Visual Generation for Understanding and Reasoning

Semantic supervision and geometric evidence

Yanwei Li
liyanwei@sjtu.edu.cn
Shanghai Jiao Tong University



OVERVIEW

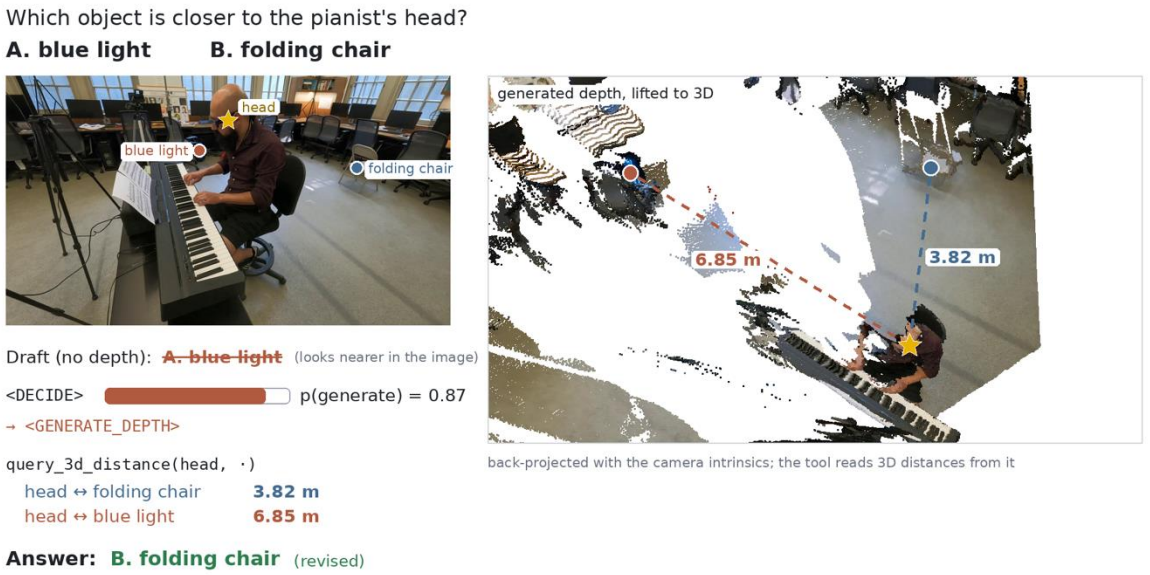
Generation as training supervision and reasoning evidence

SGT



RGB Semantic Instance

GeoUG



Semantic targets

Which targets improve understanding and generation?

Geometric supervision and evidence

What geometry should a model learn, and when should it generate?

MOTIVATION

Unified architecture, different learning signals

Understanding

Sparse text supervision

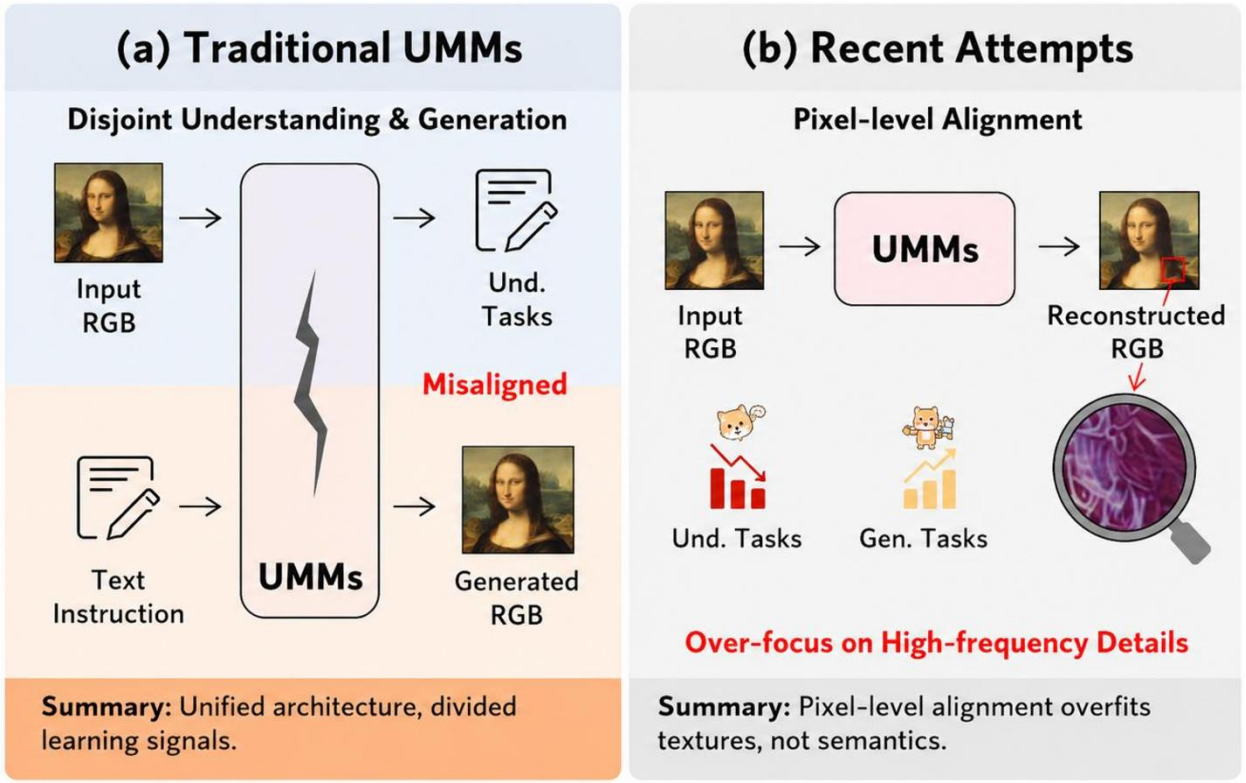
VQA teaches what to say about an image.

Generation

Dense visual supervision

Pixel objectives teach how to reproduce appearance.

The question is how to align learning targets.

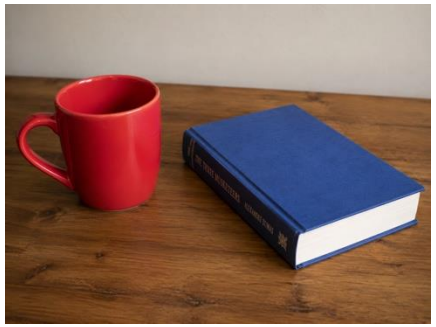


MOTIVATION

Both capabilities depend on the same visual structure

UNDERSTANDING

- What is present?
- Where is it?
- How are objects related?



GENERATION

- Which objects?
- Which attributes?
- Which spatial composition?

SHARED VISUAL SIGNAL

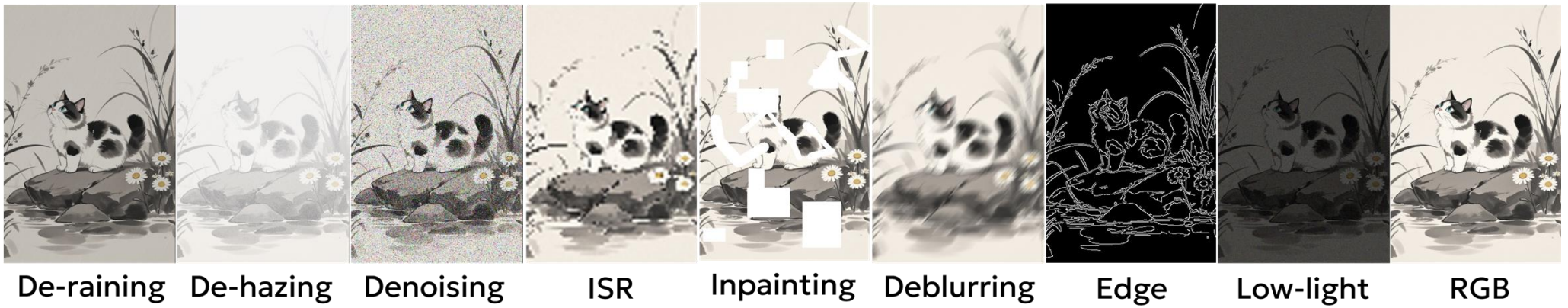
Semantics
Structure and layout

A common architecture still needs aligned learning targets.

QUESTION

Which visual target provides the right level of abstraction?



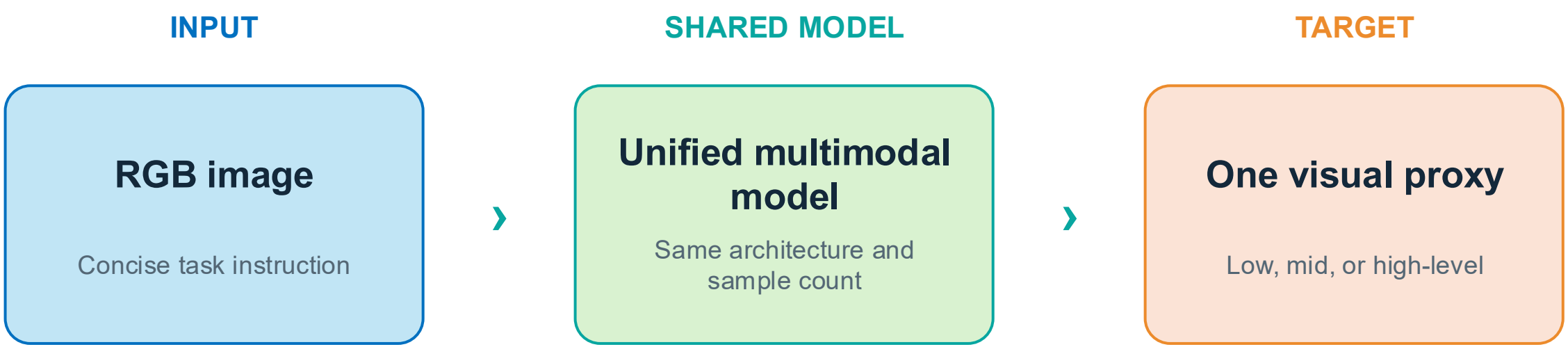


Texture → **Structure** → **Semantics**

The proxy must filter texture without discarding spatial structure

CONTROLLED STUDY

Controlled comparison of visual proxy targets

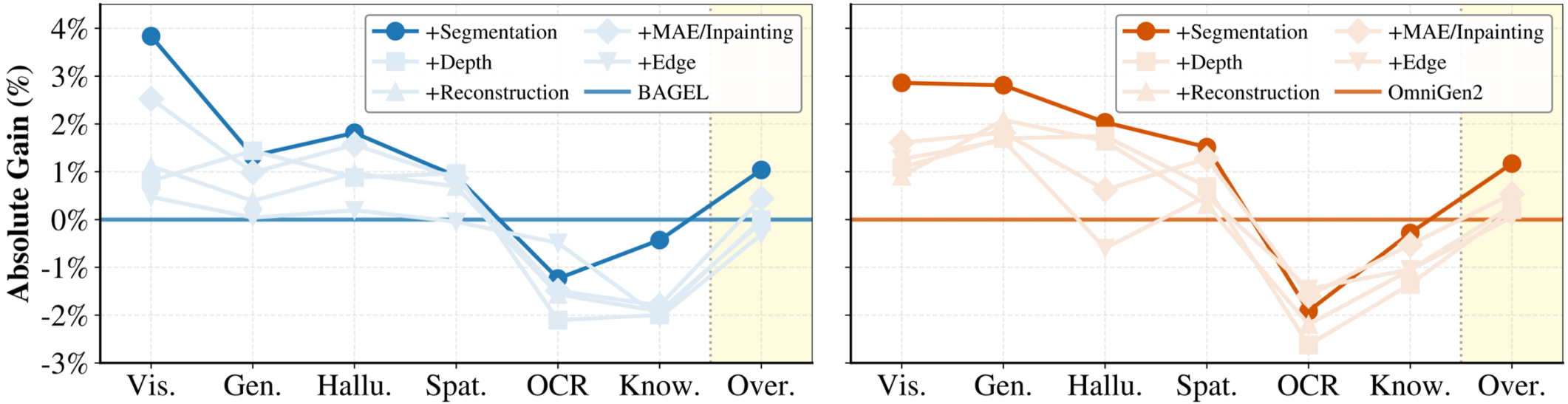


***No* VQA data, text-to-image data, and image editing data in the proxy study.**

Over 95% of source images overlap across tasks

OBSERVATION

High-level proxies deliver the largest understanding gains



Segmentation leads

Largest gains

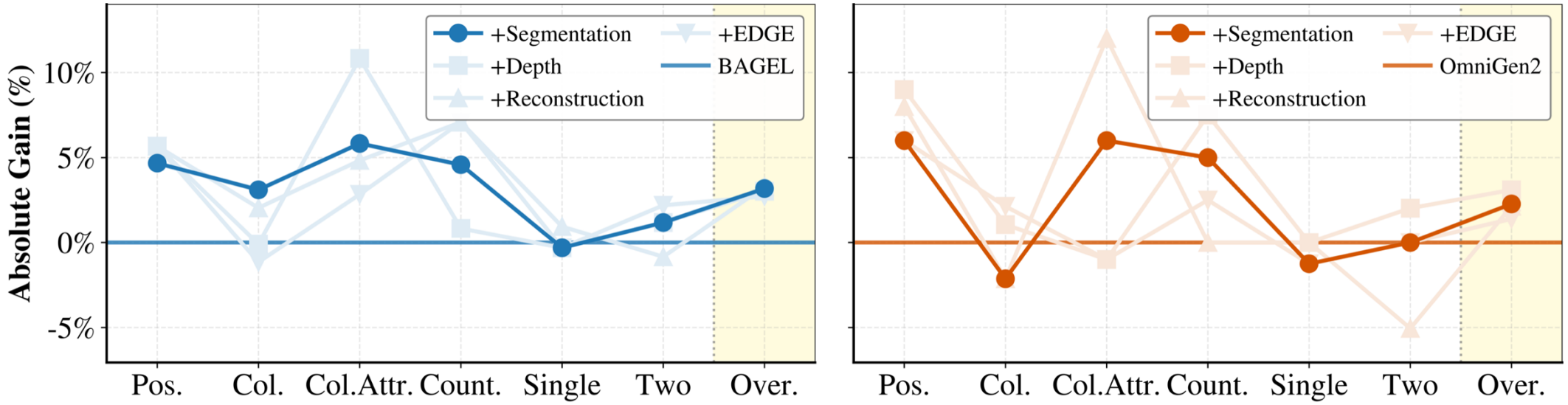
Vision-centric perception, spatial reasoning, hallucination resistance

Limited gains

OCR, document tasks, external knowledge

OBSERVATION

All four tested proxies improve position-aware generation

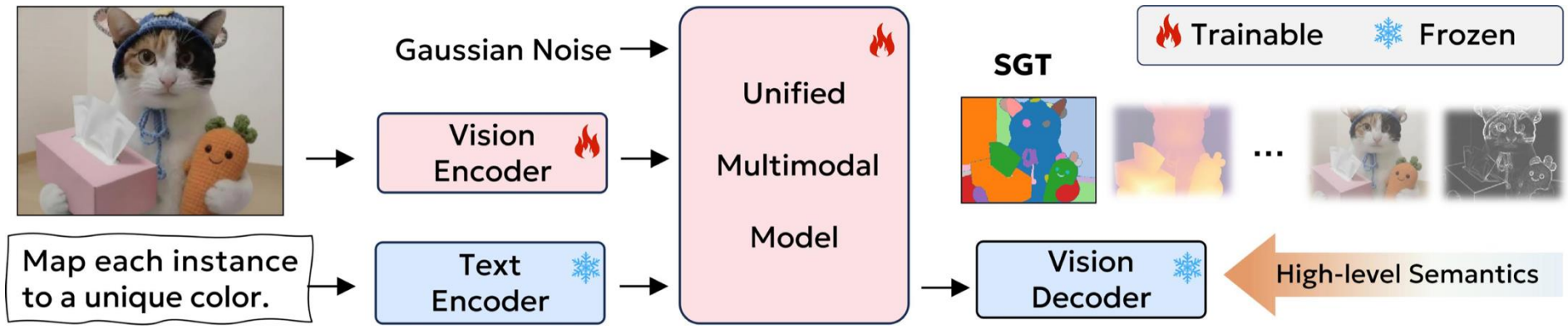


The common effect

Layout is a shared benefit of the tested visual targets

METHOD

SGT turns segmentation into a generative alignment target



1 Colorize segments

Map each segment or instance to a unique target color.

2 Condition on the image

Pair the RGB image with a concise visual-task instruction.

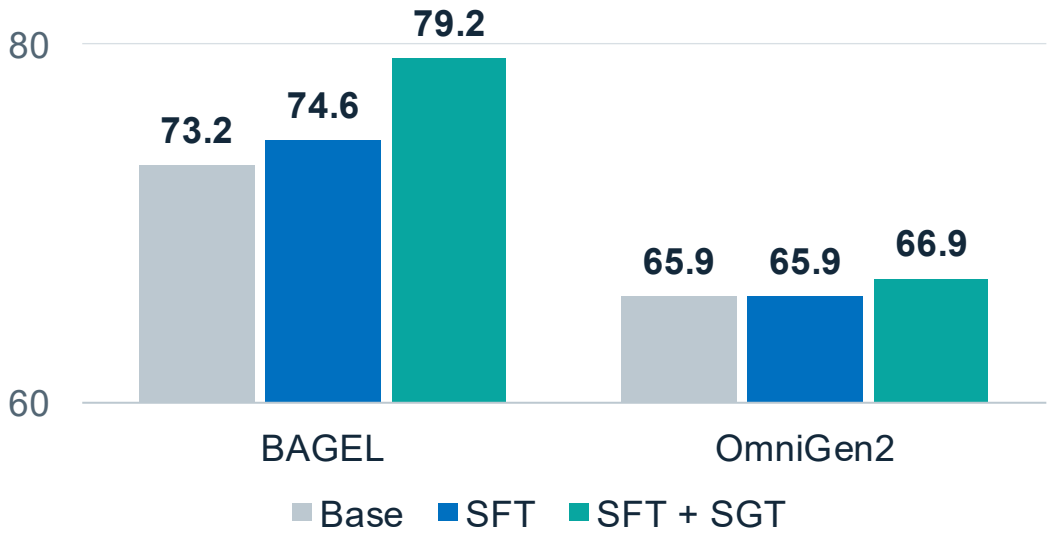
3 Generate semantics

Train the UMM to generate semantic visual targets.

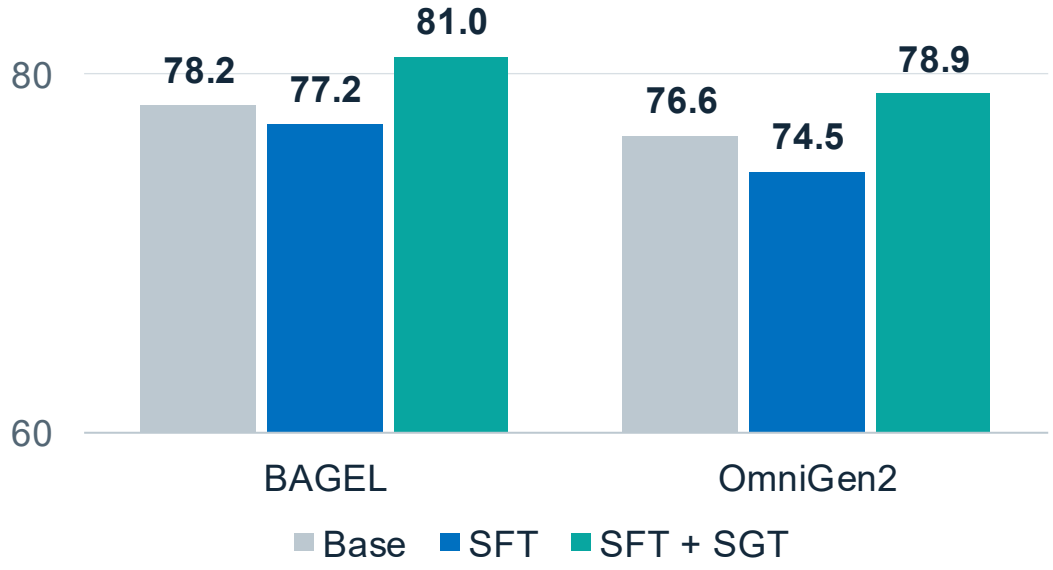
SGT RESULTS

Joint SFT + SGT improves both measured capabilities

CV-Bench



GenEval

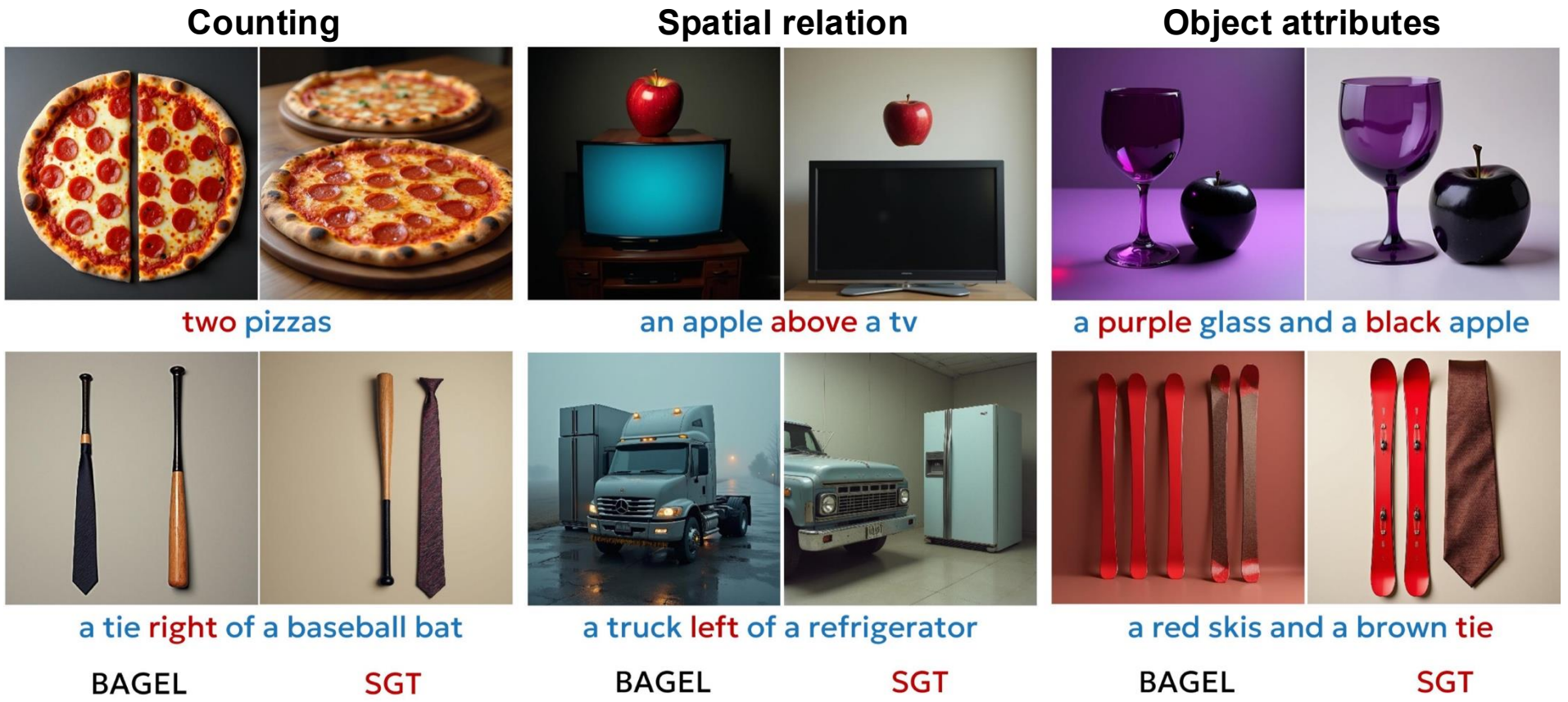


GenEval averaged over 12 seeds.

With a prompt rewriter, Table 2 reports BAGEL GenEval 88.0 to 90.0.

EXAMPLES

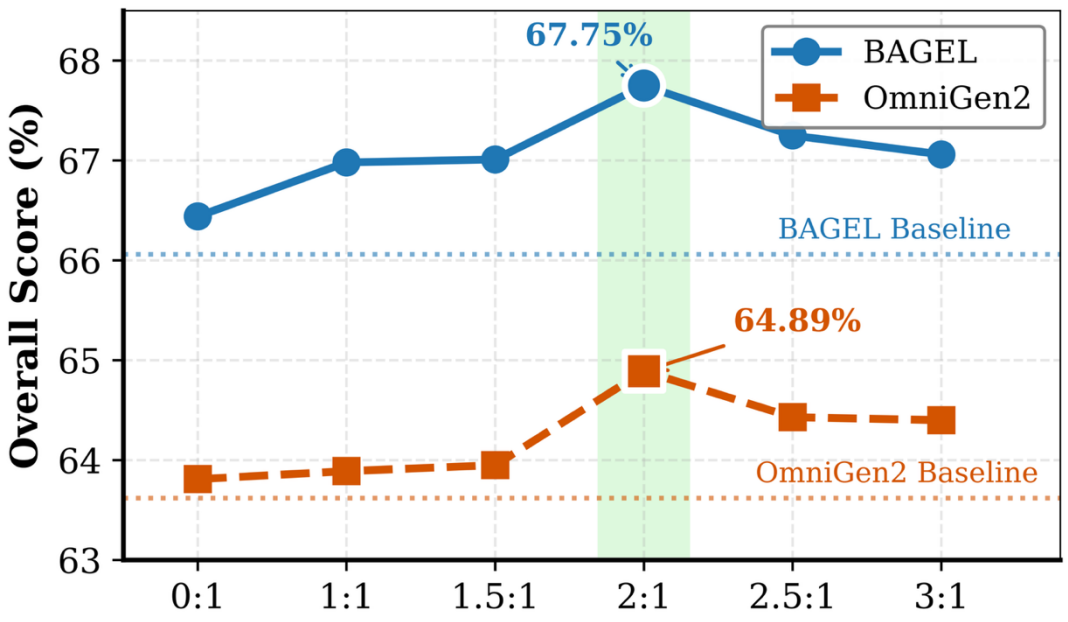
SGT follows compositional prompts more faithfully



SGT TRAINING RECIPE

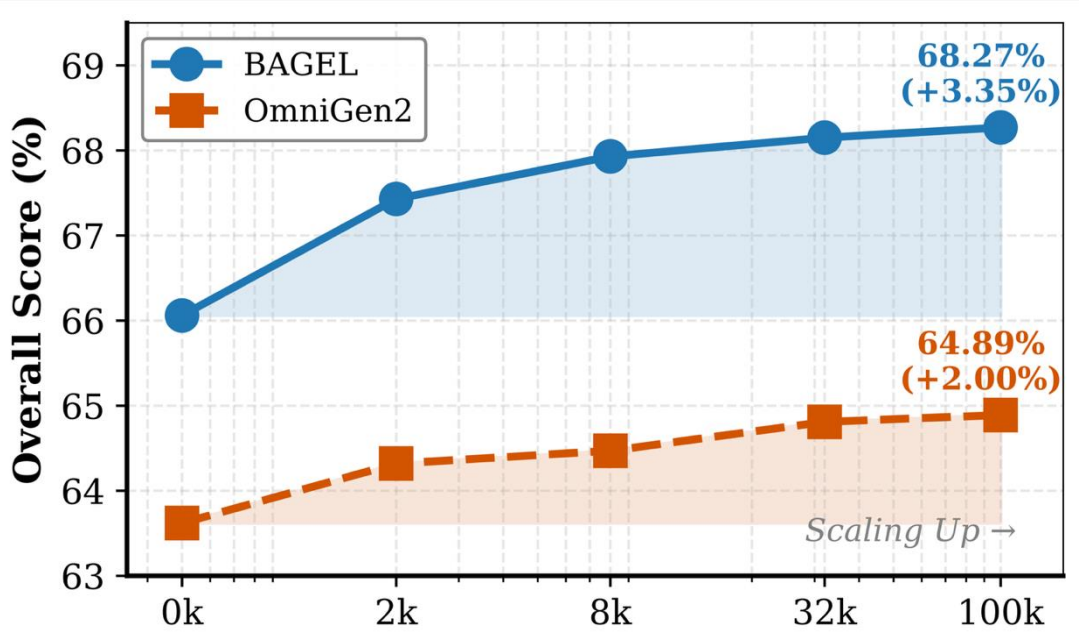
Segmentation ratio and data scale

Sampling ratio



2:1 segmentation to VQA in each batch

Segmentation data scale



Understanding gains grow, then flatten

VISUAL FEATURES

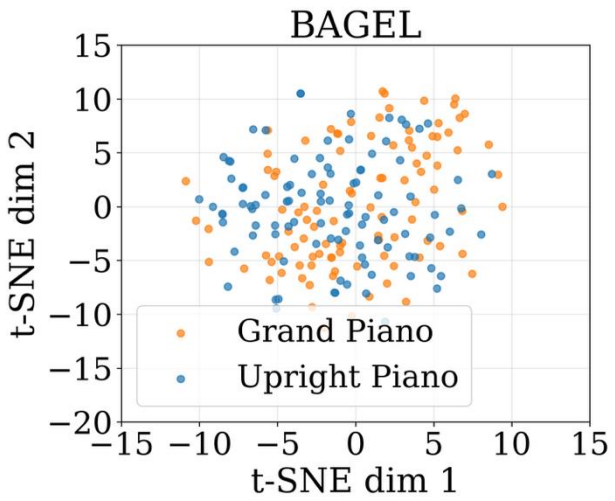
Feature projections show clearer category separation after SGT

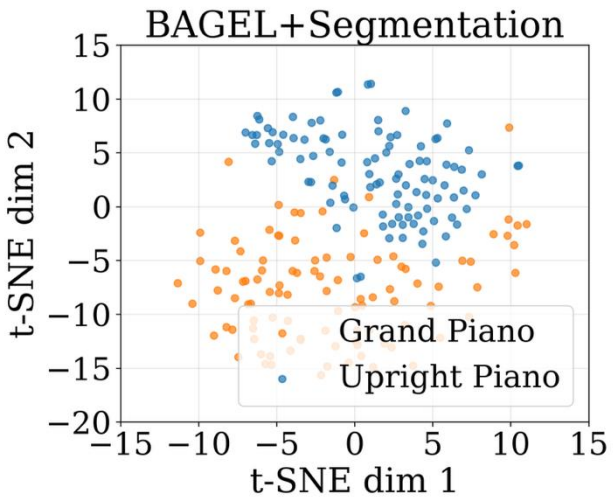
Confusable categories



Grand Piano vs Upright Piano

Visual feature embeddings





BAGEL

Overlapping clusters

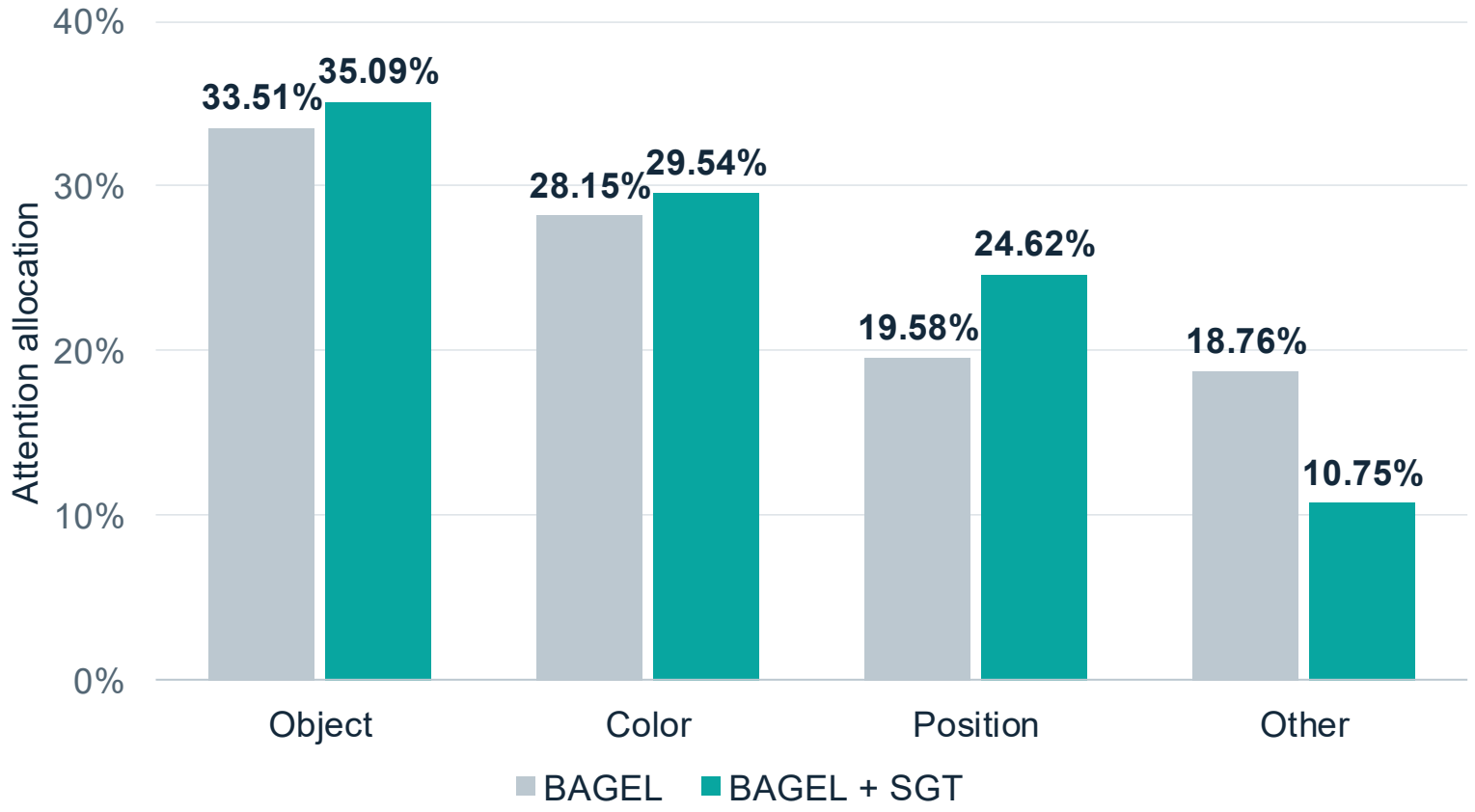
BAGEL + SGT

Clearer separation in projection

Qualitative t-SNE evidence

TOKEN ATTENTION

SGT shifts attention toward composition-critical words



+5.04 pp

Position-token attention

19.58% to 24.62%

+1.58 pp

Object-token attention

33.51% to 35.09%

Attention shifts toward object, color and position tokens

GEOMETRIC GENERATION

Visual generation captures rich spatial structure

Novel viewpoints with coherent scene layout and object identity



Input view



Generated view: move forward left

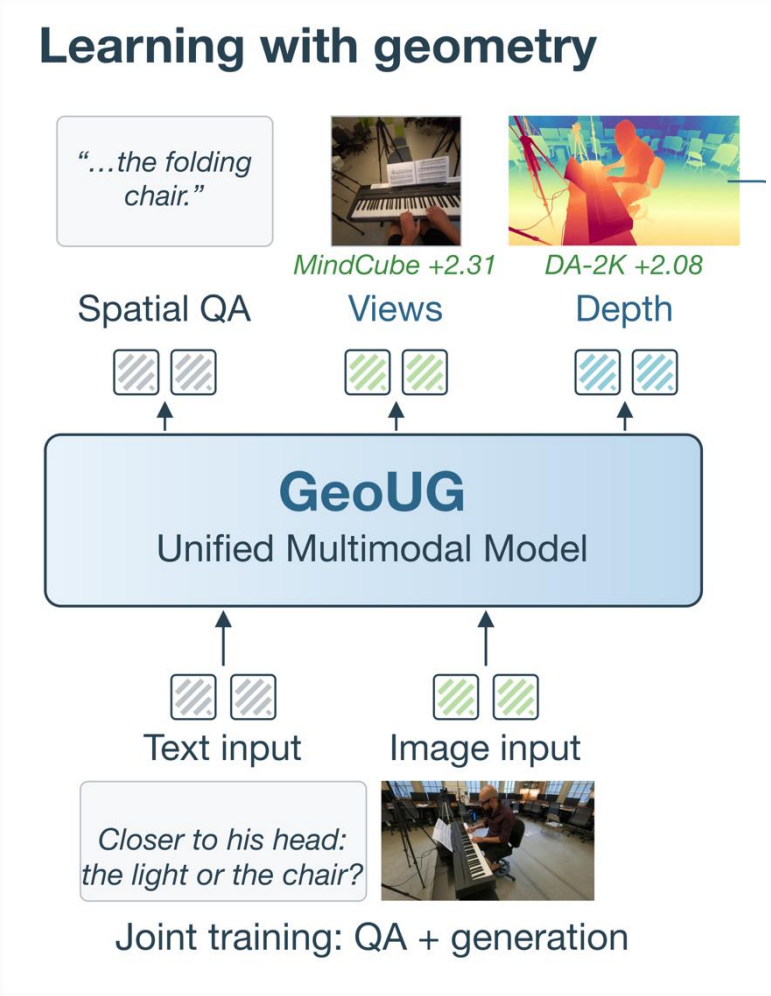


Generated view: look left

How can these generative capabilities improve spatial understanding?

GEOUG: LEARNING WITH GEOMETRY

Geometric supervision for spatial representation



Spatial QA

Supervision on the relations we ask about

Depth supervision

Dense geometry across visible surfaces

View generation

Spatial changes under a new viewpoint

Co-training on BAGEL

Spatial QA + one geometric target

MMDC · Encode: three pivots, three RGB channels

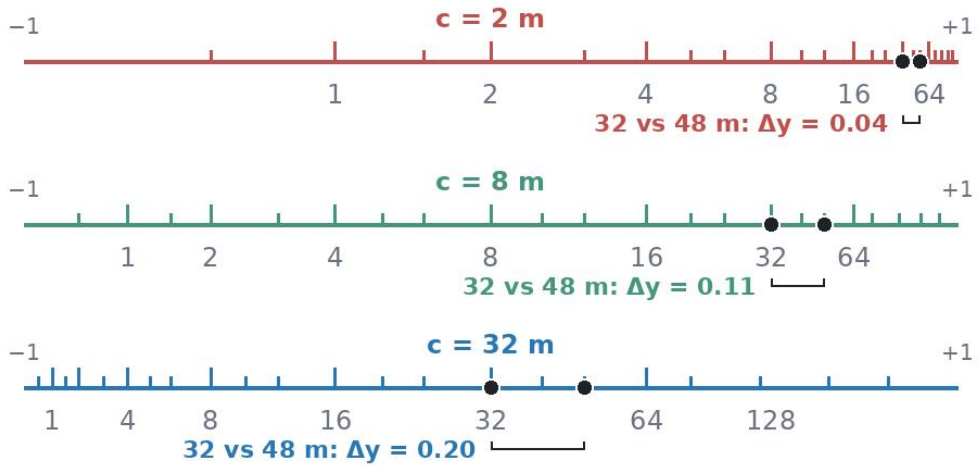


$$y = (d - c) \,/\, (d + c) \in (-1, 1) \qquad Y_{\mathrm{k}} = (d - c_{\mathrm{k}}) \,/\, (d + c_{\mathrm{k}}), \quad c_{\mathrm{k}} \in \{2, 8, 32\}\ \mathrm{m}$$

MMDC · Generate: an MMDC image, like any RGB image

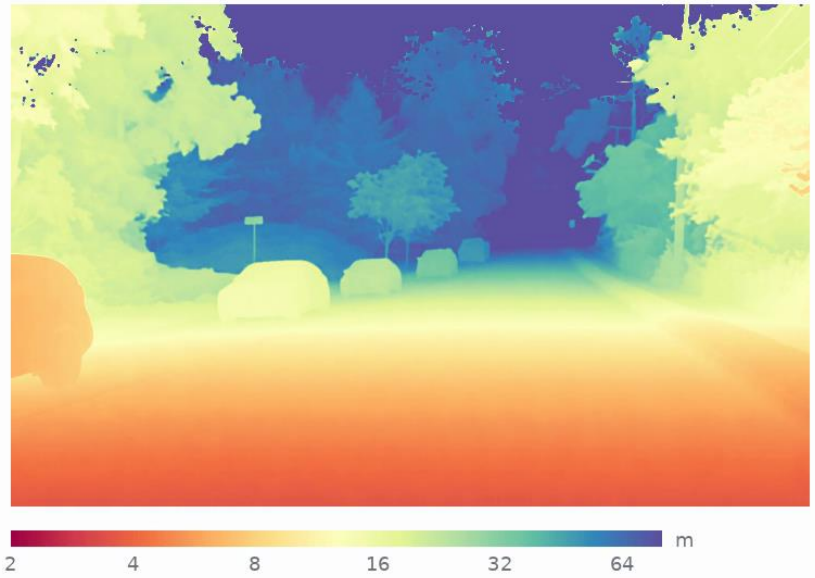


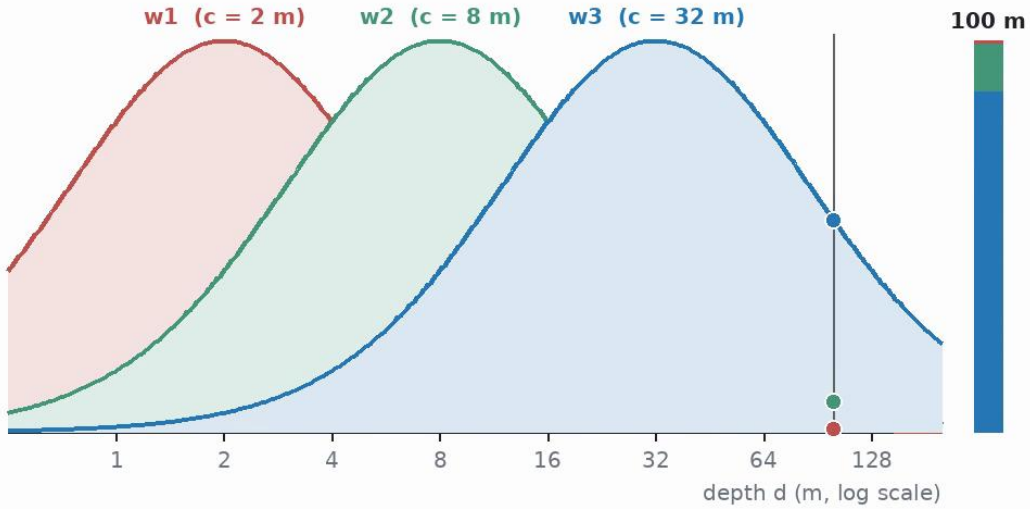
one MMDC image: channel i encodes depth with pivot c_i



Frozen VAE: $3 \times H \times W \rightarrow 16 \times (H/8) \times (W/8)$ Flow matching: noise → MMDC latent

MMDC · Decode: back to metres, no learned parameters

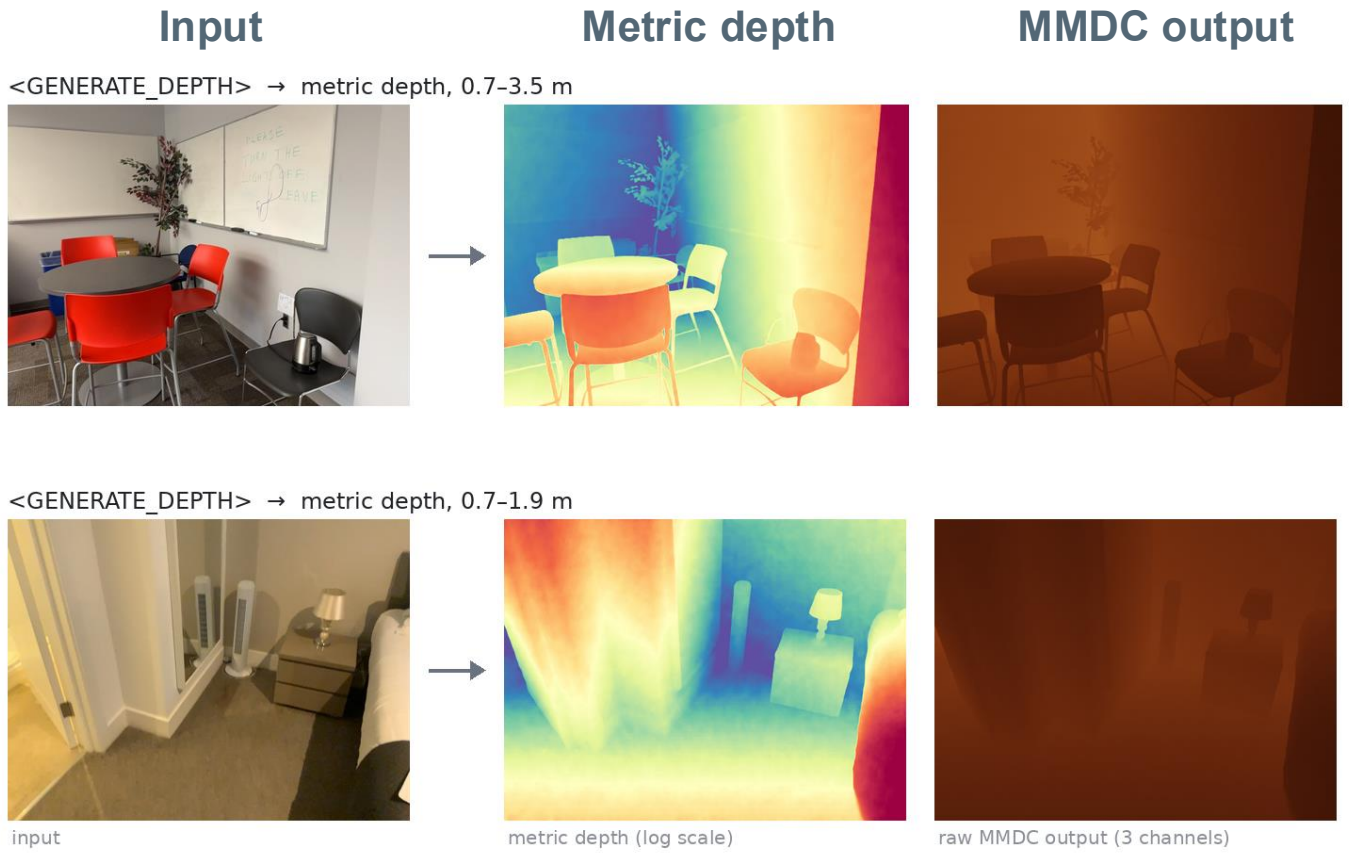




$$\hat{d}_k = c_k\,(1 + \hat{Y}_k)\,/\,(1 - \hat{Y}_k) \qquad w_k = (1 - \hat{Y}_k^{\,2})^2 \qquad \hat{d} = \sum_k w_k\,\hat{d}_k\,/\,\sum_k w_k$$

GEOUG: LEARNING WITH GEOMETRY

Depth generation learns scene geometry



AbsRel↓ / δ_1↑	NYU-v2	KITTI	ETH3D
DA V2	4.16 / 97.9	6.77 / 94.3	4.63 / 97.2
MoGe-V2	2.89 / 98.6	3.75 / 98.1	2.80 / 99.1
Marigold-v1.1	5.50 / 96.4	10.5 / 90.2	6.90 / 95.7
Lotus	5.40 / 96.8	8.50 / 92.2	5.90 / 97.0
GeoUG-Depth	**3.32 / 98.5**	**4.84 / 97.3**	**3.75 / 97.9**

Encoding (5-set avg.)	AbsRel	δ_1
single pivot (c = 5 m)	7.47	92.2
MMDC (2, 8, 32 m)	7.24	92.6
MMDC + more data	**6.23**	**93.9**

Affine-aligned, AbsRel ×100 and δ_1 (%)

GEOUG: LEARNING WITH GEOMETRY

Camera-conditioned multi-view generation

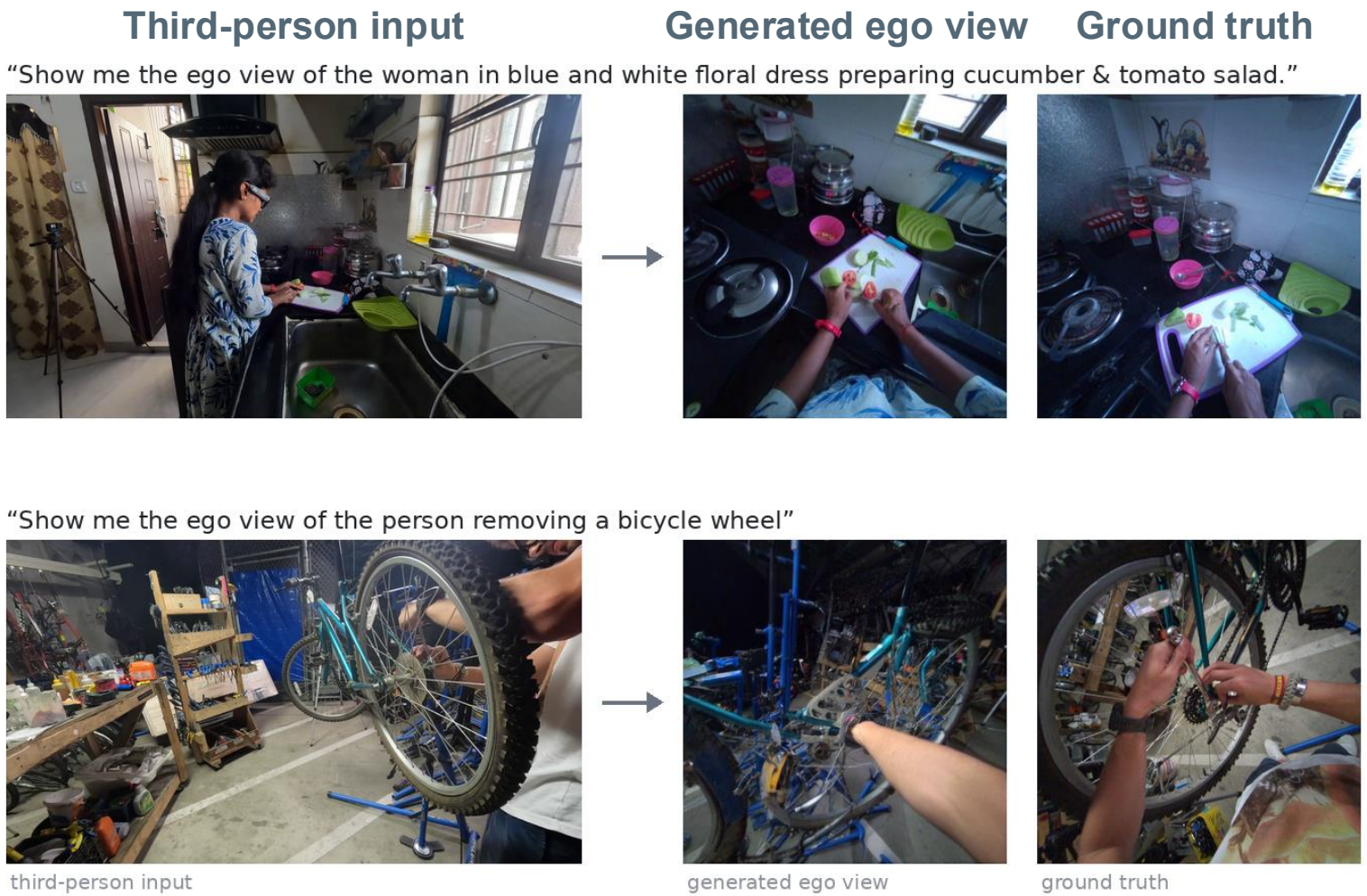


Model	PSNR ↑	SSIM ↑	LPIPS ↓
BAGEL	13.16	0.476	0.520
GeoUG	**13.85**	**0.495**	**0.477**

- Camera motion changes layout and visibility.
- Generating the target view supervises spatial change.

GEOUG: LEARNING WITH GEOMETRY

Exo-to-ego generation: perspective taking

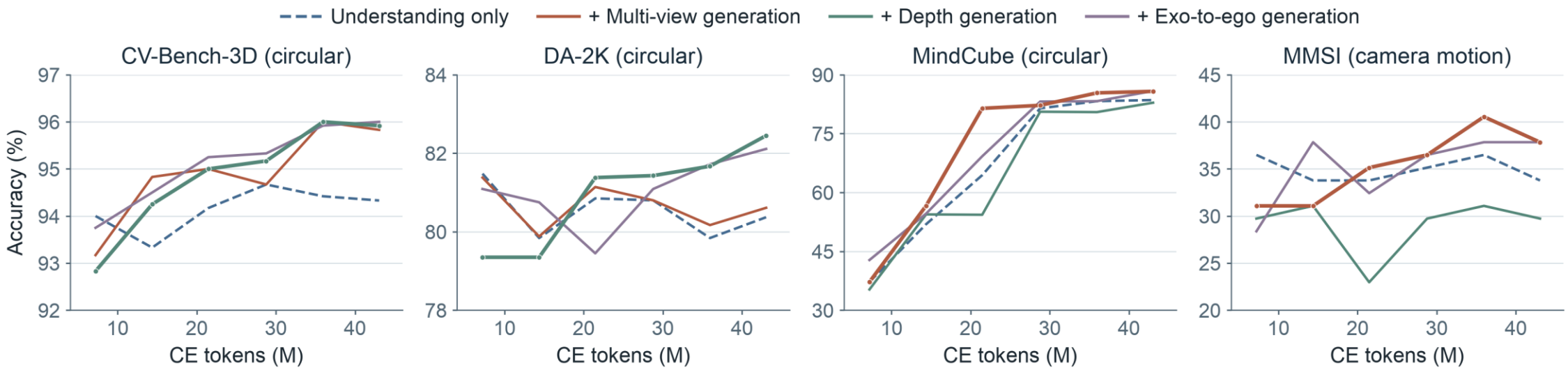


Model	PSNR ↑	SSIM ↑	LPIPS ↓
BAGEL	7.75	0.312	0.839
GeoUG	**12.41**	**0.398**	**0.670**

- GeoUG generates the scene from the person's viewpoint.
- The task requires perspective taking and scene geometry.

GEOUG: LEARNING WITH GEOMETRY

Geometric targets teach different spatial skills



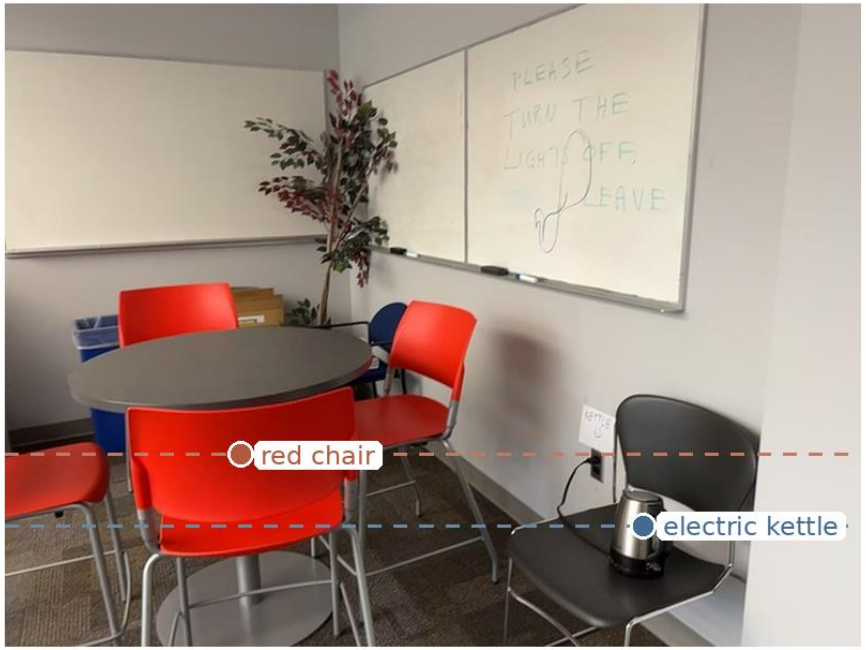
Depth supports depth judgments. View generation supports perspective reasoning.

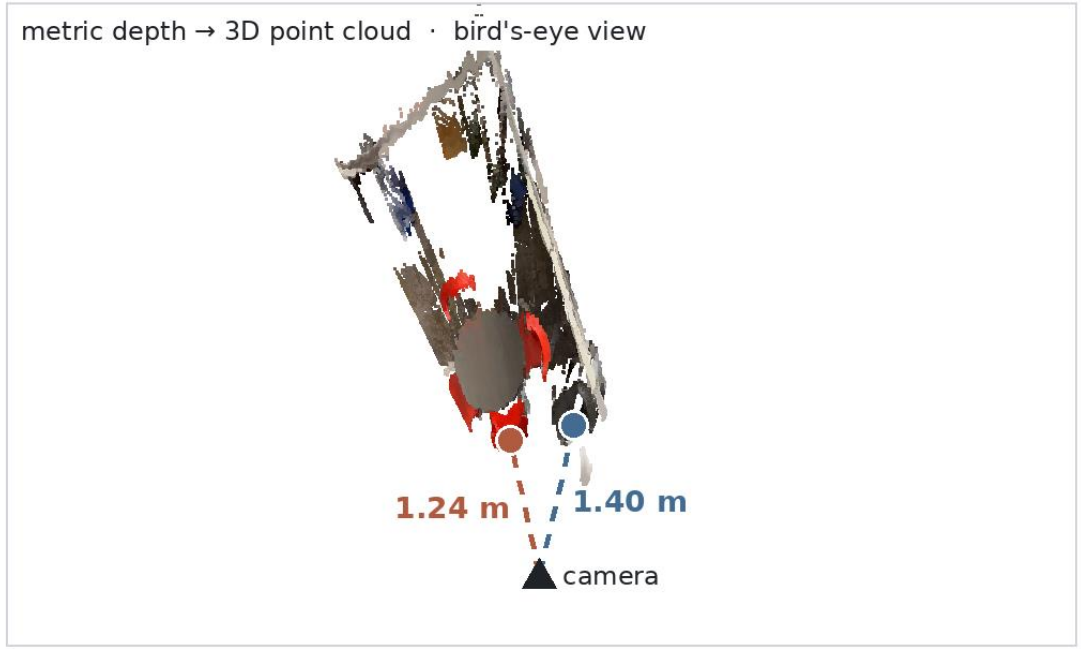
Training setup	CV-Bench-3D-Visual Circular acc. (%)	DA-2K-Visual Circular acc. (%)	MindCube Circular acc. (%)	MMSI camera motion Accuracy (%)
Understanding only	94.33	80.37	83.56	33.78
+ Depth	95.92	**82.45**	82.88	29.73
+ Multi-view	95.83	80.61	85.77	**37.84**
+ Exo-to-ego	**96.00**	82.11	**85.87**	**37.84**

GEOUG: REASONING WITH GEOMETRY

Image-plane cues can bypass 3D reasoning

Which is closer to the camera: the electric kettle or the red chair?





~~2D cue "lower in the image is closer" → electric kettle~~

3D: the red chair is closer (1.24 m vs 1.40 m)

How far do such 2D rules go on existing benchmarks?

GEOUG: REASONING WITH GEOMETRY

3DCueBench: balanced 2D cues and difficulty

Depth



Compare camera depths of two objects

cue-conflicting: lower is farther

Distance



Compare 3D distances to a reference object ★

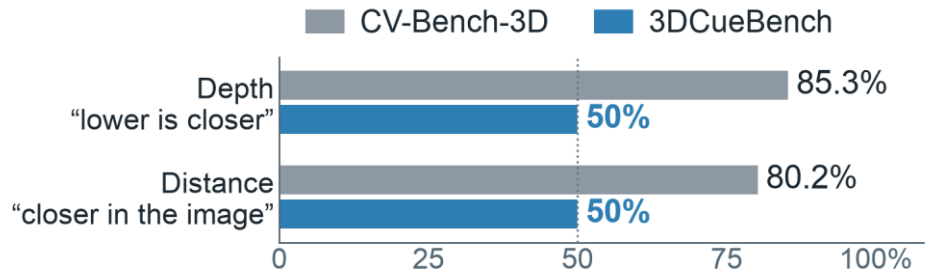
cue-conflicting: A looks closer to ★

Relative gap $g = |d_A - d_B| \,/\, \max(d_A, d_B)$

Examples: 36.6 vs 42.0 m, 5.01 vs 5.75 m $\rightarrow$ $g = 0.13$

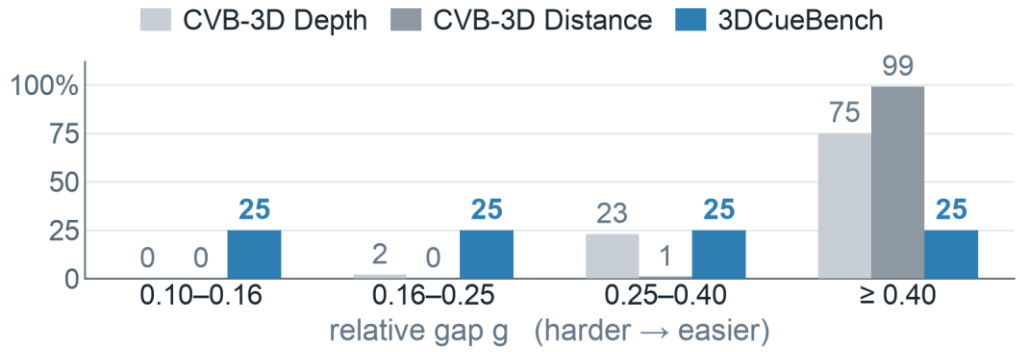
Smaller g is harder · 4 bins × 100 questions per task

Selected 2D rules score 50%



cue-consistent: rule right · cue-conflicting: rule wrong

Hard comparisons are covered

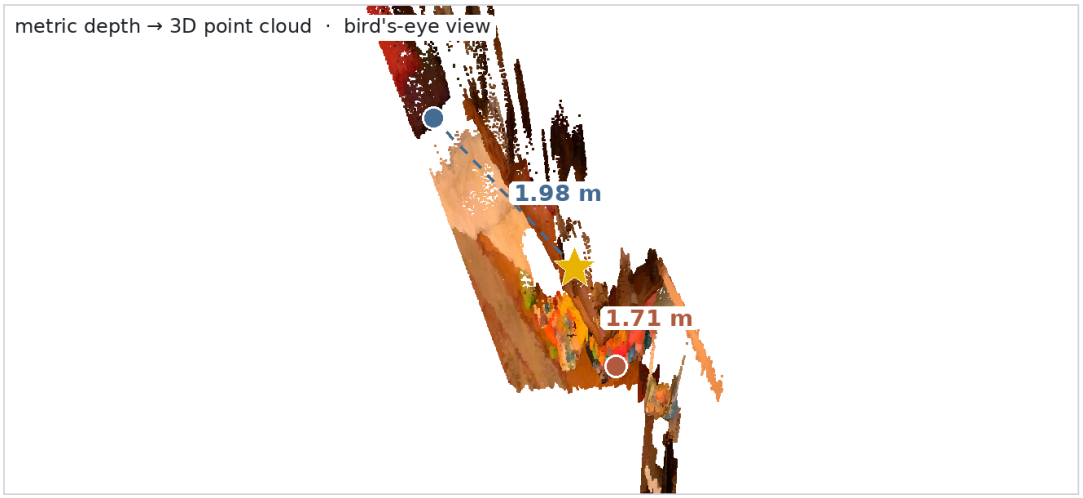


GEOUG: REASONING WITH GEOMETRY

VLM predictions follow image-plane cues

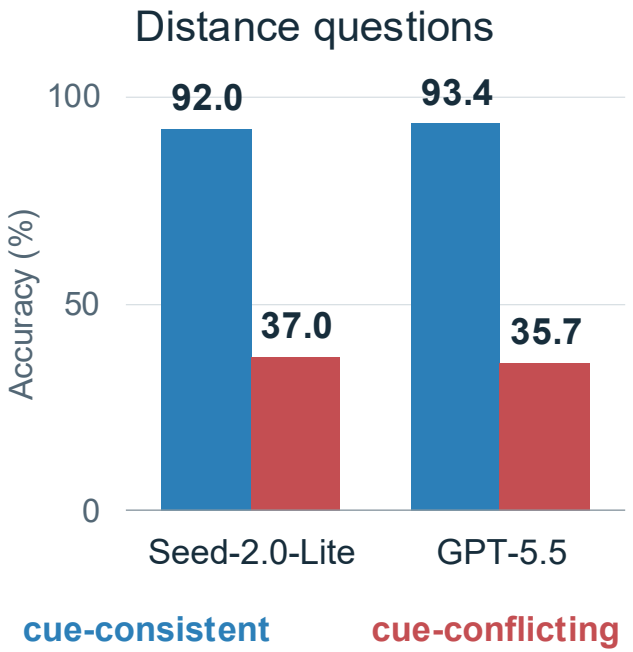
Which is closer to the flower arrangement: the stool or the toy basket?





~~2D cue "closer in the image plane" → stool~~

3D: the toy basket is closer (1.71 m vs 1.98 m)



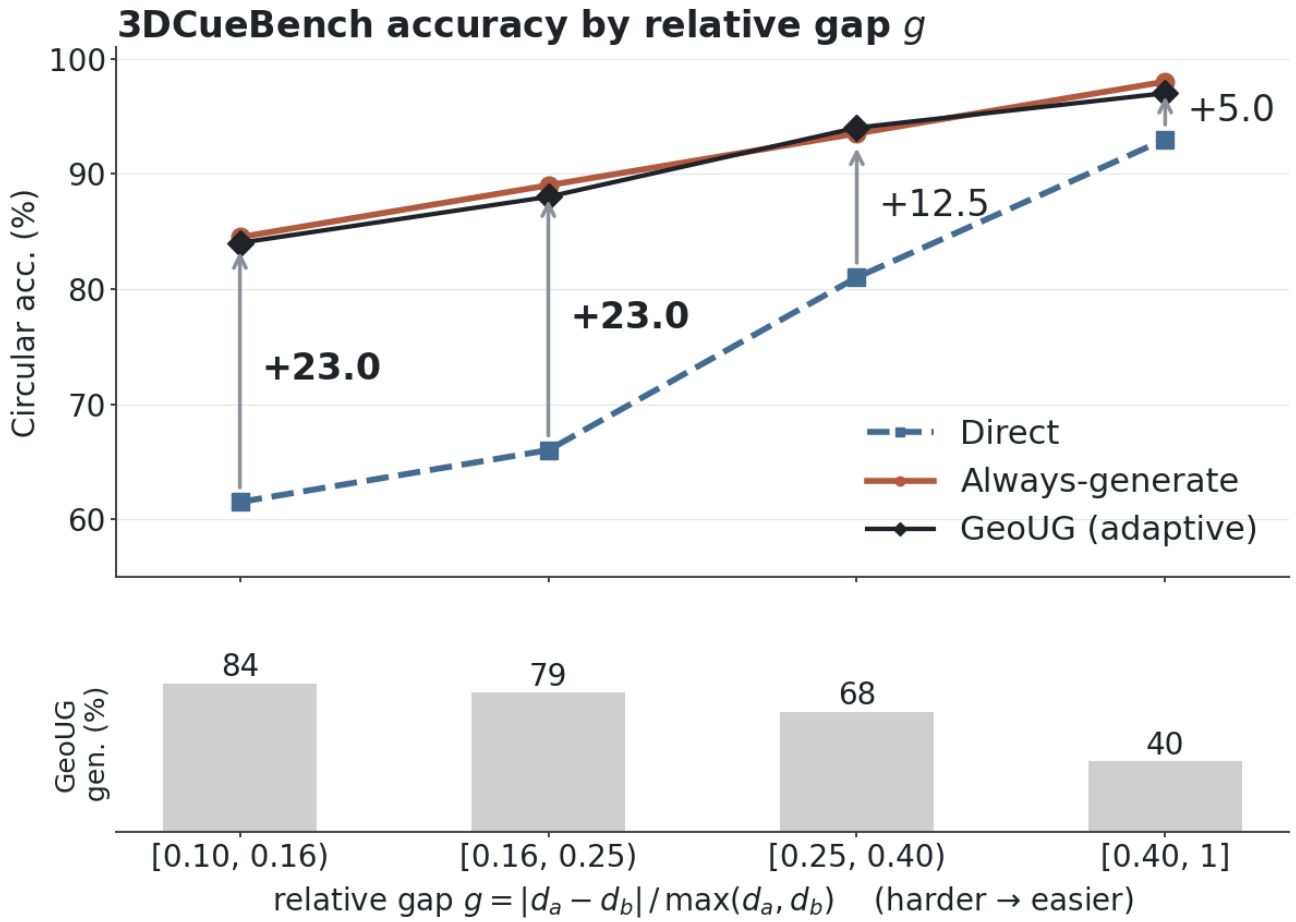
Cue-conflicting Distance questions

Seed-2.0-Lite: 37.0%

GPT-5.5: 35.7%

GEOUG: REASONING WITH GEOMETRY

When does generated depth help?



Generation helps most when the gap is small

+23 points on the two hardest bins, only +5 on the easiest

GeoUG learns to generate more often on hard questions (84% → 40%)

Accuracy curves: Table 12. Generation-rate bars: schematic.

GEOUG: REASONING WITH GEOMETRY

Draft answers inform the generation decision

Decision before answering



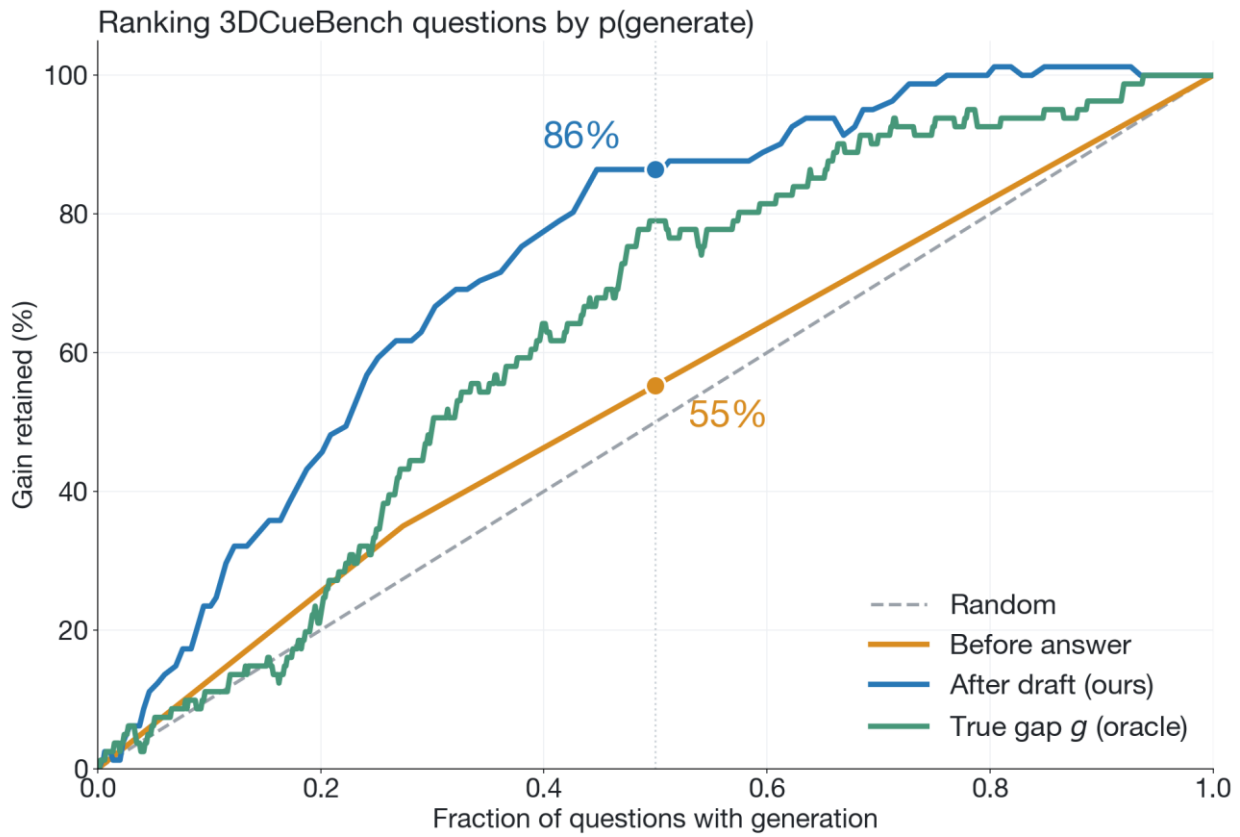
Decision after a draft answer



Draft answers provide a signal for selecting when to generate.

GEOUG: REASONING WITH GEOMETRY

Draft answers provide useful decision information



Generate for the top 50% ranked by generation probability

Before-answer policy

55% of the gain retained

Draft-conditioned policy

86% of the gain retained

Pilot: ordinary accuracy, normalized within each policy.

Gain retained = (selected accuracy − direct accuracy) / (always-generate accuracy − direct accuracy)

GEOUG: REASONING WITH GEOMETRY

Adaptive interleaved geometric reasoning

Which object is closer to the pianist's head?

A. blue light **B. folding chair**



Draft (no depth): ~~**A. blue light**~~ (looks nearer in the image)

<DECIDE> p(generate) = 0.87

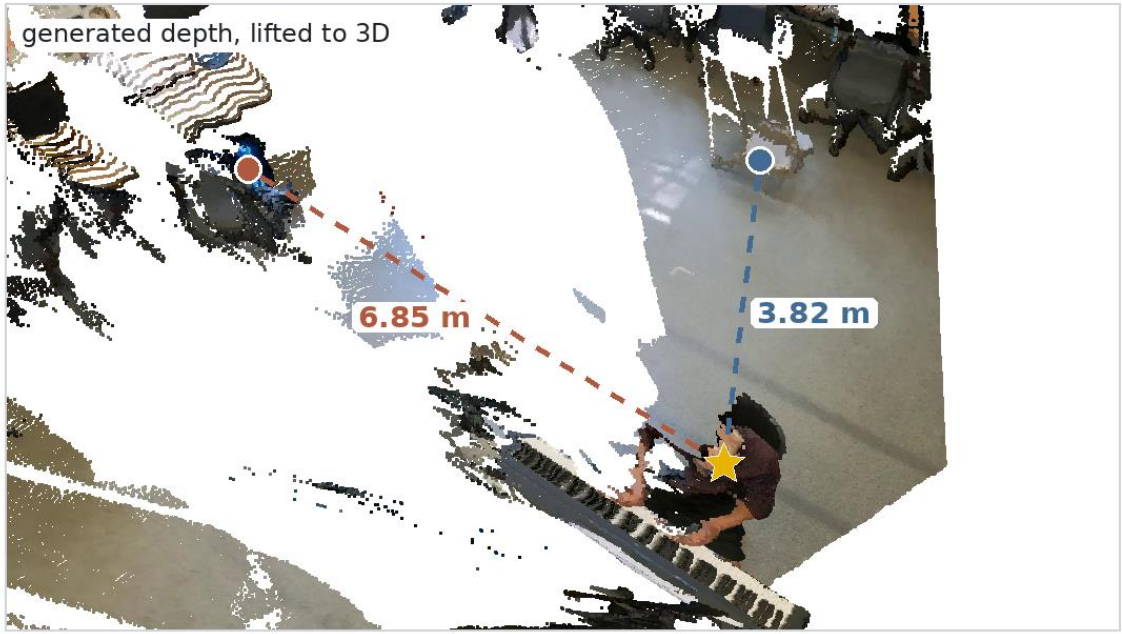
→ <GENERATE_DEPTH>

query_3d_distance(head, ·)

head ↔ folding chair **3.82 m**

head ↔ blue light **6.85 m**

Answer: **B. folding chair** (revised)



back-projected with the camera intrinsics; the tool reads 3D distances from it

GEOUG: REASONING WITH GEOMETRY

A shared generation cost disappears after centering

Illustrative all-generate group, generation cost = 0.1

Rollout	Correctness	Reward	Centered reward
1	1	0.9	+0.5
2	0	−0.1	−0.5
Group mean	0.5	0.4	0

Without a cost: [1, 0] − 0.5 = [+0.5, −0.5]

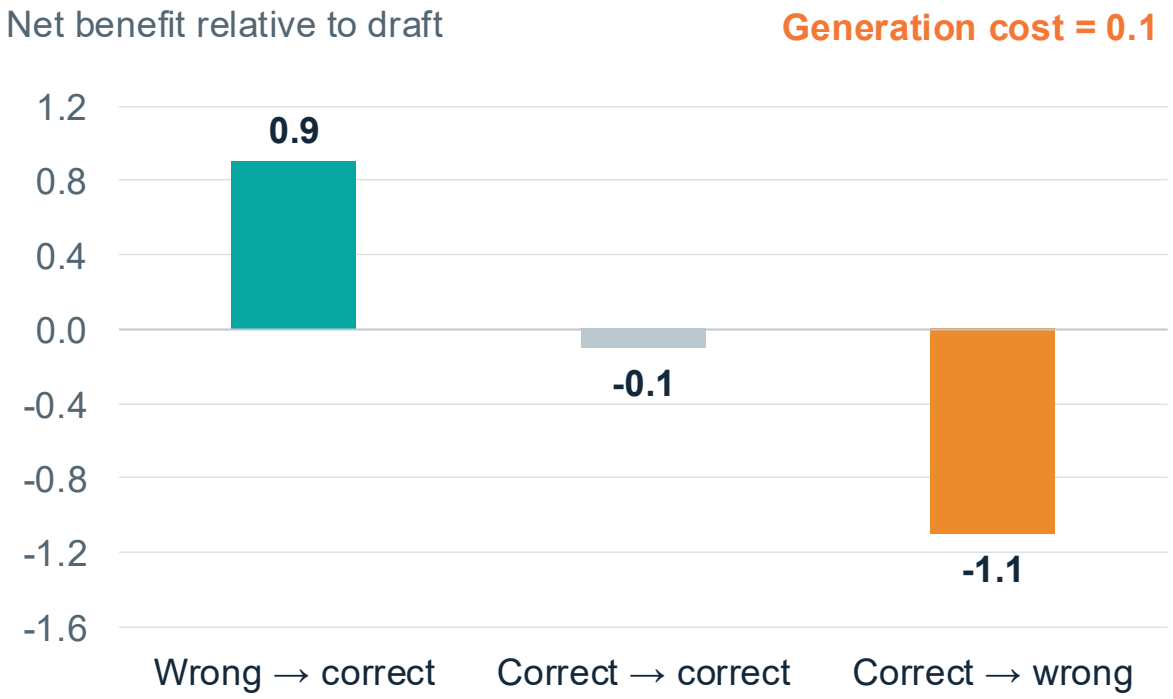
The cost leaves the centered rewards unchanged

When every rollout generates, the shared cost cancels from the centered reward.

GEOUG: REASONING WITH GEOMETRY

Generation value relative to the draft

$$\text{Net benefit} = \text{final correctness} - \text{draft correctness} - \text{generation cost}$$



Wrong → correct

The improvement outweighs the cost.

Correct → correct

No improvement, but the cost remains.

Correct → wrong

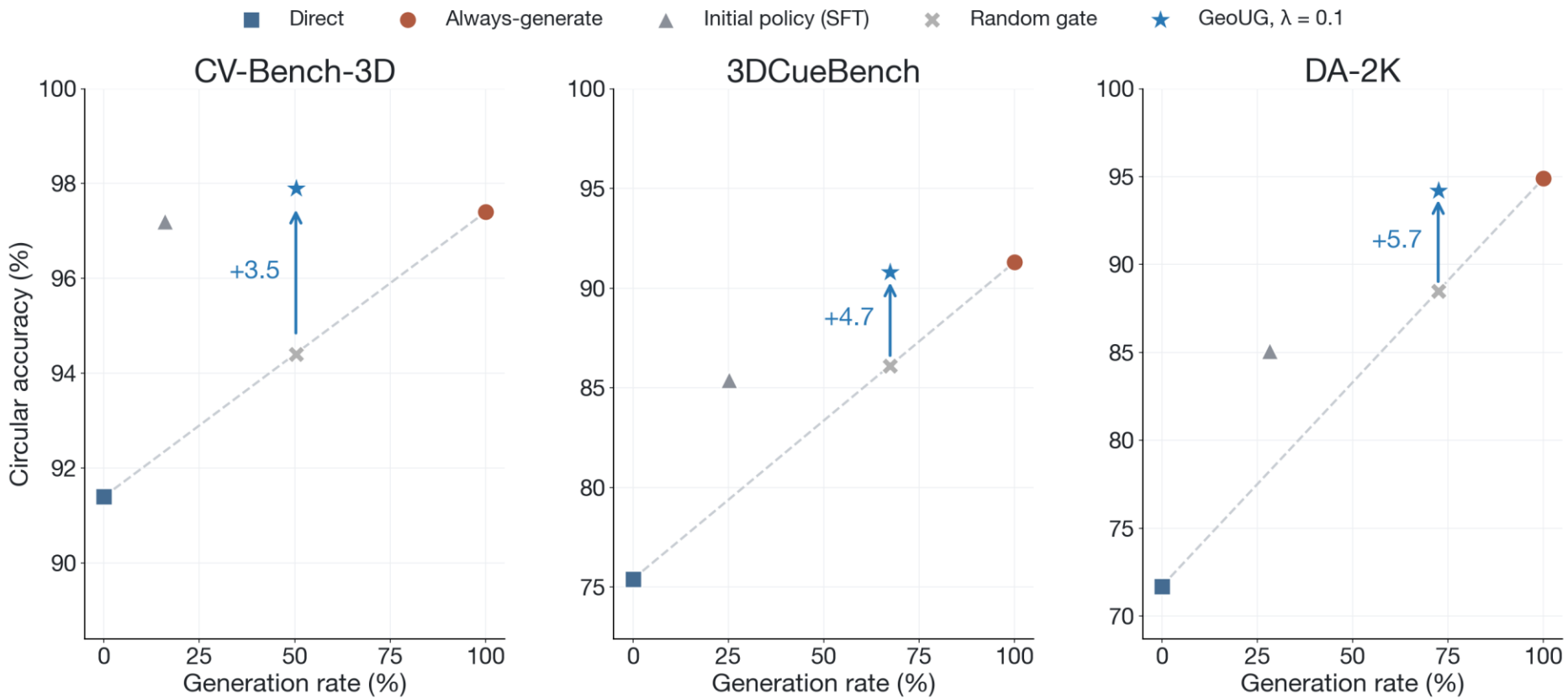
Accuracy loss adds to the cost.

Correct = 1; wrong = 0.

Illustrative training benefits. Decision advantages also include branch probabilities and a fixed scale.

GEOUG: REASONING WITH GEOMETRY

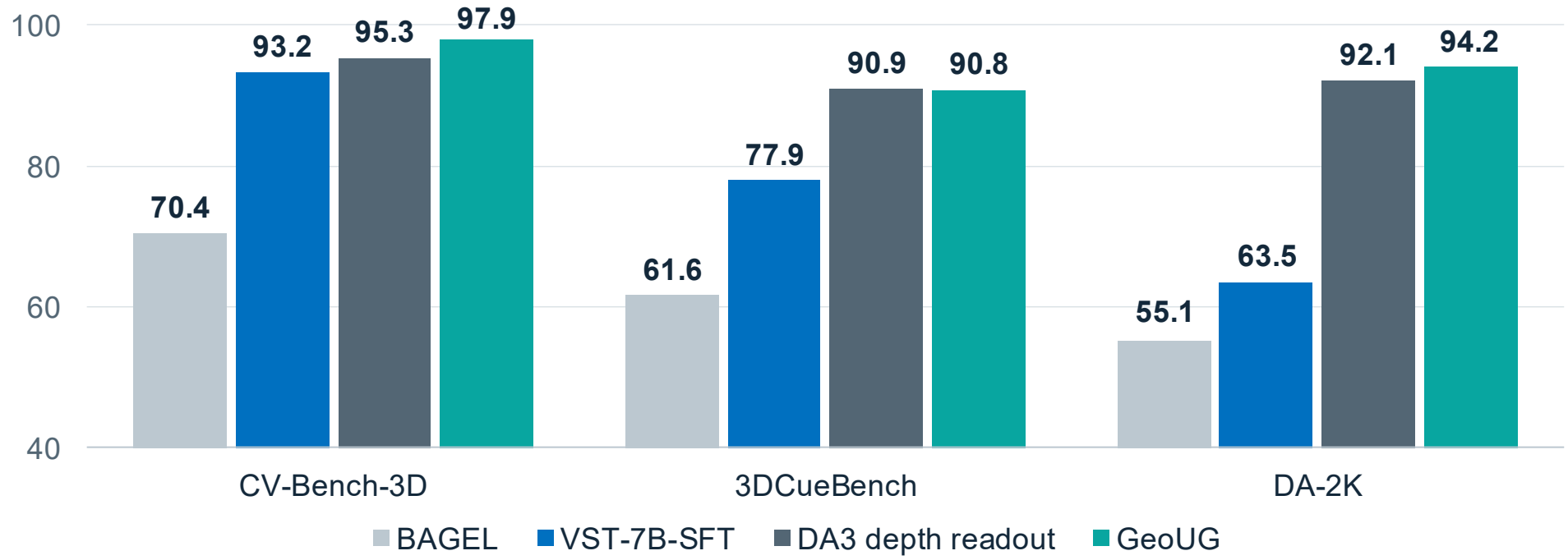
Selective generation preserves spatial accuracy



Compared with a random gate at the same generation rate: +3.5 / +4.7 / +5.7 points.
Accuracy remains within 0.7 points of always generating.

GEOUG RESULTS

Spatial accuracy with generated geometric evidence



Circular accuracy (%): both option orders must be correct.

DA3 uses the same depth-measurement procedure, without a VLM.

DISCUSSION

Generation as supervision and geometric evidence

	SGT	GeoUG
Training target	Segmentation targets	Metric depth and new views
Learning benefit	Visual perception and layout fidelity	Task-dependent spatial transfer
Inference role	Direct understanding and image generation	Selective depth generation and measurement

Generated evidence can be wrong

How should a model verify its generated evidence before revising an answer?

Questions and discussion

Email: liyanwei@sjtu.edu.cn

TRAINING AND EVALUATION

The test spans two architectures and both capabilities

190k

SGT samples from SAM

500k

LLaVA-OneVision SFT samples

2

Distinct UMM architectures

15

Understanding and generation evaluations

BAGEL

Mixture of Transformers

7B understanding + 7B generation

OmniGen2

VLM-guided diffusion

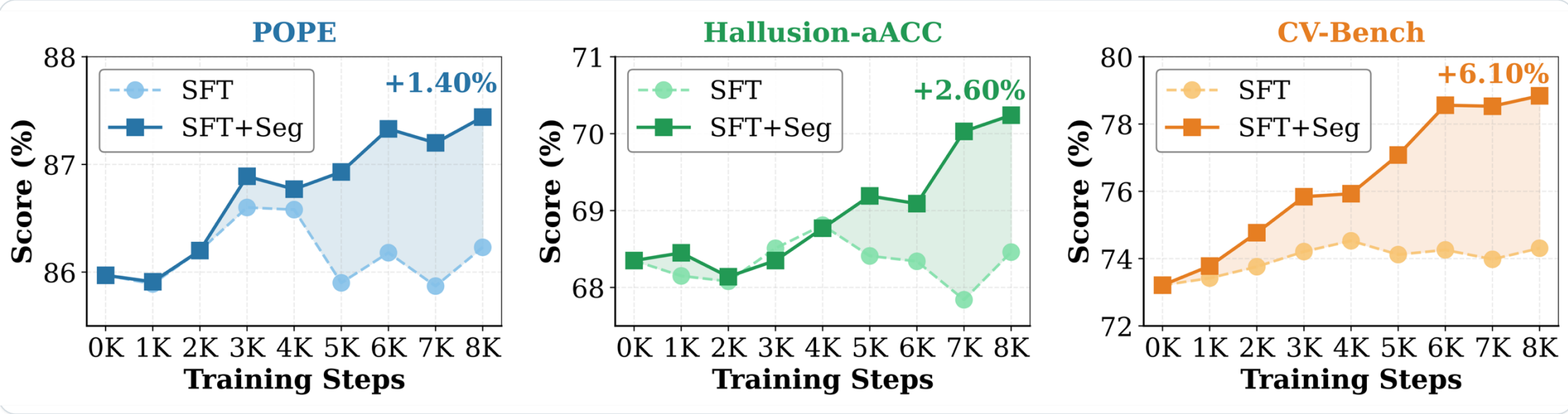
3B understanding + 4B generation

Consistent gains across two distinct UMM designs

TRAINING DYNAMICS

SGT improves learning trajectories on visual benchmarks

BAGEL: VQA-only SFT vs. SFT + SGT (VQA : segmentation = 1:2)



Higher POPE and Hallusion scores during training

The largest CV-Bench gains emerge later in optimization

MECHANISTIC ANALYSIS

What changes inside the model?

1

Features

Clearer separation in feature projections

2

Understanding

More attention to visual evidence

3

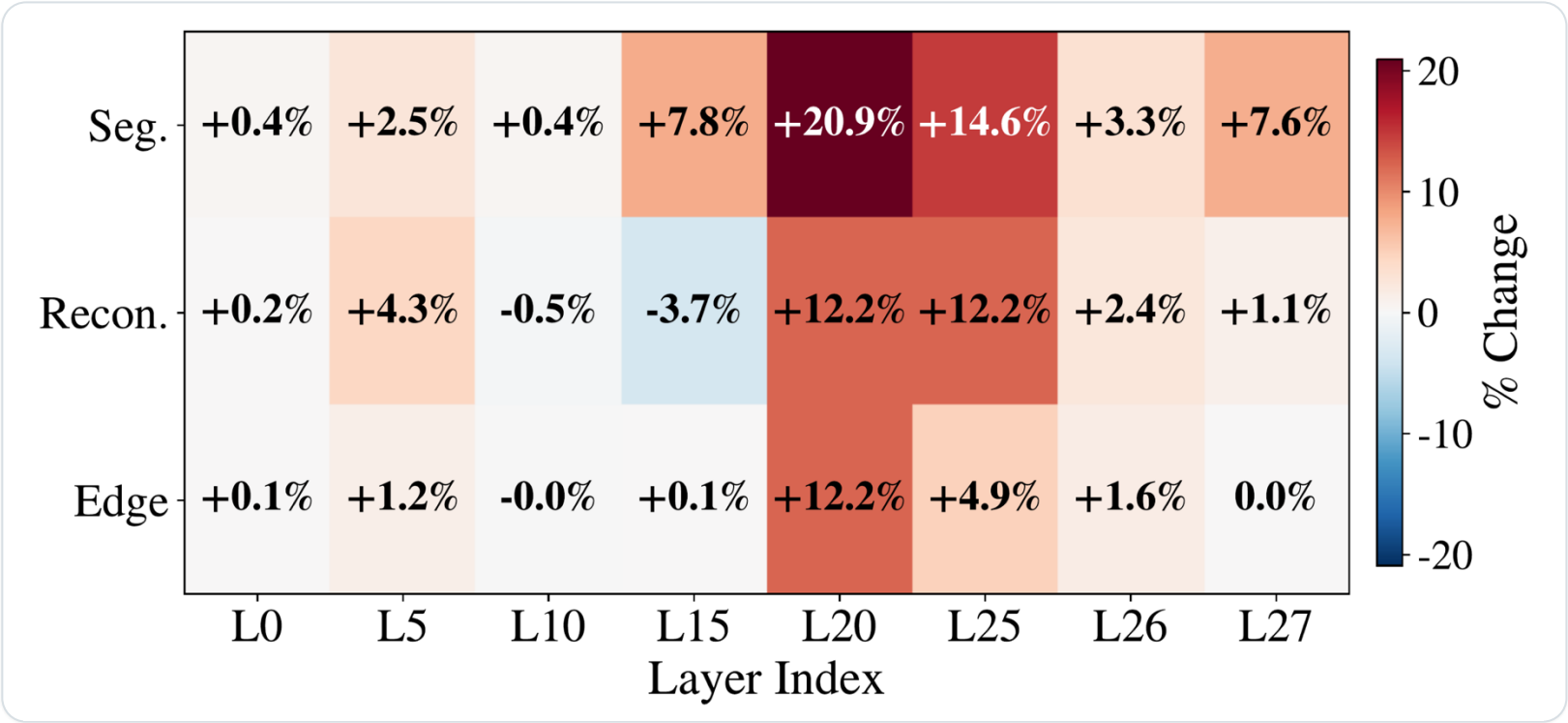
Generation

More attention to composition tokens

Evidence from feature geometry and attention patterns in BAGEL

VISUAL ATTENTION

Segmentation increases visual attention in deeper layers



Relative to BAGEL

+20.9%

Visual attention at L20

+14.6%

Visual attention at L25

Segmentation shows the largest visual-attention shift in this comparison

BACKUP: FULL RESULTS

Unified benchmark results for BAGEL and OmniGen2

OmniGen2											
Method	**CV-Bench**	**MMVP**	**VSR**	**SIBench**	**POPE**	**Hallusion**	**MMBench**	**MMMU**	**MMStar**	**GenEval**	**GEdit (En)**
Base	65.94	65.00	77.52	43.29	85.97	62.35	77.04	42.11	55.07	76.58	6.63
SFT	65.99	66.00	77.61	44.37	86.25	64.35	77.04	43.22	55.73	74.54	6.32
SFT + Edge	66.67	65.33	77.99	**45.51**	86.10	63.72	76.88	42.78	55.53	77.45	6.79
SFT + Recon.	66.71	66.33	78.18	45.41	85.92	**65.19**	77.00	44.44	55.73	77.53	6.81
SFT + SGT	**66.91**	**68.33**	**78.85**	45.37	**87.29**	64.25	**77.09**	**45.89**	**57.07**	**78.86**	**6.83**
BAGEL											
Method	**CV-Bench**	**MMVP**	**VSR**	**SIBench**	**POPE**	**Hallusion**	**MMBench**	**MMMU**	**MMStar**	**GenEval**	**GEdit (En)**
Base	73.21	83.00	80.45	48.95	85.69	68.34	81.25	46.77	67.46	78.21	6.52
SFT	74.61	82.67	80.69	49.34	86.77	67.92	81.86	47.33	66.93	77.18	6.49
SFT + Edge	74.56	**83.67**	80.83	49.51	86.48	68.66	81.89	47.56	67.20	79.96	6.72
SFT + Recon.	75.23	83.33	80.83	**50.59**	87.98	68.03	82.18	46.56	67.40	80.82	6.75
SFT + SGT	**79.23**	83.33	**81.54**	50.18	**88.32**	**70.24**	**83.84**	**48.56**	**68.33**	**80.95**	**6.94**

GenEval: average over 12 seeds. BAGEL + SGT scores 80.95 here. The 90.0 result uses a prompt rewriter (Table 2).

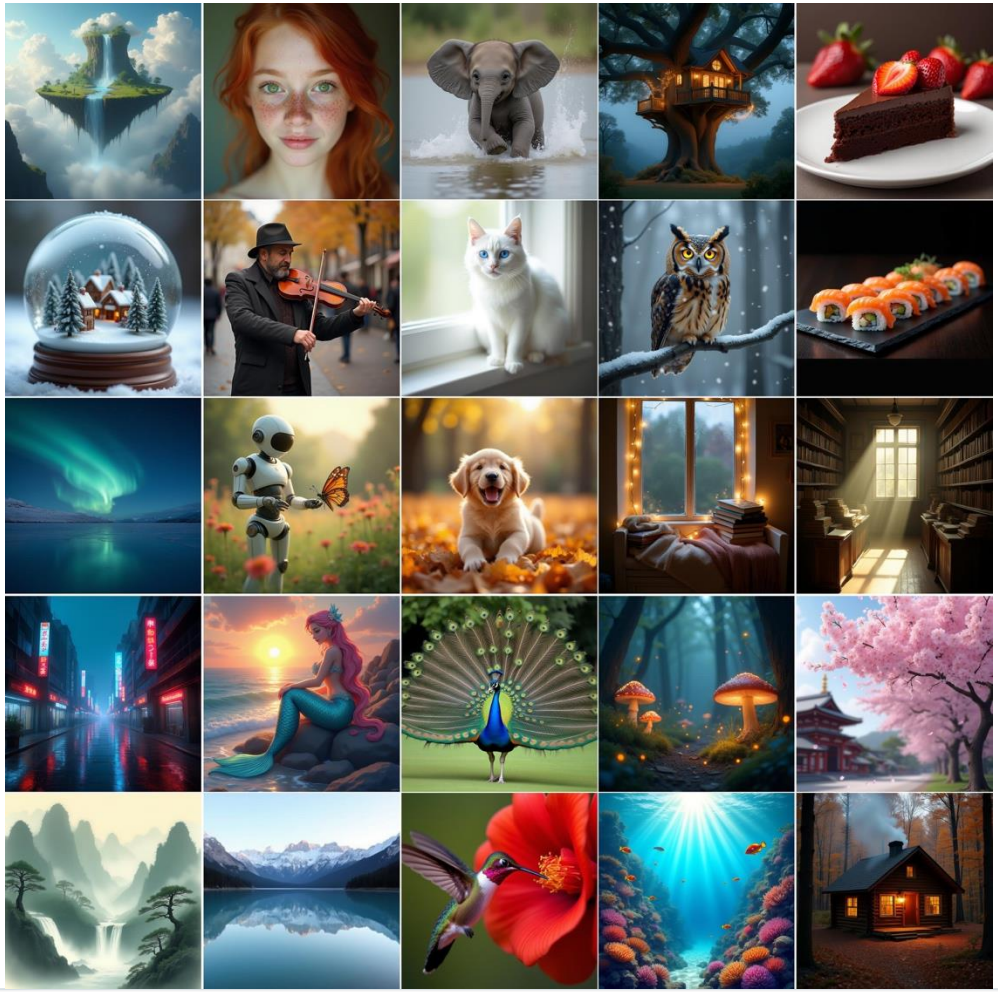
BACKUP: QUALITATIVE EXAMPLES

Qualitative range of SGT generations

Examples across subjects and visual styles

People, animals, objects, and scenes
Photographic and stylized outputs

Qualitative illustration
No formal diversity metric



BACKUP: GEOUG

3DCueBench cases: matched source and gap bin

Model input an unmarked image and this text (Depth shown; Distance asks “Which is closer to ⟨anchor⟩?” with the anchor's point)

Consider the relationship in the 3D world. Which is closer to the camera?

(A) the baseball player (represented by the point <point>445 352</point>)

(B) the spectator (represented by the point <point>571 120</point>)

Depth WildDet3D · g 0.16–0.25

rule: lower in the image is closer



A baseball player
B spectator

2D rule picks A ✓



A toothbrush
B sink faucet

2D rule picks B ✗

Distance Hypersim · g ≥ 0.40

rule: closer to ★ in the image is closer



A hanging lamp
B folded clothes

2D rule picks B ✓



A blue jacket
B standing mirror

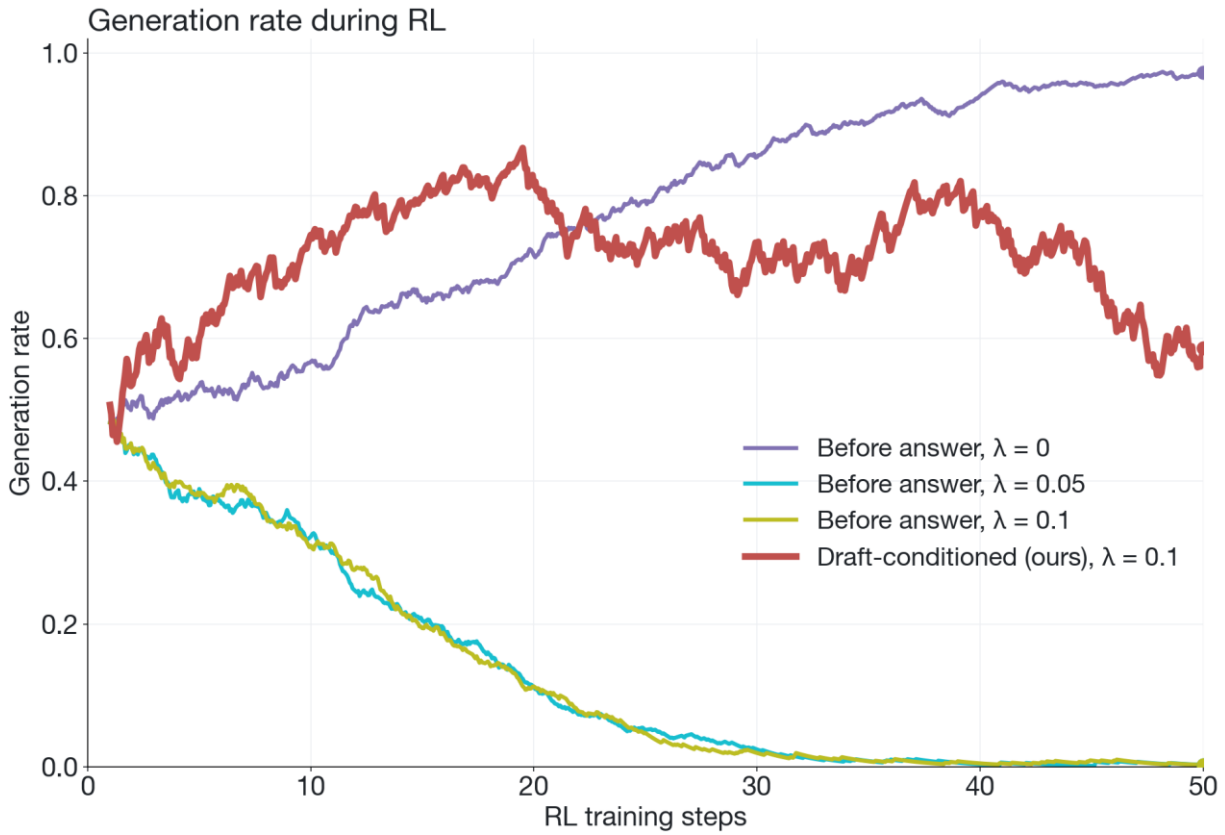
2D rule picks B ✗

Pairs match source and gap bin, with opposite cue validity.

Each selected 2D rule scores 50% across the benchmark.

BACKUP: GEOUG

Draft-conditioned RL learns selective generation



Before-answer policy in these runs

λ = 0: almost always generate
Tested λ > 0: almost never generate

Draft-conditioned policy remains selective at λ = 0.1

These runs also differ in model, training questions and credit assignment.

BACKUP: GEOUG

Why group-relative credit loses the cost signal

Reward = answer correctness − generation cost

$$r_i = c_i - \lambda a_i$$

$c_i \in \{0, 1\}$: correctness $a_i \in \{0, 1\}$: generation decision

When every rollout generates, $a_i = 1$ for all i

$$r_i - \bar{r} = (c_i - \lambda) - (\bar{c} - \lambda) = c_i - \bar{c}$$

The shared cost λ cancels from the advantage.

The decision receives no cost-dependent signal for whether generation is worthwhile.

BACKUP: GEOUG

Draft-relative credit preserves benefit and cost

Compare the expected benefit of generation with the initial answer.

Wrong draft → correct answer

$$1 - 0.1 - 0 = +0.9$$

Correct draft → correct answer

$$1 - 0.1 - 1 = -0.1$$

Decision advantages before scaling

Generate $$A_{\mathrm{i}} = (1 - p_{\mathrm{i}})\,(c_{\mathrm{i}} - \lambda - c_{\mathrm{d,i}})$$

Answer directly $$A_{\mathrm{i}} = -p_{\mathrm{i}}\,(Q_{\mathrm{gen,i}} - \lambda - c_{\mathrm{d,i}})$$

p_{i}: generation probability $c_{\mathrm{d,i}}$: draft correctness

$Q_{\mathrm{gen,i}}$: correctness estimate from forced-generation rollouts

All advantages are divided by 0.5.

BACKUP: GEOUG

Depth and distance understanding

Model	CVB Depth	CVB Dist.	CVB All	3DCue Depth	3DCue Dist.	3DCue All	DA-2K	Avg.
BAGEL-7B-MoT	81.5	59.3	70.4	73.8	49.5	61.6	55.1	62.4
Qwen2.5-VL-7B	80.2	61.5	70.8	61.8	35.5	48.6	45.3	54.9
Qwen3-VL-8B	93.8	80.3	87.1	80.0	53.5	66.8	44.6	66.2
InternVL3.5-8B	83.2	74.7	78.9	62.8	52.5	57.6	31.6	56.0
SenseNova-SI-1.4-8B	93.5	85.2	89.3	81.3	63.0	72.1	63.9	75.1
SenseNova-SI-1.5-8B	90.8	84.0	87.4	79.0	64.0	71.5	61.1	73.3
VST-7B-SFT	95.5	90.8	93.2	83.8	72.0	77.9	63.5	78.2
DA3 (depth readout)	96.0	94.5	95.3	**95.3**	**86.5**	**90.9**	92.1	92.8
GeoUG (ours)	**98.5**	**97.3**	**97.9**	**95.0**	**86.5**	**90.8**	**94.2**	**94.3**

Circular accuracy (%): both option orders must be correct. CVB = CV-Bench-3D.

GeoUG answers 32.6% of 3DCueBench questions directly, without depth generation.

BACKUP: GEOUG

Ablation of decision credit assignment

Decision advantage	CV-Bench-3D	3DCueBench	DA-2K	Avg. acc.
Always-generate	97.4 (100)	91.3 (100)	94.9 (100)	94.5
Initial policy (SFT)	97.2 (16.0)	85.4 (25.2)	85.1 (28.3)	89.2
Trajectory-level advantage	96.6 (44.1)	90.8 (66.2)	94.1 (71.1)	93.8
Draft baseline, sampled action	97.6 (41.6)	89.6 (62.8)	92.8 (65.6)	93.4
Two-branch, unforced	97.3 (29.2)	89.5 (49.3)	92.5 (53.0)	93.1
Ours: two-branch + forced rollouts	**97.6 (50.7)**	**91.0 (69.3)**	**93.9 (73.4)**	**94.2**

Circular accuracy (%) with generation rate (%) in parentheses. $\lambda = 0.1$; 20 RL updates from the same initial policy.

Without forced generation, 4% of training questions have no generated rollout.

Forced rollouts provide a generation-value estimate for the decision update.