ECCV 2026 | MALMÖ | SEPT 8–12

# Semantic Generative Tuning for Unified Multimodal Models

Align understanding and generation through semantic visual targets

**Songsong Yu** **Yuxin Chen** **Ying Shan** **Yanwei Li**

Shanghai Jiao Tong University Tencent

# MOTIVATION

***One architecture can still learn two disconnected spaces***

## Understanding

### Sparse text supervision

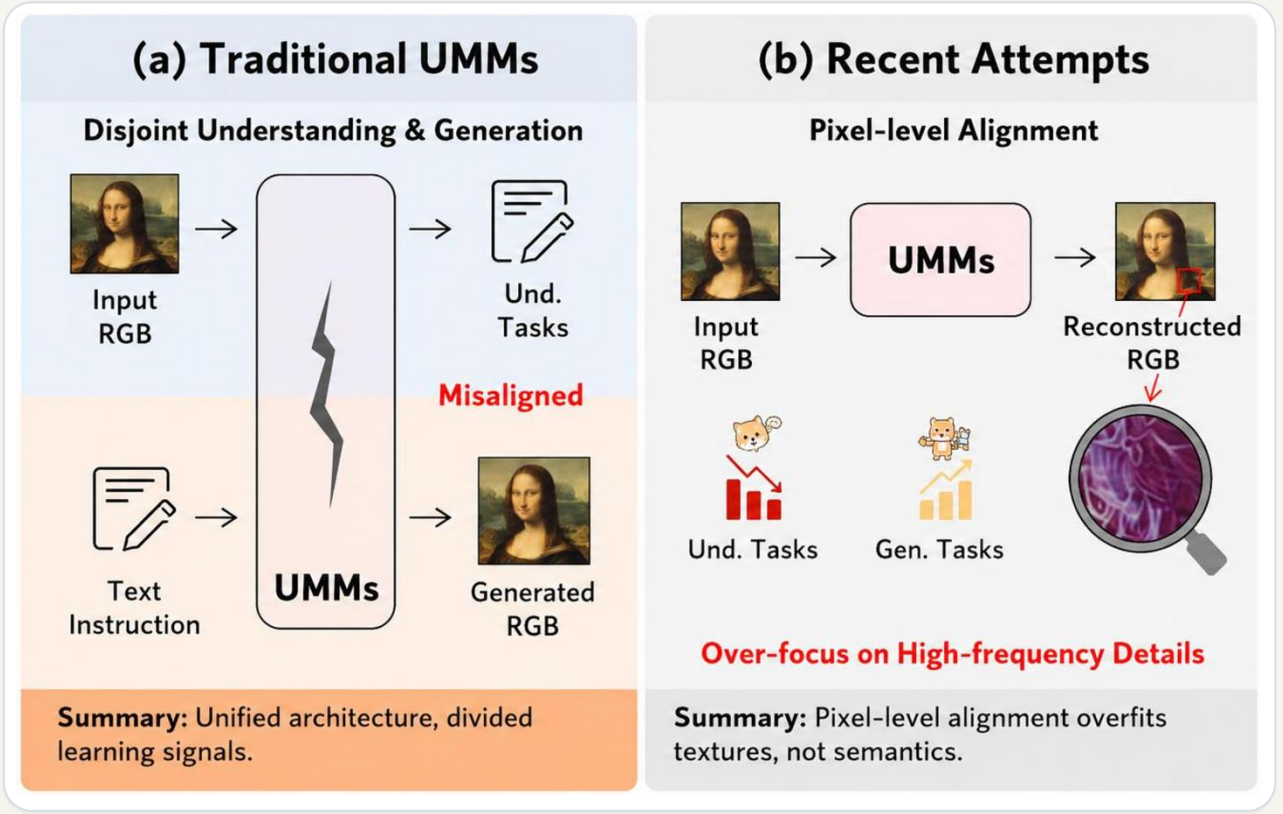
VQA teaches what to say about an image.

## Generation

### Dense visual supervision

Pixel objectives teach how to reproduce appearance.

**The bottleneck is objective alignment.**

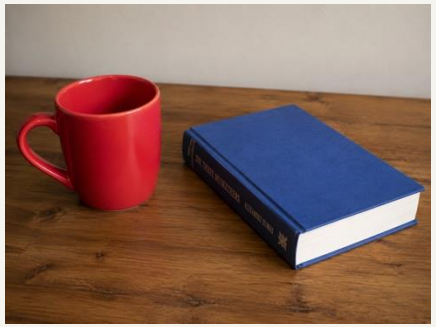


# MOTIVATION

***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?

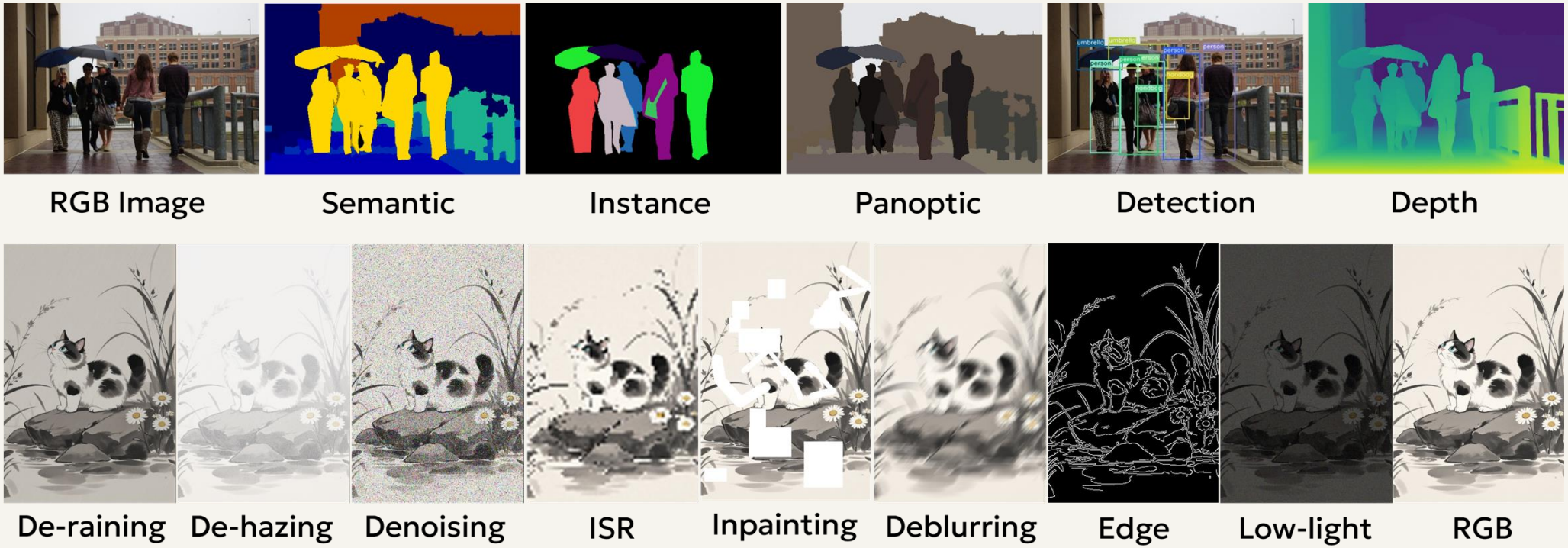
**SHARED VISUAL SIGNAL**

**Semantics**
**Structure and layout**

**A useful proxy should preserve the structure that both tasks can exploit.**

# QUESTION

***Which visual target provides the right level of abstraction?***

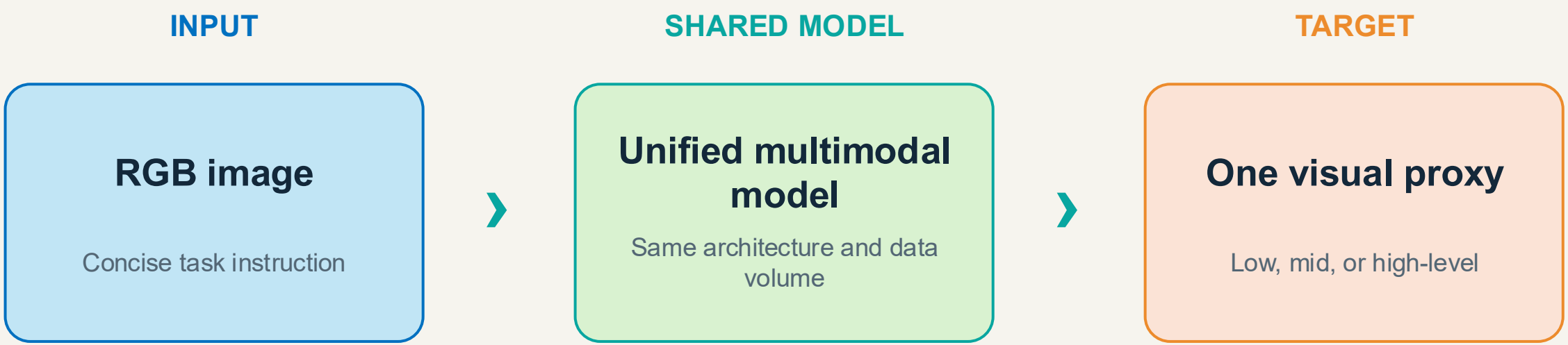



**Texture** → **Structure** → **Semantics**

**The proxy must filter texture without discarding spatial structure**

## CONTROLLED STUDY

***Matched data and architecture isolate the effect of the visual target***

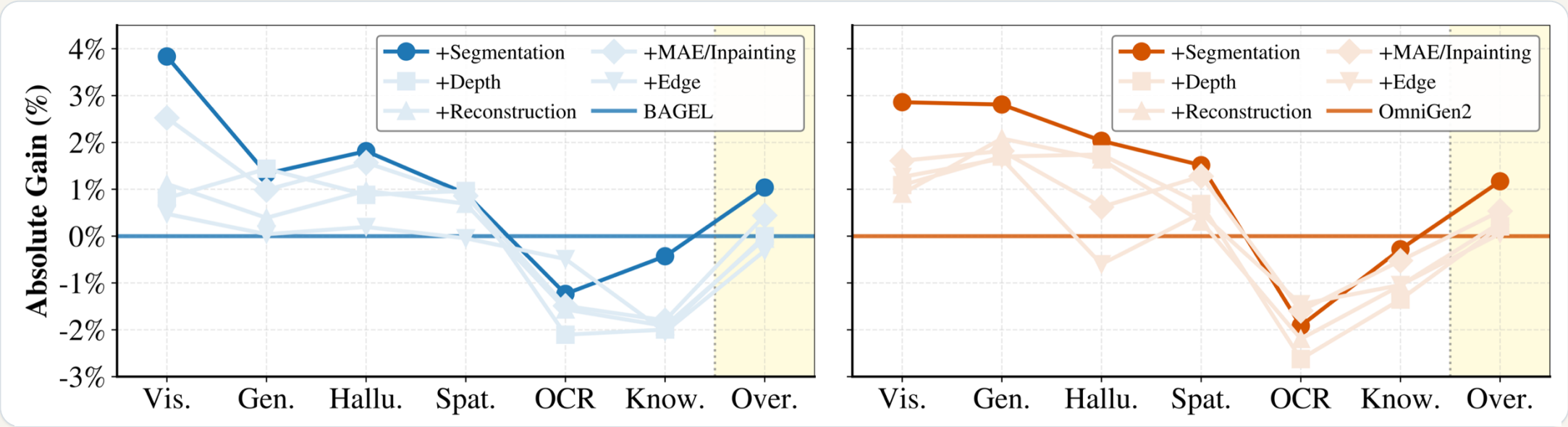



**No VQA data, text-to-image data, and image editing data in the proxy study.**

**This isolates the effect of proxy granularity**

# OBSERVATION

## *High-level proxies deliver the largest understanding gains*




### Segmentation leads

### Largest gains

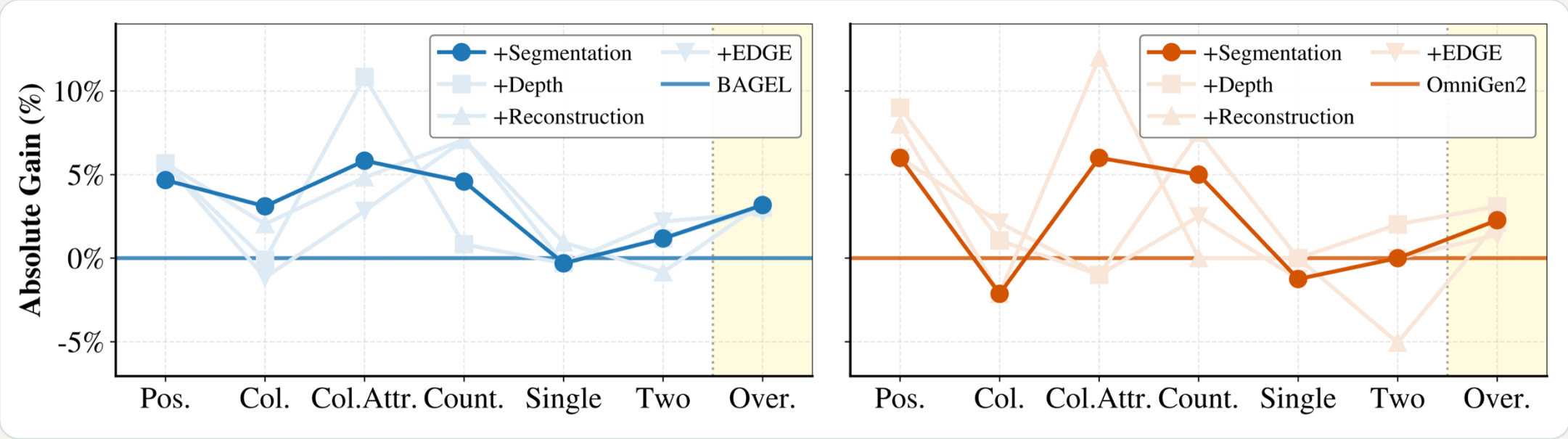
**Vision-centric perception, spatial reasoning, hallucination resistance**

### Limited gains

**OCR, document tasks, external knowledge**

# OBSERVATION

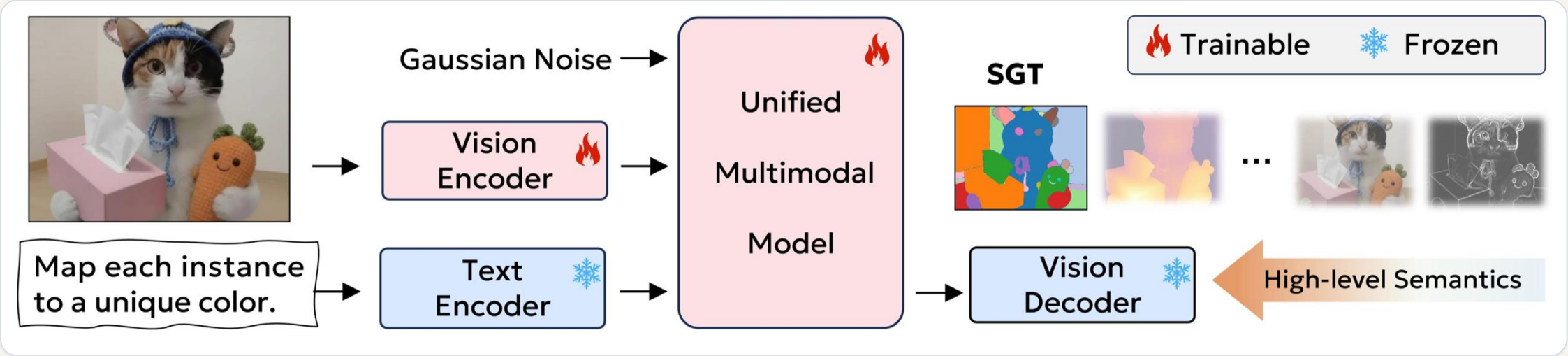
## *Every proxy improves position-aware generation*




## *The common effect*

**Visual targets force the model to preserve layout constraints**

## 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.

# 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

**Frozen VLM guides diffusion**

3B understanding + 4B generation

**Consistent gains across two distinct UMM designs**

# MAIN RESULTS

***SGT improves understanding and generation on both UMMs***

## UNDERSTANDING

**+6.0**

### CV-Bench

BAGEL: 73.2 -> +SGT: 79.2

### +3.3 MMVP on OmniGen2

## GENERATION

**90.0**

### GenEval

BAGEL: 88.0 -> +SGT: 90.0

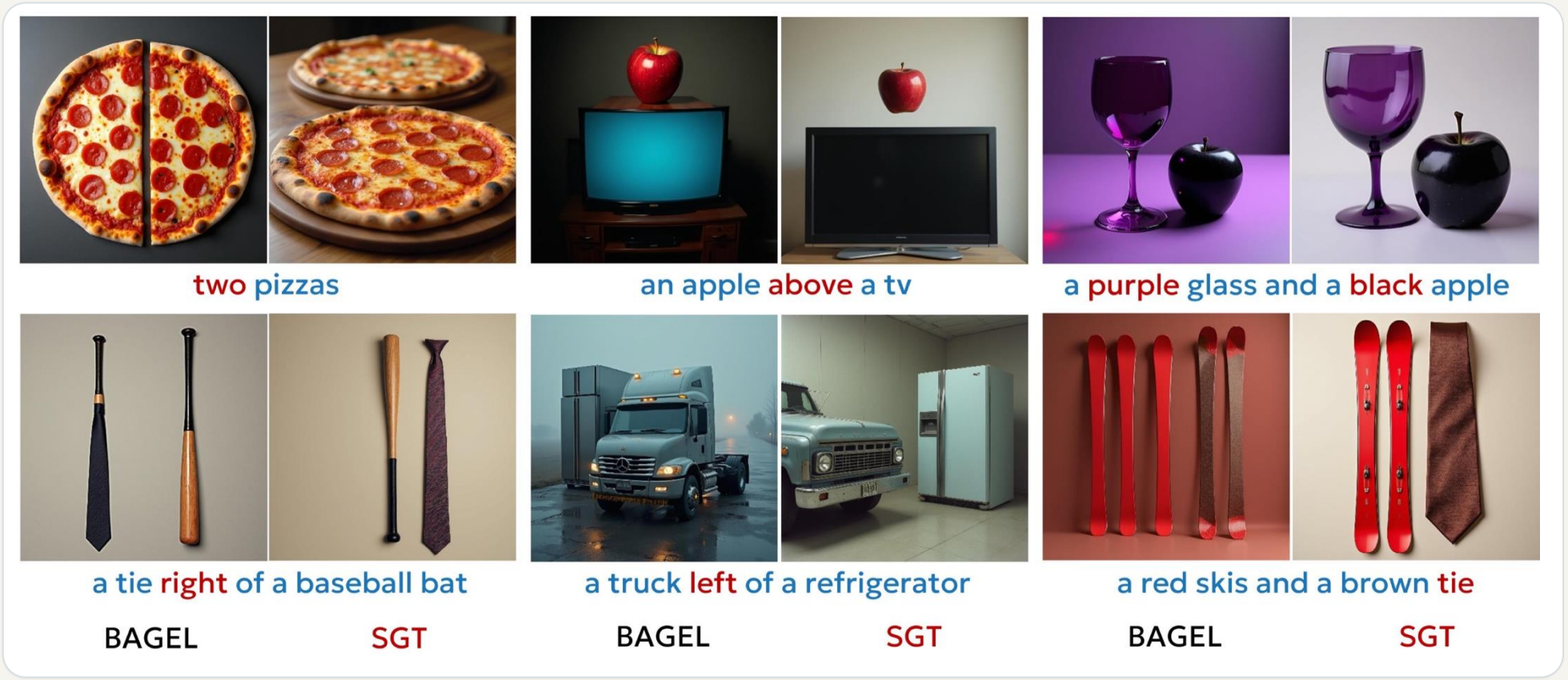
### +2.3 GenEval on OmniGen2

BAGEL GenEval scores use an LLM prompt rewriter (Table 2).

**Segmentation provides the strongest joint gain across both architectures**

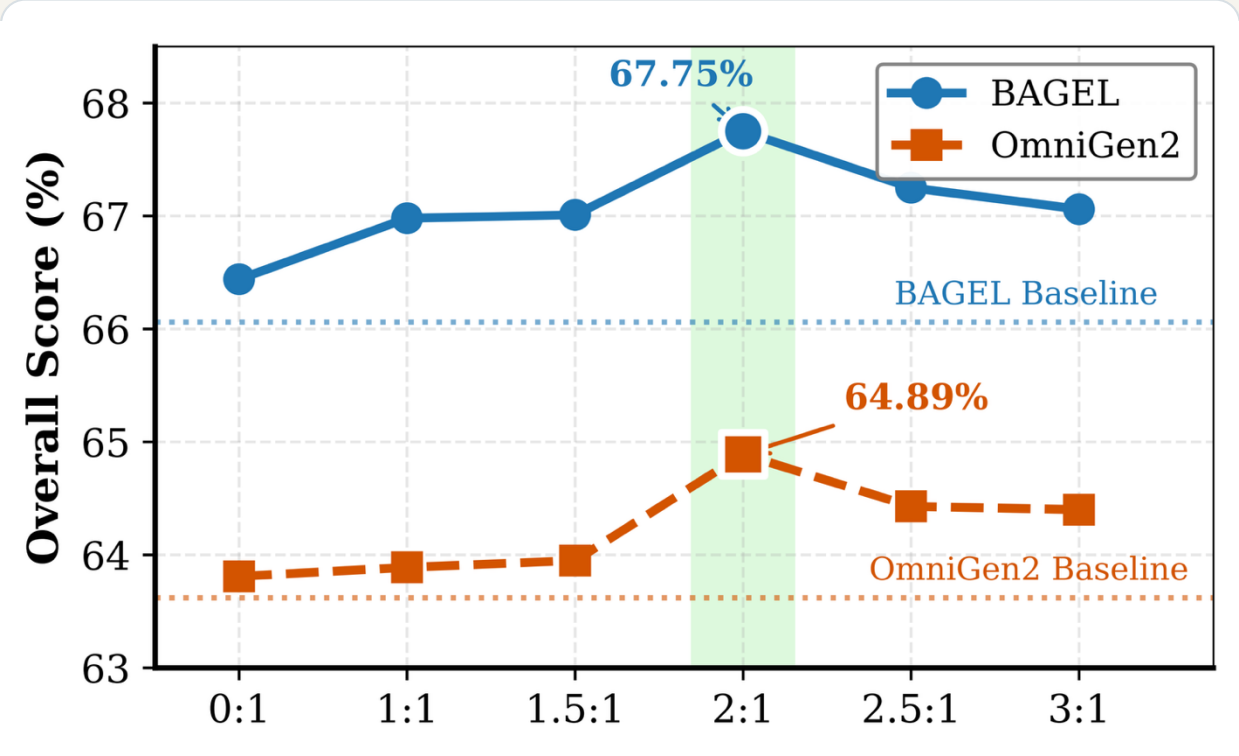
# EXAMPLES

## *SGT follows compositional prompts more faithfully*




Counting | Spatial relation | Object attributes

## A 2:1 segmentation-to-VQA ratio balances both objectives




**2:1**

**Segmentation : VQA**

Sampling ratio within each batch

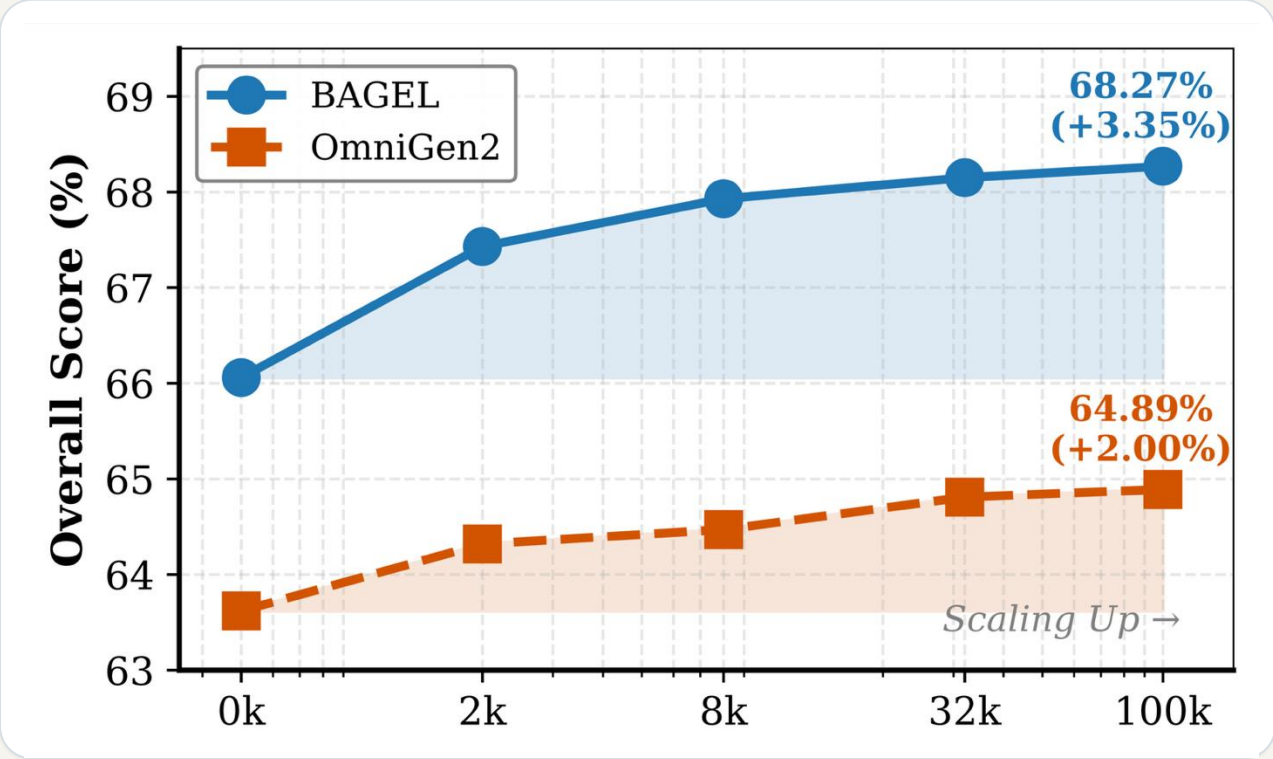
**Both models peak at 2:1**

Higher segmentation proportions continue to favor generation

**The final recipe balances both capabilities**

# DATA SCALING

## *Understanding improves as segmentation data scales*




**2k to 100k**

**Segmentation examples**

VQA SFT data held fixed

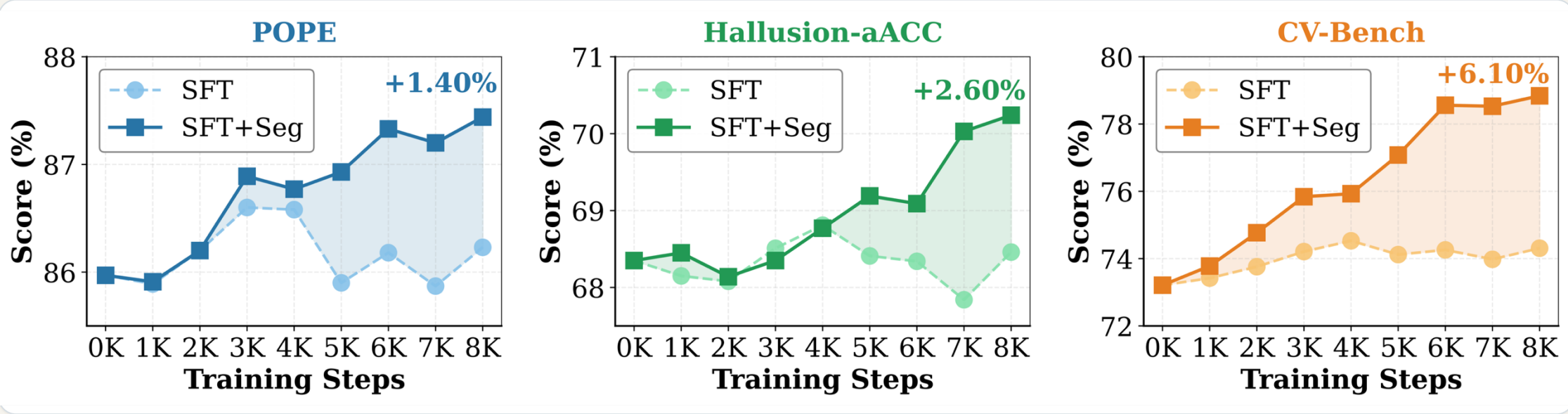
**Monotonic gains on both models**

Gains flatten toward the high-data end

**Average normalized score across eight understanding benchmarks**

## 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

## What changes inside the model?

1

### Features

More separable visual representations

2

### Understanding

More attention to visual evidence

3

### Generation

More attention to composition tokens

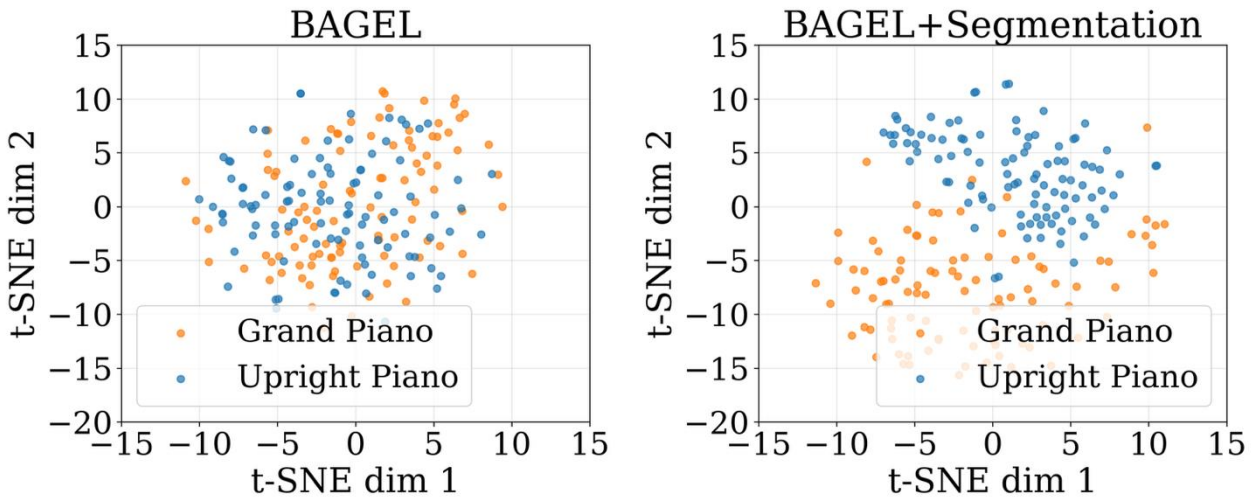
**Evidence from feature geometry and attention patterns in BAGEL**

# MECHANISM 1

## *SGT separates visually confusable categories in feature space*

### Confusable categories


### Visual feature embeddings




Qualitative t-SNE evidence

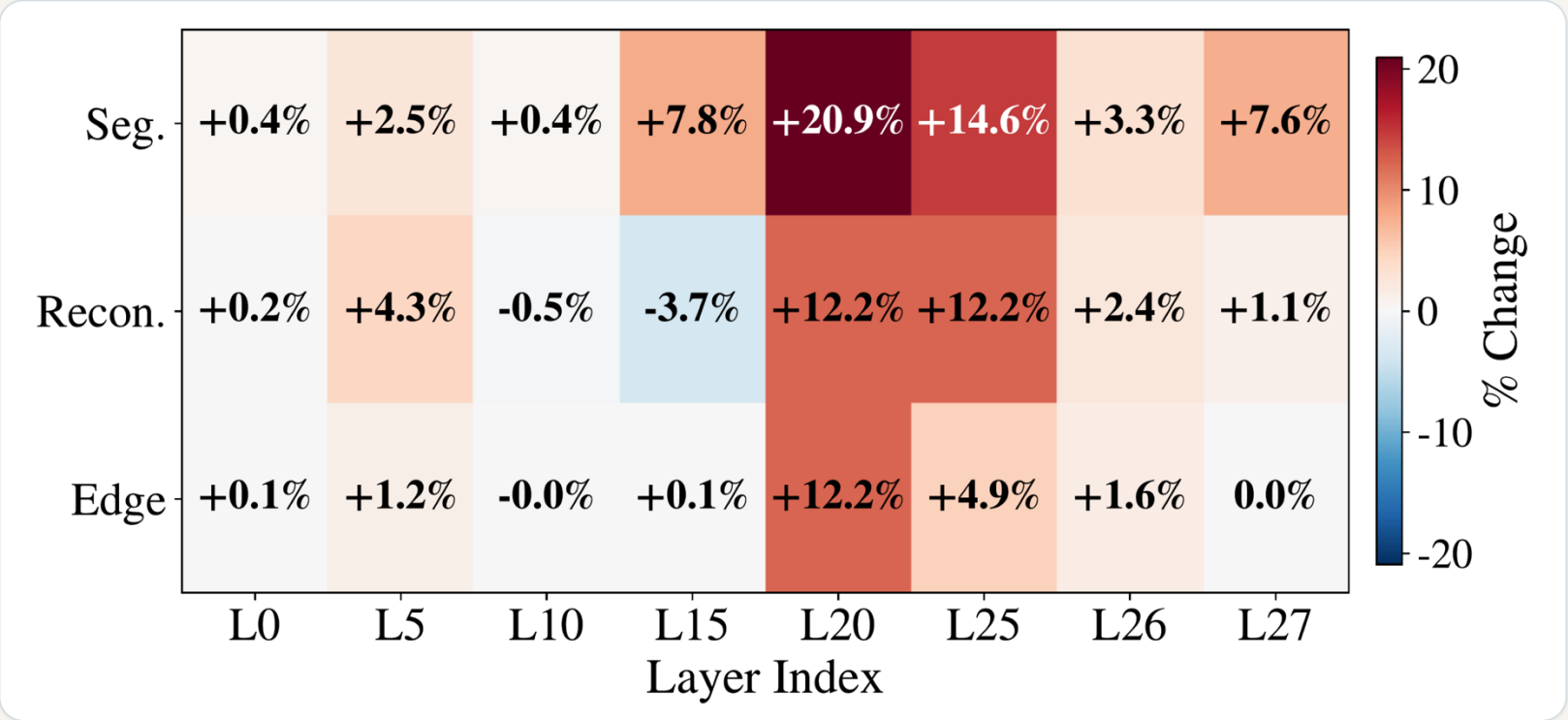
**BAGEL**

Overlapping clusters

**BAGEL + SGT**

**Clearer class separation**

## Segmentation increases visual attention in deeper layers




Relative to BAGEL

**+20.9%**

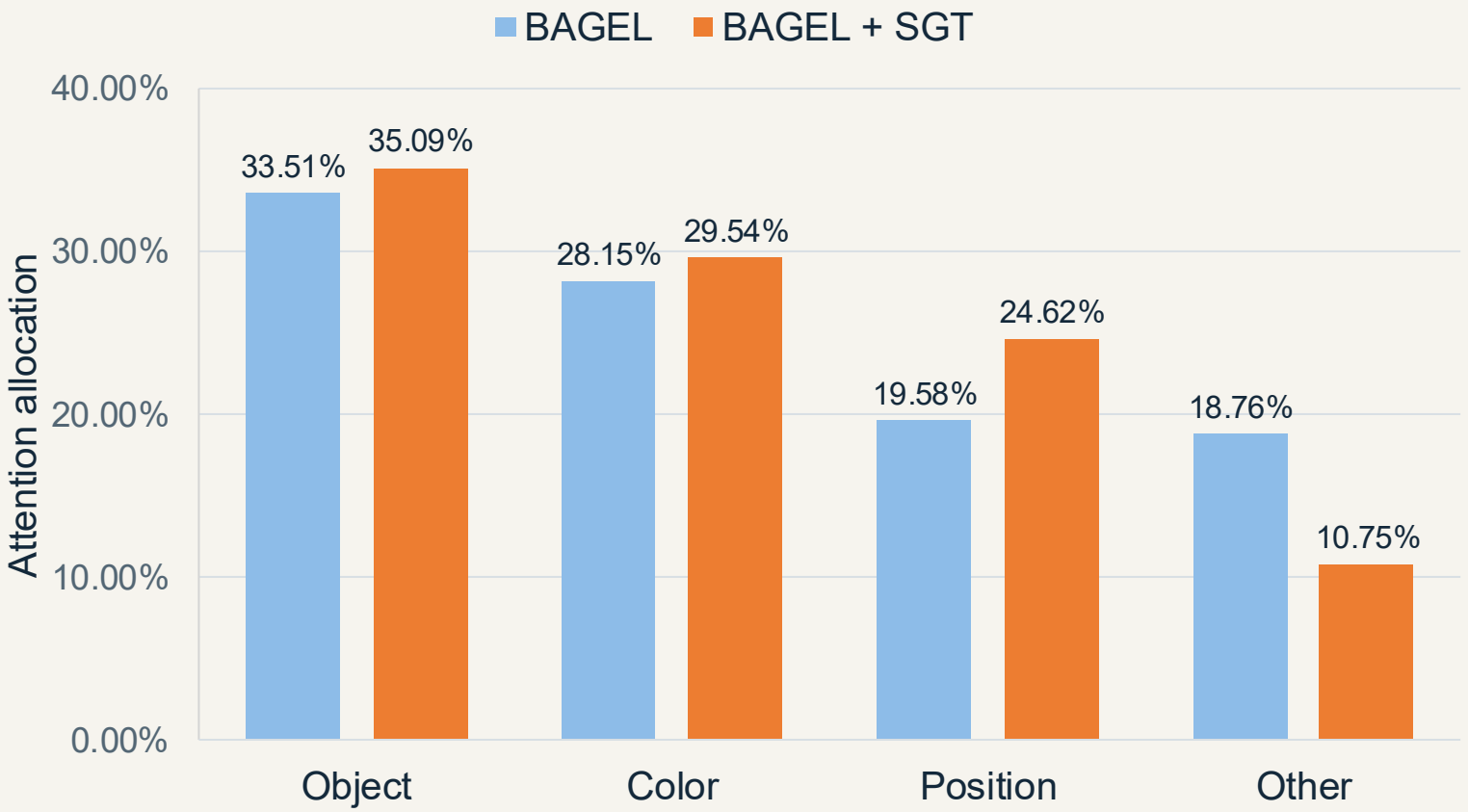
**Visual attention at L20**

**+14.6%**

**Visual attention at L25**

**Segmentation induces the strongest deep-layer shift toward visual tokens**

## 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%

**More attention to object, color, and position tokens, less to other words**

# *LIMITS AND FUTURE WORK*

***SGT complements task-specific supervision***

## STRONGEST TRANSFER

- **Natural-scene perception**
- **Spatial grounding**
- **Hallucination resistance**

## LIMITED TRANSFER

- **OCR and document reading**
- **External knowledge**
- **Abstract reasoning**

## Semantic alignment + task-specific supervision

**VQA helps preserve symbolic skills while semantic targets improve visual grounding.**

**Future work: joint post-training with generation, editing, and reinforcement learning**

# TAKEAWAY

*Semantic structure is the useful bridge*

1. **High-level proxies give the strongest joint transfer in our study.**

2. **Segmentation preserves object boundaries and spatial layout.**

3. **SGT improves both capabilities on BAGEL and OmniGen2.**

# Thank you!

Paper

Website

Code

Connect me if you have more question:
liyanwei@sjtu.edu.cn

# BACKUP: FULL RESULTS

## Unified benchmark results for BAGEL and OmniGen2

| OmniGen2 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **CV-Bench** | **MMVP** | **VSR** | **SIBench** | **POPE** | **Hallusion** | **MMBench** | **MMMU** | **MMStar** | **GenEval** | **GEdit-Bench-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 + Reconstruction | 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-Bench-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 + Reconstruction | 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