	57.89	17.68	28.00	83.71	14.61	15.17	0.00	68.95	16.37	19.68
Qwen3.5-397B-A17B	33.26	12.63	27.79	96.63	39.33	40.17	0.00	60.41	24.07	28.95
Qwen3.5-35B-A3B	20.63	6.32	18.11	96.35	39.61	40.17	0.84	53.07	20.58	24.21
Kimi K2.5	3.79	1.68	4.84	**100.00**	77.25	77.25	28.57	45.01	34.06	34.95
Qwen3-VL-30B-A3B-T	22.53	1.47	3.16	89.89	56.74	59.55	10.92	51.39	25.15	25.26
Qwen3-VL-235B-A22B-T	7.79	1.26	5.05	96.91	35.67	35.96	11.76	45.97	16.00	17.47
Qwen3-VL-8B-T	3.79	0.42	2.74	98.60	38.48	38.76	10.08	44.41	16.73	17.16
GLM-4.6V	0.42	0.21	0.21	**100.00**	**97.47**	**97.47**	**84.87**	43.08	41.88	**47.26**
Open-weight (Instruct)										
Qwen3-VL-30B-A3B-I	18.32	2.74	7.16	97.47	0.00	0.00	0.00	52.23	1.56	3.58
Qwen3-VL-235B-A22B-I	1.47	0.42	1.47	**100.00**	92.70	92.70	84.03	43.68	39.95	46.00
Qwen3-VL-8B-I	0.42	0.21	1.05	99.44	91.85	92.42	68.91	42.84	39.47	43.79
Llama-4-Maverick-17B-128E-I	3.16	0.00	0.21	99.72	75.84	76.12	57.98	44.53	32.49	35.89
Llama-4-Scout-17B-16E-I	0.63	0.00	0.00	**100.00**	82.30	82.30	74.79	43.20	35.26	40.21

Accuracy (%). All Q1–4: all question correct on the same image.

Model	Artifact Q1 recall	Artifact All Q1–4	Ref-PG All Q2–4	Overall All Q1–4
Gemini 3.7 Flash	39.58	27.16	1.68	**56.20**
Gemini 3.8 Flash	33.68	23.37	3.36	**54.51**
GPT-6 Astra	24.00	21.05	1.68	**52.59**
GPT-6 Luna	6.53	3.16	5.88	**39.71**
GPT-6 Sol	10.74	5.47	3.36	**36.70**
Gemini 3.5 Flash Lite	22.32	7.79	0.00	**21.30**

Paper's 20 VLMs: best overall All Q1–4 = 48.0% (Claude 4.6 Opus)
Newest models: best overall All Q1–4 = 56.2% (Gemini 3.7 Flash)

Overall evaluation results

Question / Data type / Model	Artifact images			Reference images				Overall		
	Artifact			Ref-Real			Ref-PG	All splits		
	Q1 detection recall	All_{Q1-4} complete accuracy	All_{Q2-4} grounding accuracy	Q1 no artifact specificity	All_{Q1-4} complete accuracy	All_{Q2-4} grounding accuracy	All_{Q2-4} grounding accuracy	Q1 detection accuracy	All_{Q1-4} complete accuracy	All_{Q2-4} grounding accuracy
Random	50.00	0.40	0.80	50.00	0.40	0.80	0.80	50.00	0.40	0.80
Human	100.00	100.00	100.00	99.53	99.53	99.53	99.16	99.80	99.80	99.72
Closed-source										
Gemini 3.1 Pro	**99.37**	**53.26**	**53.89**	50.28	7.87	8.71	0.00	78.34	33.81	30.21
Gemini 3.1 Flash Lite	83.37	32.63	37.47	93.26	16.01	16.01	0.00	**87.61**	25.51	24.73
Claude 4.6 Opus	47.58	11.79	17.47	**100.00**	96.35	96.35	0.84	70.04	**48.01**	44.95
GPT-5.4	4.63	1.89	6.32	**100.00**	95.79	95.79	20.17	45.49	42.12	41.58
Claude 4.5 Haiku	0.42	0.00	0.84	99.44	49.16	49.16	7.56	42.84	21.06	19.79
GPT-5.4-nano	0.21	0.00	0.42	99.72	78.65	78.93	19.33	42.84	33.69	32.21
Open-weight (Thinking / Hybrid-thinking by default)										
Gemma-4-31B-it	86.32	33.26	37.68	68.26	3.93	3.93	0.00	78.58	20.70	20.31
Gemma-4-26B-it-A4B	57.89	17.68	28.00	83.71	14.61	15.17	0.00	68.95	16.37	19.68
Qwen3.5-397B-A17B	33.26	12.63	27.79	96.63	39.33	40.17	0.00	60.41	24.07	28.95
Qwen3.5-35B-A3B	20.63	6.32	18.11	96.35	39.61	40.17	0.84	53.07	20.58	24.21
Kimi K2.5	3.79	1.68	4.84	**100.00**	77.25	77.25	28.57	45.01	34.06	34.95
Qwen3-VL-30B-A3B-T	22.53	1.47	3.16	89.89	56.74	59.55	10.92	51.39	25.15	25.26
Qwen3-VL-235B-A22B-T	7.79	1.26	5.05	96.91	35.67	35.96	11.76	45.97	16.00	17.47
Qwen3-VL-8B-T	3.79	0.42	2.74	98.60	38.48	38.76	10.08	44.41	16.73	17.16
GLM-4.6V	0.42	0.21	0.21	**100.00**	**97.47**	**97.47**	**84.87**	43.08	41.88	**47.26**
Open-weight (Instruct)										
Qwen3-VL-30B-A3B-I	18.32	2.74	7.16	97.47	0.00	0.00	0.00	52.23	1.56	3.58
Qwen3-VL-235B-A22B-I	1.47	0.42	1.47	**100.00**	92.70	92.70	84.03	43.68	39.95	46.00
Qwen3-VL-8B-I	0.42	0.21	1.05	99.44	91.85	92.42	68.91	42.84	39.47	43.79
Llama-4-Maverick-17B-128E-I	3.16	0.00	0.21	99.72	75.84	76.12	57.98	44.53	32.49	35.89
Llama-4-Scout-17B-16E-I	0.63	0.00	0.00	**100.00**	82.30	82.30	74.79	43.20	35.26	40.21

Accuracy (%). All Q1–4: all question correct on the same image.

Gemini and Gemma perform best on artifact images

High recall ≠ true understanding

A model having high Q1 recall does not necessarily understand the artifact in the image.

Overall evaluation results

Question / Data type / Model	Artifact images			Reference images				Overall		
	Artifact			Ref-Real			Ref-PG	All splits		
	Q1 detection recall	All_{Q1-4} complete accuracy	All_{Q2-4} grounding accuracy	Q1 no artifact specificity	All_{Q1-4} complete accuracy	All_{Q2-4} grounding accuracy	All_{Q2-4} grounding accuracy	Q1 detection accuracy	All_{Q1-4} complete accuracy	All_{Q2-4} grounding accuracy
Random	50.00	0.40	0.80	50.00	0.40	0.80	0.80	50.00	0.40	0.80
Human	100.00	100.00	100.00	99.53	99.53	99.53	99.16	99.80	99.80	99.72
Closed-source										
Gemini 3.1 Pro	**99.37**	**53.26**	**53.89**	50.28	7.87	8.71	0.00	78.34	33.81	30.21
Gemini 3.1 Flash Lite	83.37	32.63	37.47	93.26	16.01	16.01	0.00	**87.61**	25.51	24.73
Claude 4.6 Opus	47.58	11.79	17.47	**100.00**	96.35	96.35	0.84	70.04	**48.01**	44.95
GPT-5.4	4.63	1.89	6.32	**100.00**	95.79	95.79	20.17	45.49	42.12	41.58
Claude 4.5 Haiku	0.42	0.00	0.84	99.44	49.16	49.16	7.56	42.84	21.06	19.79
GPT-5.4-nano	0.21	0.00	0.42	99.72	78.65	78.93	19.33	42.84	33.69	32.21
Open-weight (Thinking / Hybrid-thinking by default)										
Gemma-4-31B-it	86.32	33.26	37.68	68.26	3.93	3.93	0.00	78.58	20.70	20.31
Gemma-4-26B-it-A4B	57.89	17.68	28.00	83.71	14.61	15.17	0.00	68.95	16.37	19.68
Qwen3.5-397B-A17B	33.26	12.63	27.79	96.63	39.33	40.17	0.00	60.41	24.07	28.95
Qwen3.5-35B-A3B	20.63	6.32	18.11	96.35	39.61	40.17	0.84	53.07	20.58	24.21
Kimi K2.5	3.79	1.68	4.84	**100.00**	77.25	77.25	28.57	45.01	34.06	34.95
Qwen3-VL-30B-A3B-T	22.53	1.47	3.16	89.89	56.74	59.55	10.92	51.39	25.15	25.26
Qwen3-VL-235B-A22B-T	7.79	1.26	5.05	96.91	35.67	35.96	11.76	45.97	16.00	17.47
Qwen3-VL-8B-T	3.79	0.42	2.74	98.60	38.48	38.76	10.08	44.41	16.73	17.16
GLM-4.6V	0.42	0.21	0.21	**100.00**	**97.47**	**97.47**	**84.87**	43.08	41.88	**47.26**
Open-weight (Instruct)										
Qwen3-VL-30B-A3B-I	18.32	2.74	7.16	97.47	0.00	0.00	0.00	52.23	1.56	3.58
Qwen3-VL-235B-A22B-I	1.47	0.42	1.47	**100.00**	92.70	92.70	84.03	43.68	39.95	46.00
Qwen3-VL-8B-I	0.42	0.21	1.05	99.44	91.85	92.42	68.91	42.84	39.47	43.79
Llama-4-Maverick-17B-128E-I	3.16	0.00	0.21	99.72	75.84	76.12	57.98	44.53	32.49	35.89
Llama-4-Scout-17B-16E-I	0.63	0.00	0.00	**100.00**	82.30	82.30	74.79	43.20	35.26	40.21

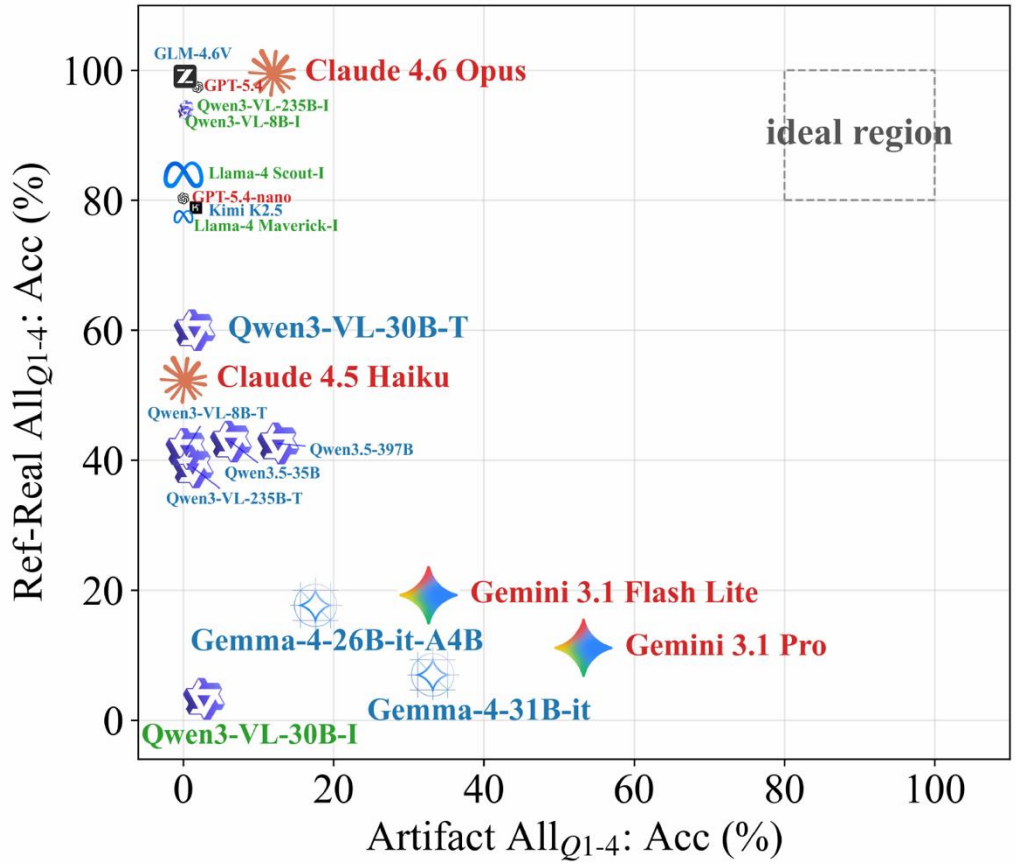
Accuracy (%). All Q1–4: all question correct on the same image.

Ref-PG is much harder than Ref-Real: models that reject artifacts on real photos still pick the removed defect in the same generated scene.

No model reaches the ideal region

Question / Data type / Model	Artifact images: Artifact: Q1 detection recall	Artifact images: Artifact: All_{Q1-4} complete accuracy	Artifact images: Artifact: All_{Q2-4} grounding accuracy	Reference images: Ref-Real: Q1 no artifact specificity	Reference images: Ref-Real: All_{Q1-4} complete accuracy	Reference images: Ref-Real: All_{Q2-4} grounding accuracy	Reference images: Ref-PG: All_{Q2-4} grounding accuracy	Overall: All splits: Q1 detection accuracy	Overall: All splits: All_{Q1-4} complete accuracy	Overall: All splits: All_{Q2-4} grounding accuracy
Random	50.00	0.40	0.80	50.00	0.40	0.80	0.80	50.00	0.40	0.80
Human	100.00	100.00	100.00	99.53	99.53	99.53	99.16	99.80	99.80	99.72
Closed-source										
Gemini 3.1 Pro	**99.37**	**53.26**	**53.89**	50.28	7.87	8.71	0.00	78.34	33.81	30.21
Gemini 3.1 Flash Lite	83.37	32.63	37.47	93.26	16.01	16.01	0.00	**87.61**	25.51	24.73
Claude 4.6 Opus	47.58	11.79	17.47	**100.00**	96.35	96.35	0.84	70.04	**48.01**	44.95
GPT-5.4	4.63	1.89	6.32	**100.00**	95.79	95.79	20.17	45.49	42.12	41.58
Claude 4.5 Haiku	0.42	0.00	0.84	99.44	49.16	49.16	7.56	42.84	21.06	19.79
GPT-5.4-nano	0.21	0.00	0.42	99.72	78.65	78.93	19.33	42.84	33.69	32.21
Open-weight (Thinking / Hybrid-thinking by default)										
Gemma-4-31B-it	86.32	33.26	37.68	68.26	3.93	3.93	0.00	78.58	20.70	20.31
Gemma-4-26B-it-A4B	57.89	17.68	28.00	83.71	14.61	15.17	0.00	68.95	16.37	19.68
Qwen3.5-397B-A17B	33.26	12.63	27.79	96.63	39.33	40.17	0.00	60.41	24.07	28.95
Qwen3.5-35B-A3B	20.63	6.32	18.11	96.35	39.61	40.17	0.84	53.07	20.58	24.21
Kimi K2.5	3.79	1.68	4.84	**100.00**	77.25	77.25	28.57	45.01	34.06	34.95
Qwen3-VL-30B-A3B-T	22.53	1.47	3.16	89.89	56.74	59.55	10.92	51.39	25.15	25.26
Qwen3-VL-235B-A22B-T	7.79	1.26	5.05	96.91	35.67	35.96	11.76	45.97	16.00	17.47
Qwen3-VL-8B-T	3.79	0.42	2.74	98.60	38.48	38.76	10.08	44.41	16.73	17.16
GLM-4.6V	0.42	0.21	0.21	**100.00**	**97.47**	**97.47**	**84.87**	43.08	41.88	**47.26**
Open-weight (Instruct)										
Qwen3-VL-30B-A3B-I	18.32	2.74	7.16	97.47	0.00	0.00	0.00	52.23	1.56	3.58
Qwen3-VL-235B-A22B-I	1.47	0.42	1.47	**100.00**	92.70	92.70	84.03	43.68	39.95	46.00
Qwen3-VL-8B-I	0.42	0.21	1.05	99.44	91.85	92.42	68.91	42.84	39.47	43.79
Llama-4-Maverick-17B-128E-I	3.16	0.00	0.21	99.72	75.84	76.12	57.98	44.53	32.49	35.89
Llama-4-Scout-17B-16E-I	0.63	0.00	0.00	**100.00**	82.30	82.30	74.79	43.20	35.26	40.21

Accuracy (%). All Q1–4: all question correct on the same image.



Sensitive models over-claim (bottom right)
Calibrated models under-detect (top left)
No model is both (upper right is empty)

Conclusion

- SalArt-VQA turns each artifact claim into a hierarchy — presence, region, box, evidence — and pairs every artifact image with an artifact-free reference to test abstention.
- Image-level detection hides failures: the best VLM detects 99.4% of artifacts but gets the full chain right on only 53.3% of images.
- No current VLM is both sensitive and calibrated: sensitive models over-claim on references, while calibrated models miss real artifacts.

Dataset: huggingface.co/datasets/salartvqa/SalArt-VQA **Project:** zhangruotian.github.io/SalArt-VQA

Thank you