ComputerVisionLab
Seoul National University

NEURAL INFORMATION PROCESSING SYSTEMS

Structural Self-Teaching for Compositional Generation in Unified Multimodal Models

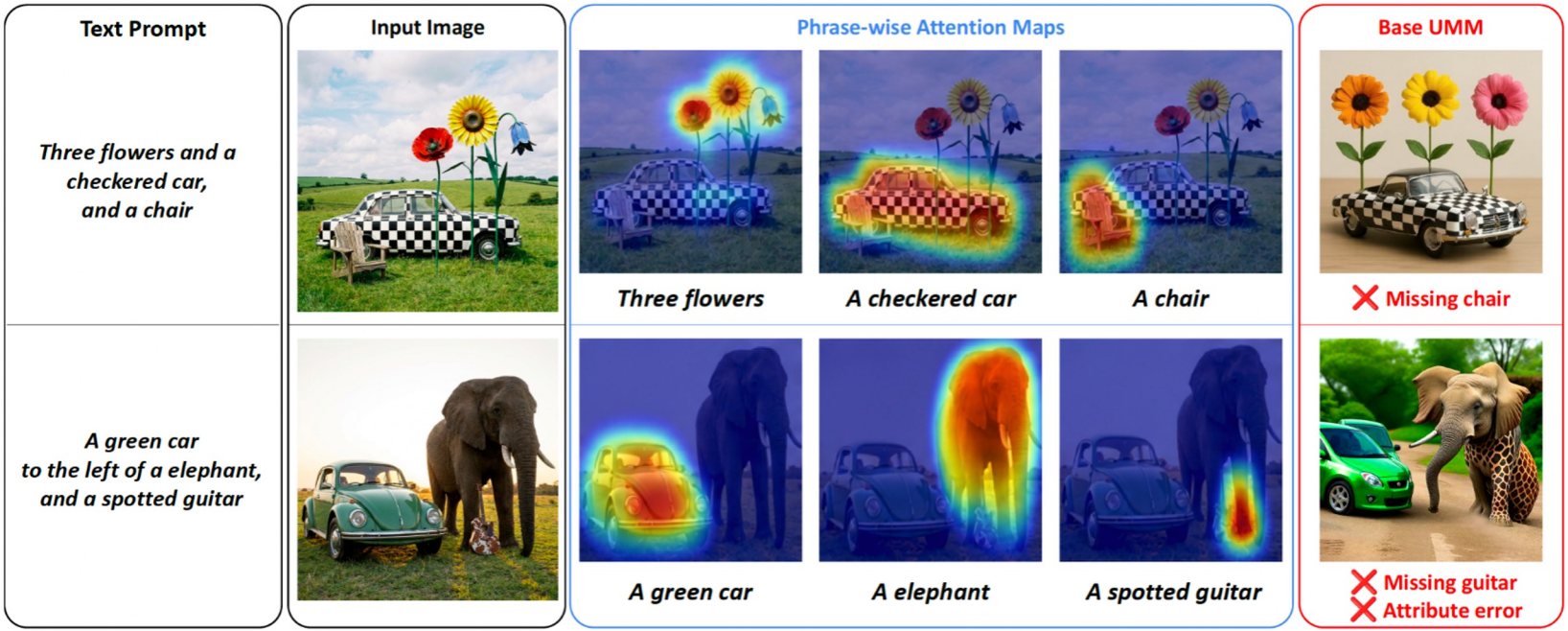
Jaeyoo Park
Seoul National University

Bohyung Han
Seoul National University

bellos1203.github.io/sst

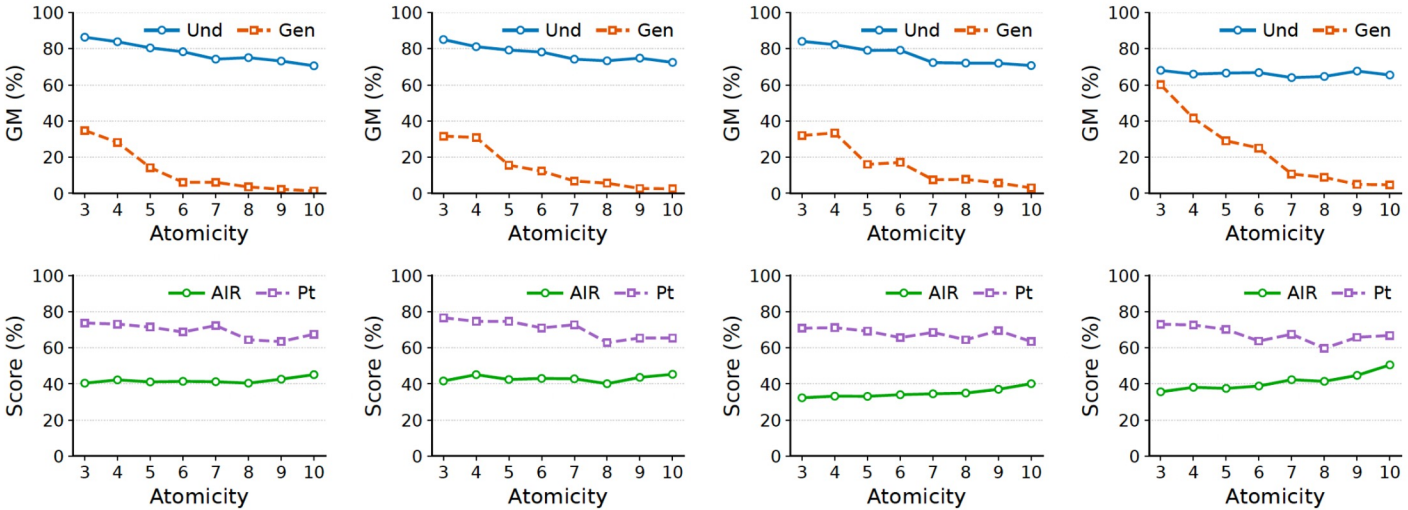
Understanding Does Not Guarantee Generation

- UMMs localize phrases they fail to generate
  - Can generation learn from its own understanding path?



Understanding as a source of structure

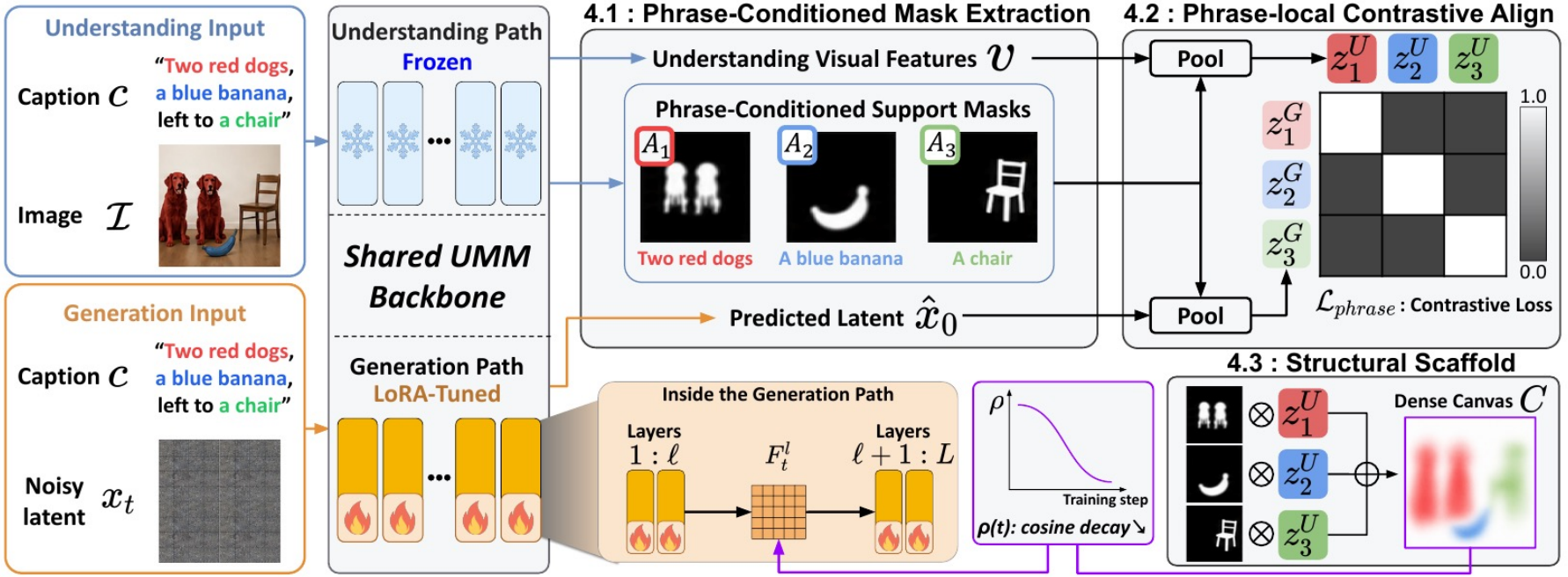
- Understanding stays stable as atomicity grows, while generation declines
  - Top: Understanding on oracle images (**Und**) and prompt-only generation (**Gen**)
  - Bottom: Phrase localization remains stable across prompt complexity



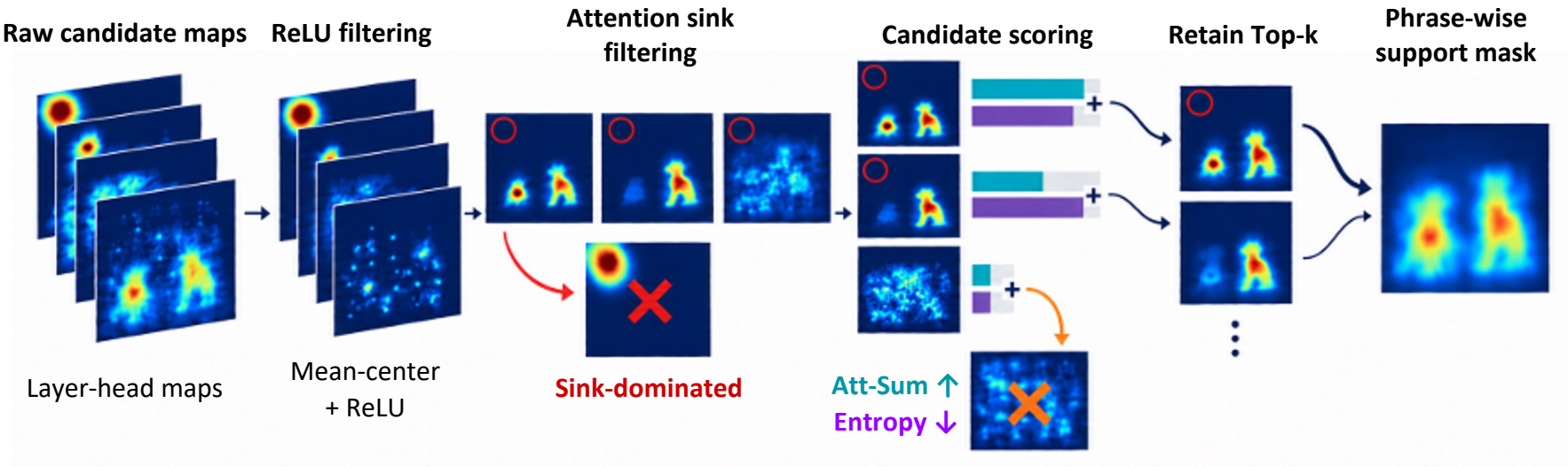
(a) BLIP3o 4B (b) BLIP3o 8B (c) Show-o2 7B (d) BAGEL

Overall Framework

- Phrase-conditioned masks connect understanding to generation
  - Frozen understanding path and LoRA-tuned generation path



Phrase-Conditioned Support Mask Extraction

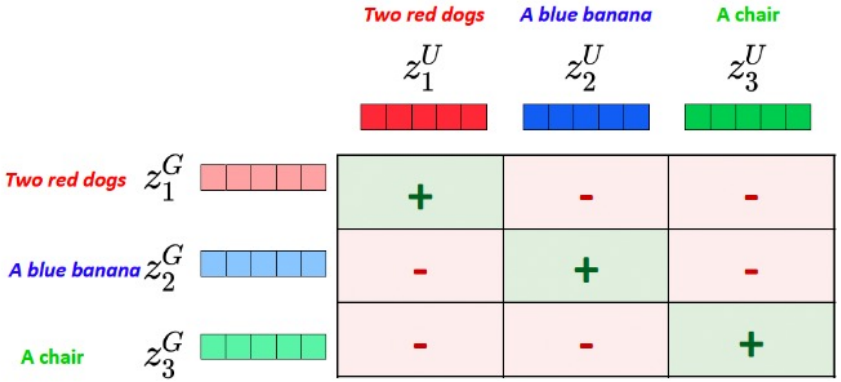


1. Suppress weak activations with mean-centered ReLU.
2. Discard sink-dominated maps, then zero attention sinks and renormalize.
3. Rank maps per phrase by attention mass and spatial concentration.
4. Softmax-weight the top-K maps and normalize into one soft phrase mask.

Phrase Alignment and Structural Scaffolding

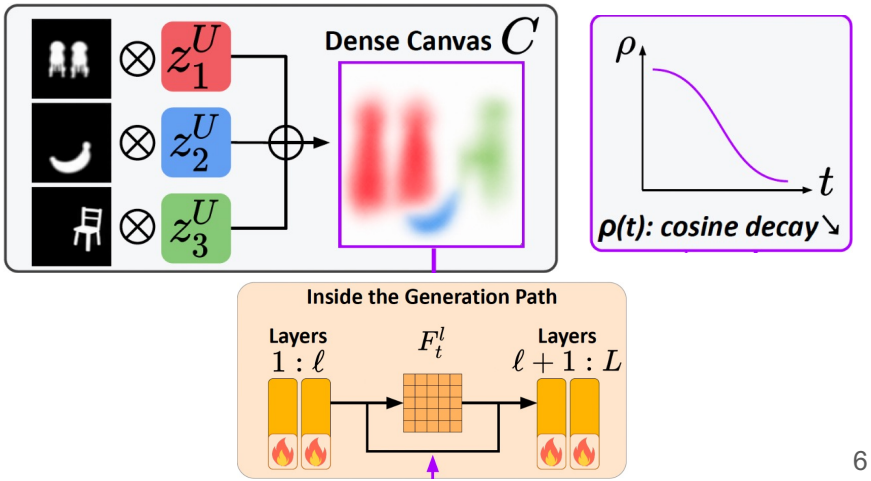
Phrase-local contrastive alignment

- Pool each phrase region from generated and reference features
- Align each phrase with its reference and use other phrases as negatives



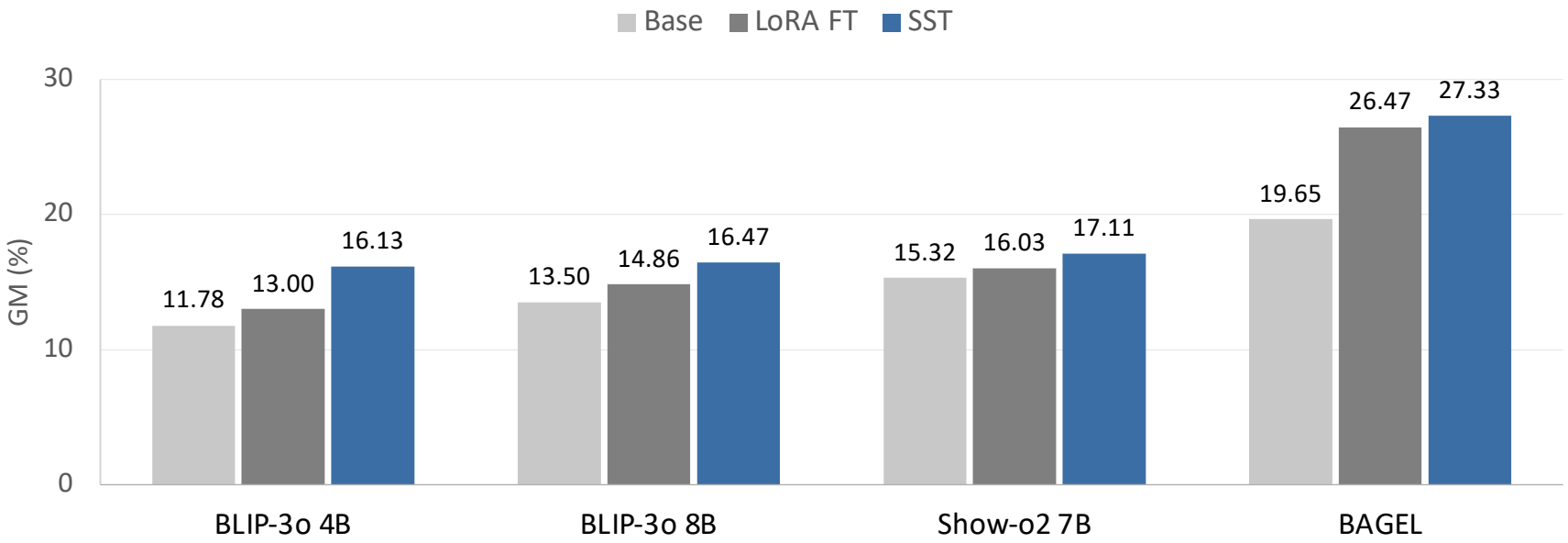
Attenuated structural scaffold

- Build a canvas from phrase masks and reference features
- Inject during training, then phase out and remove at inference



Main Results on GenEval2

- SST improves compositional generation across all four backbones
  - Same training data and LoRA settings. Only the supervision changes.



GenEval2 GM (%). Average over three training runs.

Qualitative Results

- SST reduces omission, entanglement, and spatial errors

BLIP-3o 4B :
"a penguin and three pink cows"

Base | LoRA FT | SST (Ours)



BLIP-3o 8B : "a candle to the left of a brown kangaroo to the left of a toy"

Base | LoRA FT | SST (Ours)



Show-o2 7B :
"a white violin and a red bicycle"

Base | LoRA FT | SST (Ours)



BAGEL :
"a metal clock on top of a bear"

Base | LoRA FT | SST (Ours)



Conclusion

- The understanding path retains useful structure even when generation struggles.
  - Phrase-conditioned support masks reveal where each phrase is visually grounded.

- SST transfers this structure into generation training.
  - Alignment separates phrase features, while scaffolding provides spatial guidance.

- SST outperforms matched LoRA fine-tuning across four backbones.
  - Gains reflect structural supervision, while exact counting remains challenging.