NeurIPS 2026 · Evaluations & Datasets Track · **Spotlight**

KnowVis

Dual-View Evaluation and Diagnosis of World-Knowledge Grounding in Text-to-Image Models

Yonghang Yu* Ruoyu Wang* Haoyu Zhang Lin Qi Yawei Zhou Changqian Yu

KlingAI Research

**Equal contribution*

Project page quantumwhisper42.github.io/KnowVis

Code github.com/QuantumWhisper42/KnowVis **Data** huggingface.co/datasets/QuantumWhisper42/KnowVis

Drawing what is mentioned ≠ drawing it correctly

EXPLICIT

"a red sphere on a table"

Tests rendered objects, attributes and spatial relations

IMPLICIT KNOWLEDGE

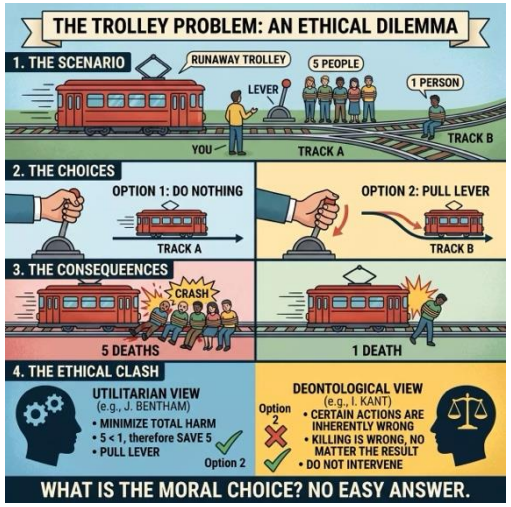
"a convex lens converges parallel light rays"

The image must also obey an optical rule

Real prompts carry implicit facts, physical commonsense, symbolic conventions, domain rules and cultural knowledge.

Same prompt, two models

"Use an image to represent the trolley problem, and add concise text explanations when necessary."



Nano Banana 2

✓ Knowledge 5/5 · Aesthetics 5/5



Midjourney v7

✗ Knowledge 1/5 · Aesthetics 4/5

Beautiful ≠ knowledge-correct

Human MOS scores from the representative scoring example (Appendix F.4)

Existing evaluations leave three questions open

Prior work covers pieces

- **Compositional binding**
- **Physical commonsense**
- **Factual generation**
- **Beyond-prompt knowledge**

Shared limitations

- Specific knowledge types or task formats
- Holistic, score-based judgments
- Little attribution: difficulty, education level, text / annotation
- Limited validation of the automatic judge

1

What is world-knowledge grounding?

A task definition with a clear boundary vs. alignment and image quality

2

How to cover atomic and open-ended use?

Verify single facts, and apply them in realistic, composed scenes

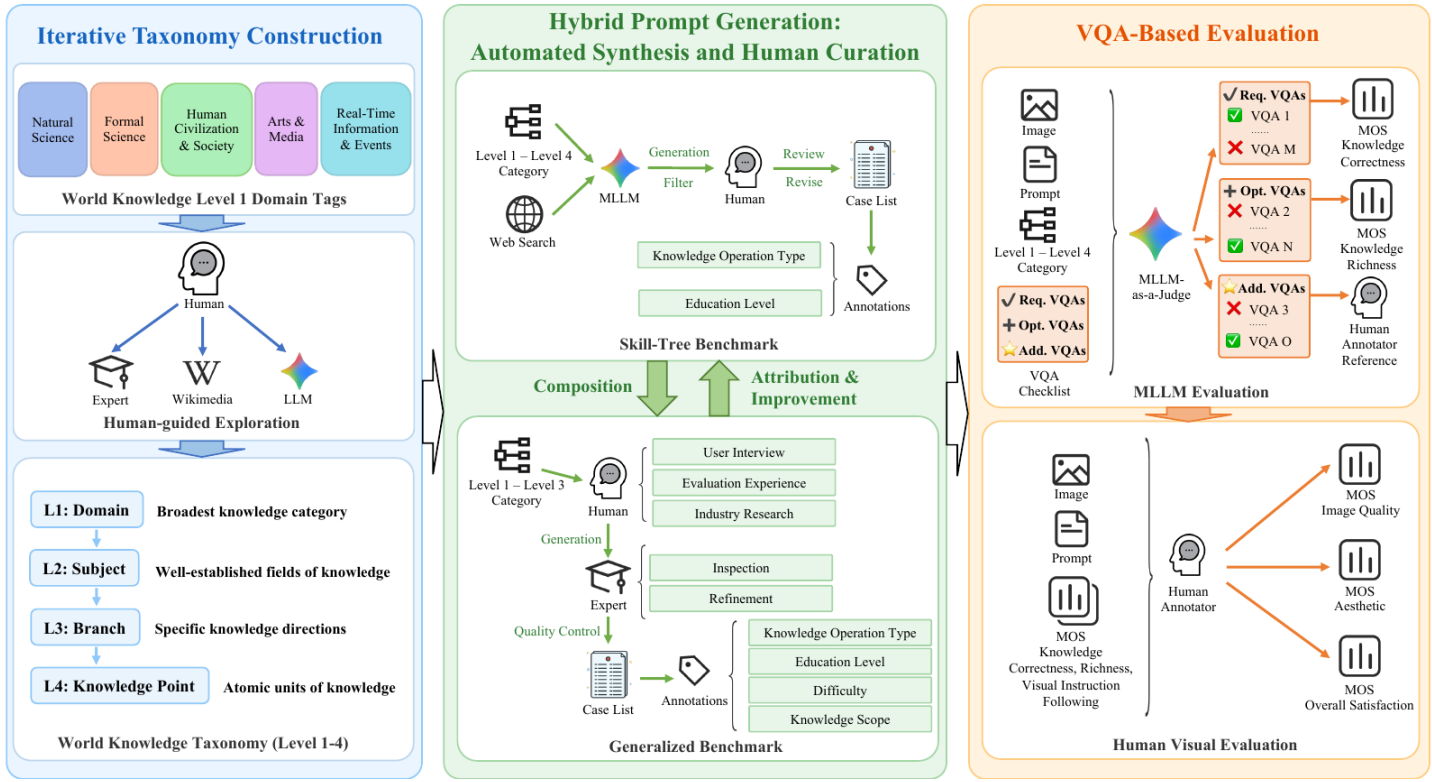
3

How to evaluate reliably and diagnostically?

Signals that agree with humans and localize where failures occur

KnowVis: definition and framework

World-knowledge grounding = the ability to **identify, retrieve, reason over and visually express** the explicit or implicit knowledge constraints required by a prompt, in a **visually verifiable** way.



Three contributions

1 Unified task formulation

Definition and boundary vs. alignment, quality and instruction following

2 Dual-view benchmark

Skill-Tree + Generalized on one shared L1–L4 taxonomy

3 Attributable protocol

Structured VQA-based MLLM-as-a-Judge with diagnostic signals

An L1–L4 taxonomy organizes world knowledge

Level	Unit	Example path (torii case)	Role
L1	**Domain**	*Human Civilization & Society*	Cross-domain coverage and comparison
L2	**Subject**	*Humanities & Architecture*	Stable discipline: structural backbone
L3	**Branch**	*Architectural Styles*	Scenario unit → Generalized benchmark
L4	**Knowledge point**	*Japanese Architecture and Key Components*	Atomic claim with a stable criterion → Skill-Tree, VQAs, attribution

5 L1 domains

Natural Science | **Formal Science** | **Human Civilization & Society** | **Arts & Media** | **Real-Time Information & Events**

Subjective or time-sensitive domains stay at L2/L3 granularity. Built iteratively from encyclopedias, library classifications, LLM expansion, user interviews and expert review.

Two complementary views on one taxonomy

Skill-Tree Benchmark

like a unit test

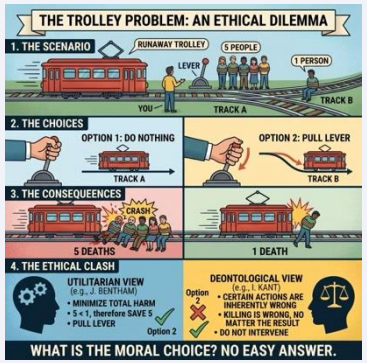


"Can the model express this specific piece of knowledge correctly?"

Unit	L4 atomic knowledge point
Setting	Controlled, one focused target
Judging	Automatic · Required VQAs · strict

Generalized Benchmark

like an integration test

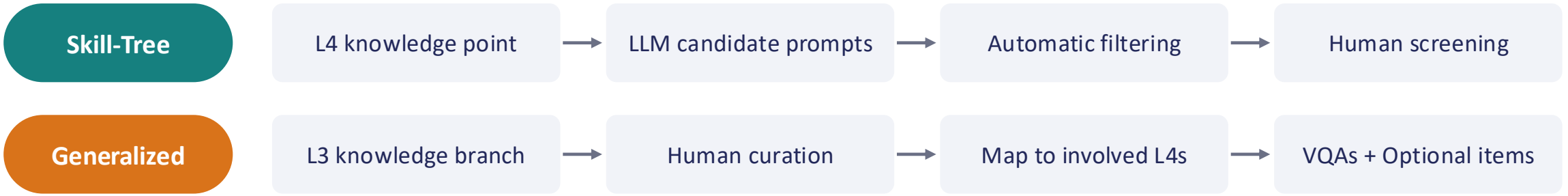


"Can the model apply and compose knowledge in realistic scenes?"

Unit	L3 knowledge branch
Setting	Open-ended, multi-knowledge
Judging	Human–judge collaborative (VQA + MOS)

Shared taxonomy: Generalized cases are mapped to involved L4 points, enabling cross-view attribution.

Separate construction, unified attribution



The L4 mapping gives partial attribution for open-ended failures and reveals taxonomy coverage gaps.

Case-level metadata

Knowledge structure

- Skill-Tree: one L1–L4 path; L4 is the target point
- Generalized: L3 branch + set of involved L4 points (L2 if no stable split)

Task attributes

- Education level: primary school → professional areas
- Knowledge operation: Explicit Rendering · Explicit Recall · Implicit Recall · Reasoning
- Difficulty and knowledge quantity (Generalized)
- Bilingual prompts (English / Chinese)

Scale and coverage



Taxonomy coverage up to L3 (Figure 2)

1,622

Skill-Tree cases

- 240 L4 points
- 3 L1 domains
- ≈ 6.8 cases per L4
- 3 operation types

550

Generalized cases

- 70 L3 branches
- 20 L2 subjects
- 5 L1 domains
- 4 operation types + difficulty

Coverage- and evaluability-driven, not an exhaustive distribution of world knowledge.

Decomposing knowledge judgment into verifiable VQAs

Prompt

"Draw a torii gate at the entrance of a traditional Japanese shrine."



Visual grounding: a Wikimedia reference photo is attached for the judge.

VQA checklist (Appendix E, Table 12)

1. Gate-like structure: two pillars and a top crossbar	REQUIRED	VISUAL
2. Matches torii traits (kasagi, nuki), not a Chinese paifang	REQUIRED	KNOWLEDGE
3. Stands at a shrine entrance: path, forest, lanterns	REQUIRED	KNOWLEDGE
4. If text is added: names torii / shrine, spelled correctly	OPTIONAL	KNOWLEDGE
5. Added text is bound to the right structure	OPTIONAL	KNOWLEDGE

Required

Minimal evidence for the core constraint → correctness

Optional

Predefined extras → richness; absent = NA, wrong = FAIL

Additional

Found on the fly → refines the Optional pool (not scored)

Fields: component · knowledge_hint · visual_grounding
Judge: Gemini 3.1 Pro Preview → PASS / FAIL / NA + rationale (protocol is judge-agnostic)

Two views, two scoring rules

Skill-Tree

Case correct ⇔ every Required VQA = PASS

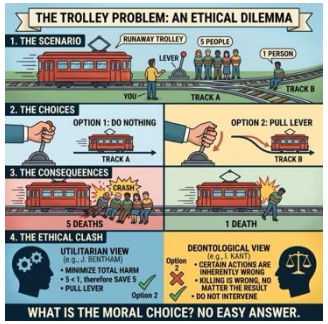
Strict pass/fail on the L4 point's minimal knowledge frame; fully automatic and scalable.

Generalized

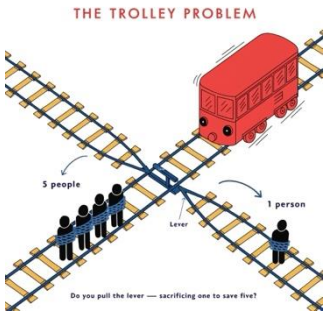
Judge (VQA)	**Knowledge correctness**	**Knowledge richness**	**Instruction following**
Human MOS	**Image quality**	**Aesthetics**	**Overall satisfaction**

Richness: missing or wrong = 0 · judge scores mapped to the MOS scale · a factual error alone never lowers image quality

One Generalized prompt on the trolley problem, four models (human scores, 1–5)



Nano Banana 2
Knowledge 5
Aesthetics 5
Overall 5



Recraft v4
Knowledge 5
Aesthetics 3
Overall 4



Midjourney v7
Knowledge 1
Aesthetics 4
Overall 2



FLUX.2 Pro
Knowledge 3
Aesthetics 2
Overall 3

From scores to diagnosis

Signal level

- Required pass / fail
- Optional coverage
- Instruction following
- Annotation failures
- Human MOS

×

Slice level

- L1–L4 taxonomy path
- Education level
- Knowledge operation
- Difficulty
- Knowledge quantity

Skill-Tree attribution

1. Find low-scoring L4 points (case level)
2. Group failed Required VQAs by component
3. Map back to L3: local gap or branch weakness?

Generalized attribution

Correctness, richness, instruction following and MOS across L2/L3 branches and task metadata

Which component fails tells why the model fails (VQA component labels from the Skill-Tree checklists)

1 Cannot recall the concept

ENTITY · ENTITY_TRAIT

e.g., a torii drawn as a Chinese paifang

2 Wrong attributes or relations

ATTRIBUTE_BINDING · ENTITY_RELATION

e.g., a "falling" apple resting on the ground

3 Cannot express via annotations

ANNOTATION · ANNOTATION_BINDING · ANNOTATION_RELATION

e.g., wrong labels, formulas, arrows, callouts

Experimental setup

Skill-Tree · 480 cases

240 visible dev cases (judge-prompt iteration) + 240 unseen validation cases (model ranking)

Generalized · 100 cases

Covers all 20 L2 subjects and 70 L3 branches; human–judge collaborative

General-Domain T2I · 100 cases

Non-knowledge control set, sampled from a 600-case pool

10 text-to-image models

Nano Banana 2	Seedream 5.0 Lite
Seedream 4.5	Recraft v4
HunyuanImage 3.0	Qwen Image 2.0
Wan 2.7 Pro	FLUX.2 Pro
Kling Image 3.0 Omni	Midjourney v7

Bilingual prompts: each model uses its best-matching language; case, intent, metadata and protocol stay fixed.

The judge agrees with human annotators

94.51%

Skill-Tree VQA-level accuracy
Cohen's κ = 0.824

88.12%

Skill-Tree case-level accuracy
Cohen's κ = 0.756

98.41%

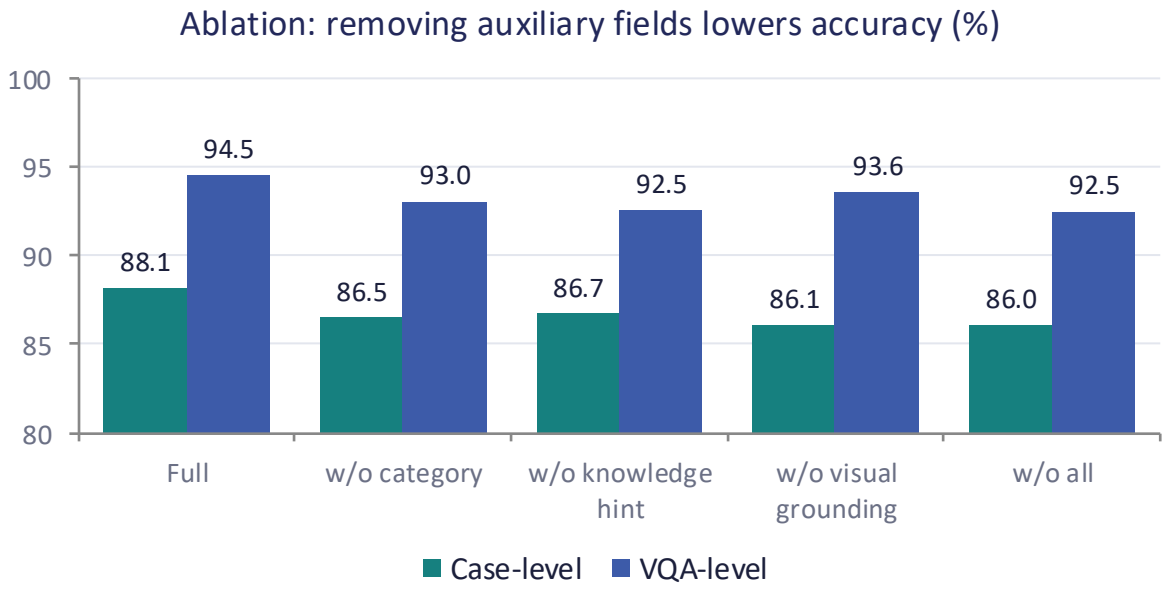
Generalized Required-VQA accuracy

98.53%

Generalized Optional-VQA accuracy

How reliable, and where it errs

- 480 images from 6 models; FAIL is the positive class
- Precision 92.07%: flagged errors are real
- Recall 80.22%: ~1 in 5 human-labeled errors missed
- Human reference: 5 annotators, 288 VQAs, Krippendorff's α = 0.766, 82.64% unanimous



Rankings are stable across judges

Judge setting	Acc.	Prec.	Rec.	κ	Kendall τb	Spearman ρ
Gemini 3.1 Pro (original)	94.51	92.07	80.22	0.824	–	–
Gemini 3.1 Pro (rerun)	91.74	83.63	74.44	0.737	0.822	0.927
Gemini 3.5 Flash	92.21	86.74	73.41	0.748	0.822	0.939
GPT 5.5	91.33	86.03	69.11	0.714	0.822	0.939

> 91%

VQA accuracy in every judge setting

Top-2

unchanged in every setting

≤ 2

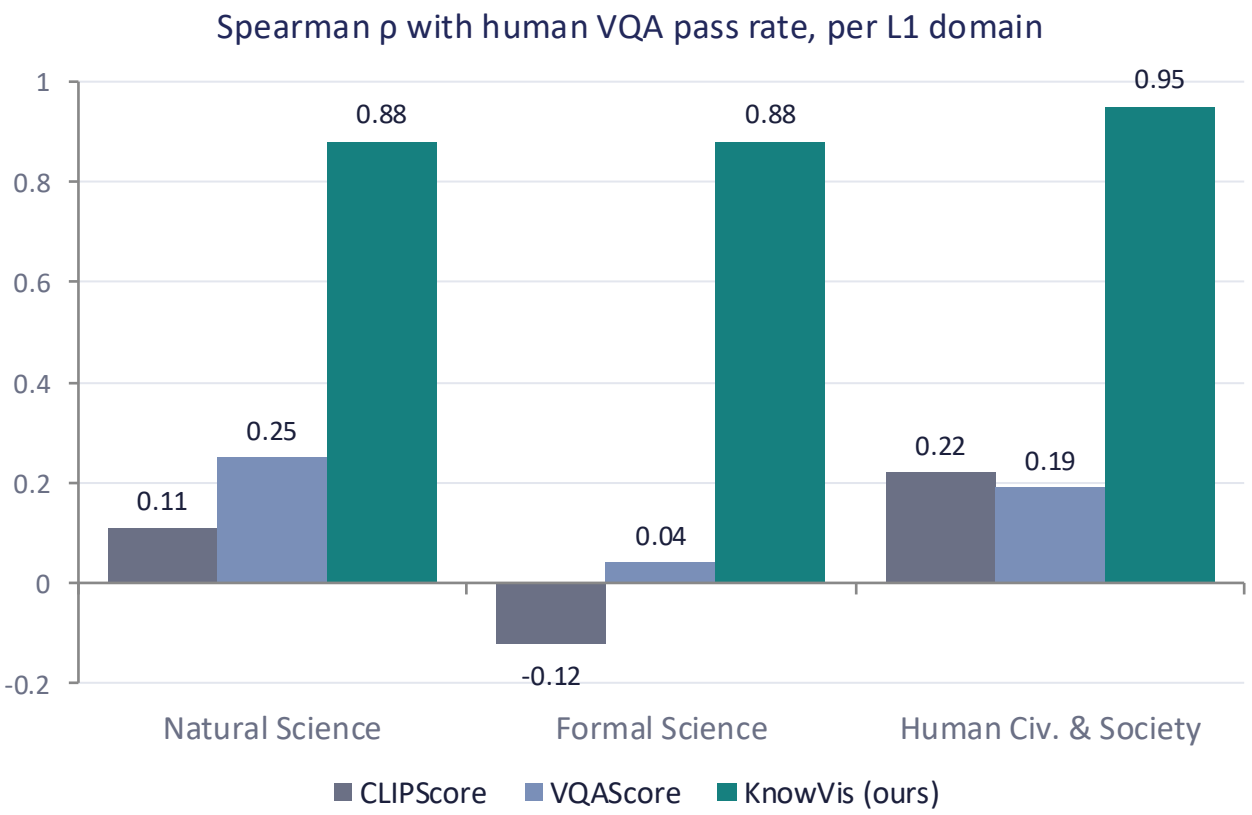
positions moved by any model

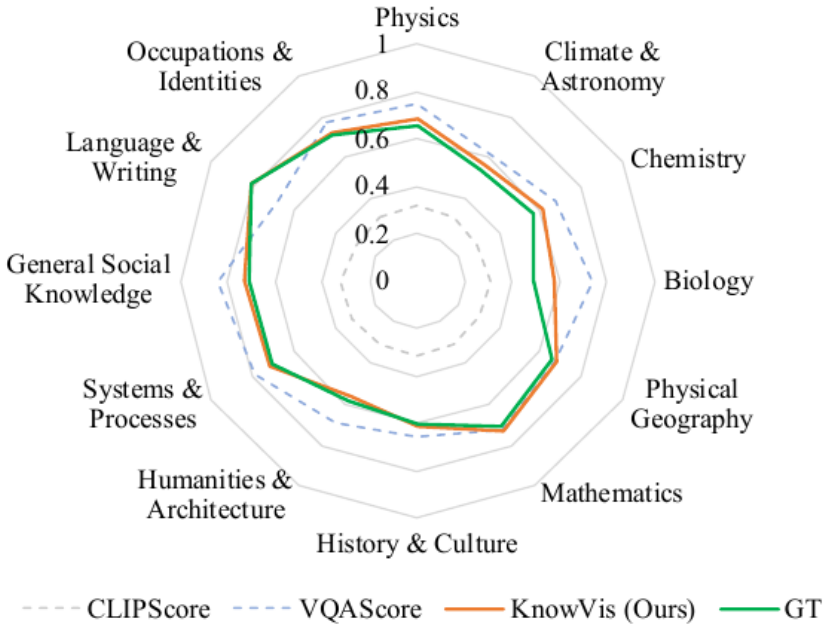
Human-correction audit (50 prompt–image pairs, 5 models): fixing judge errors keeps the top model and only swaps #2 and #3 ($\tau = 0.80$, $\rho = 0.90$).

Where the judge still errs

- Visual reasoning: human orientation, coordinates, text and graphic recognition, structure
- Too strict on decorative or handwritten text
- Too lenient on duplicated elements
- Lowest agreement in relation-, binding- and annotation-type VQAs

General alignment metrics miss knowledge errors

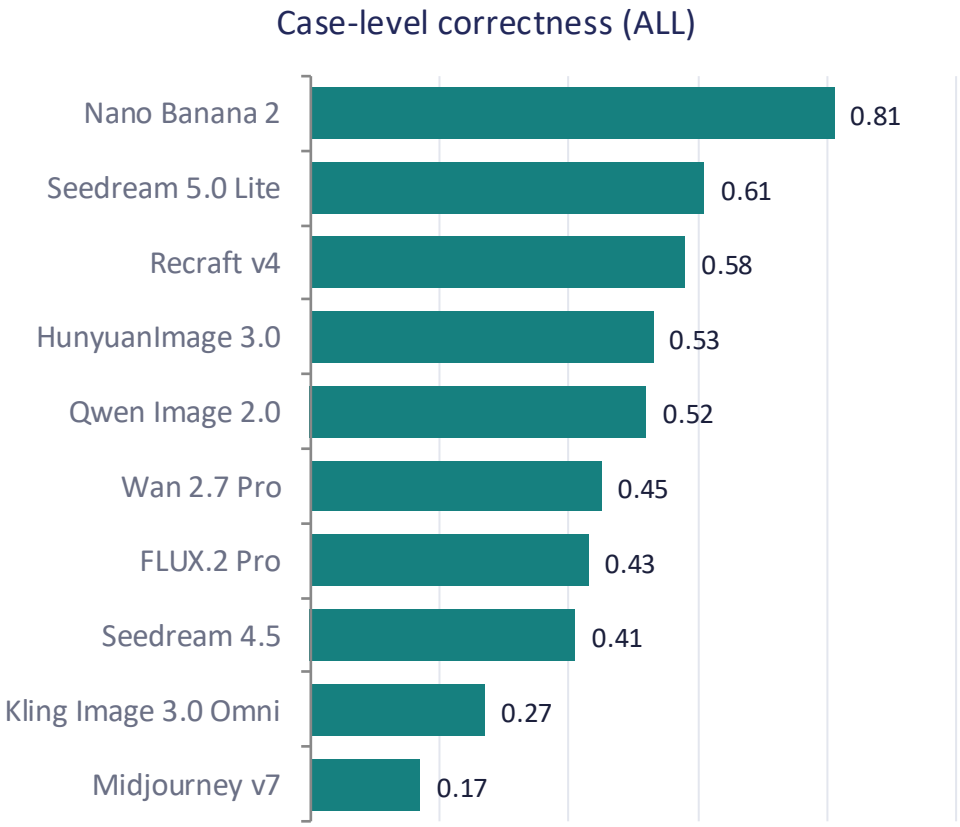




L2-level profile over 12 subjects: KnowVis (orange) tracks human GT (green); CLIPScore and VQAScore (dashed) do not.

Formal Science: CLIPScore is negatively correlated with humans ($\tau = -0.09$, $\rho = -0.12$). KnowVis reaches τ 0.74–0.85 and ρ 0.88–0.95 across domains.

Skill-Tree: objects are easy, annotations are hard



	FLUX.2 Pro	HunyuanImage 3.0	Midjourney v7	Nano Banana 2	Qwen Image 2.0	Recraft v4	Seedream 4.5	Seedream 5.0 lite	Wan 2.7 Pro	Kling Image 3.0 Omni
ALL	0.59	0.59	0.34	0.80	0.59	0.69	0.58	0.66	0.68	0.46
ENTITY	0.88	0.86	0.69	0.94	0.82	0.90	0.81	0.83	0.89	0.77
ENTITY_TRAIT	0.67	0.64	0.39	0.84	0.68	0.68	0.51	0.70	0.53	0.42
ENTITY_RELATION	0.60	0.67	0.30	0.85	0.69	0.75	0.57	0.73	0.70	0.51
ATTRIBUTE_BINDING	0.68	0.74	0.39	0.83	0.77	0.74	0.73	0.79	0.78	0.55
ANNOTATION	0.38	0.35	0.09	0.68	0.36	0.54	0.46	0.48	0.54	0.27
ANNOTATION_BINDING	0.29	0.26	0.05	0.67	0.33	0.48	0.26	0.49	0.48	0.14
ANNOTATION_RELATION	0.16	0.27	0.03	0.67	0.25	0.43	0.32	0.39	0.46	0.10
TEXT	0.38	0.37	0.07	0.68	0.38	0.56	0.47	0.55	0.57	0.28
NON_TEXT	0.67	0.67	0.44	0.85	0.68	0.74	0.62	0.70	0.72	0.54

1.00

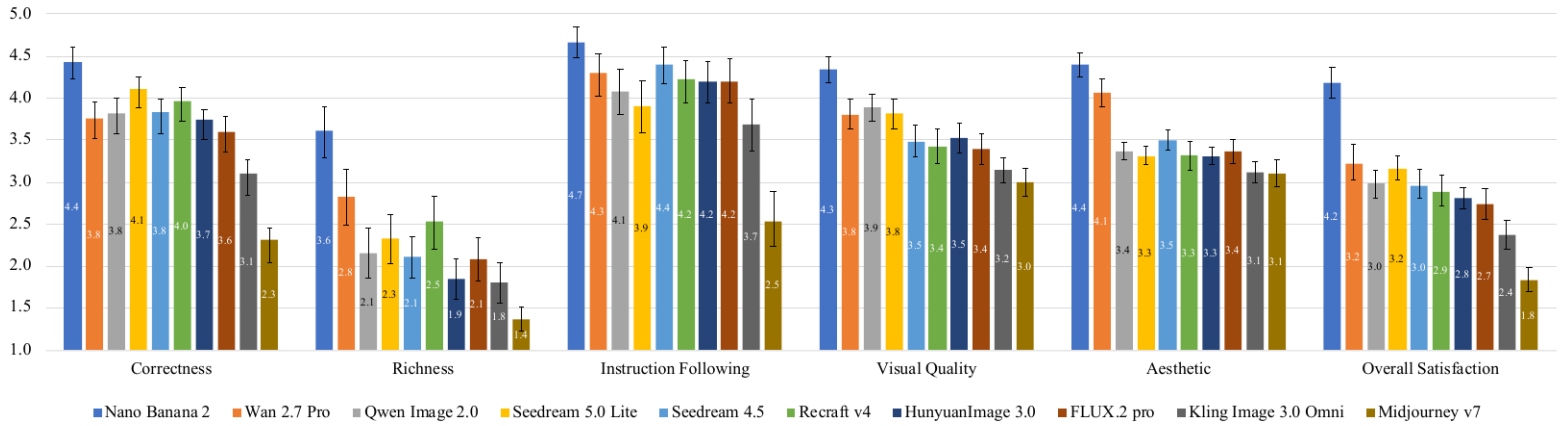
0.00

ENTITY: 0.69 – 0.94
Relevant objects are usually drawn

ANNOTATION_RELATION: 0.03 – 0.67
Labels, formulas, arrows, bindings fail

Formal Science: 0.02 – 0.92
One overall score hides domain gaps

Generalized: aesthetics ≠ knowledge



MOS (1–5): correctness · richness · instruction following · visual quality · aesthetics · overall satisfaction

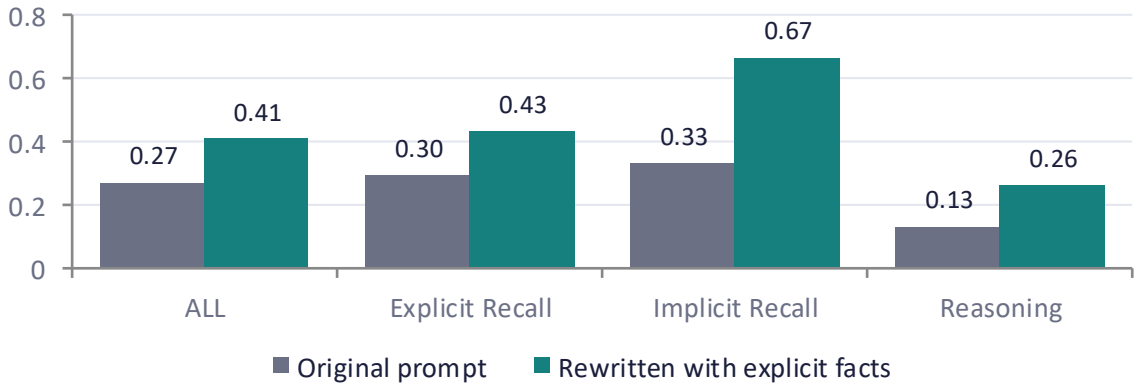
Midjourney v7
Competitive aesthetics, lowest correctness and richness

Nano Banana 2
Leads correctness, richness, instruction following and overall

Consistent across views

- Weak Skill-Tree Formal Science → lower Generalized MOS on formal-science cases
- General-domain quality and aesthetics do not imply knowledge grounding
- Some failures come from knowledge invocation, not rendering (preliminary)

Prompt rewriting on 240 unseen cases: pass rate



Take-aways

1. World-knowledge grounding needs dedicated diagnostic evaluation, beyond alignment, image quality or aesthetics

2. KnowVis: L1–L4 taxonomy + Skill-Tree + Generalized + structured VQA judge, reliable vs. humans and robust across judges

3. Key bottleneck: relations and annotations (labels, formulas, arrows, text–image binding)

Limitations and next steps

- Refine knowledge-richness modeling
- Residual knowledge / quality coupling in human ratings
- Relies on strong MLLM judges; small judges not yet reliable
- Static images only → editing, video, multi-turn generation

Project quantumwhisper42.github.io/KnowVis
Code github.com/QuantumWhisper42/KnowVis
Data huggingface.co/datasets/QuantumWhisper42/KnowVis

Thank you. Come visit our poster!