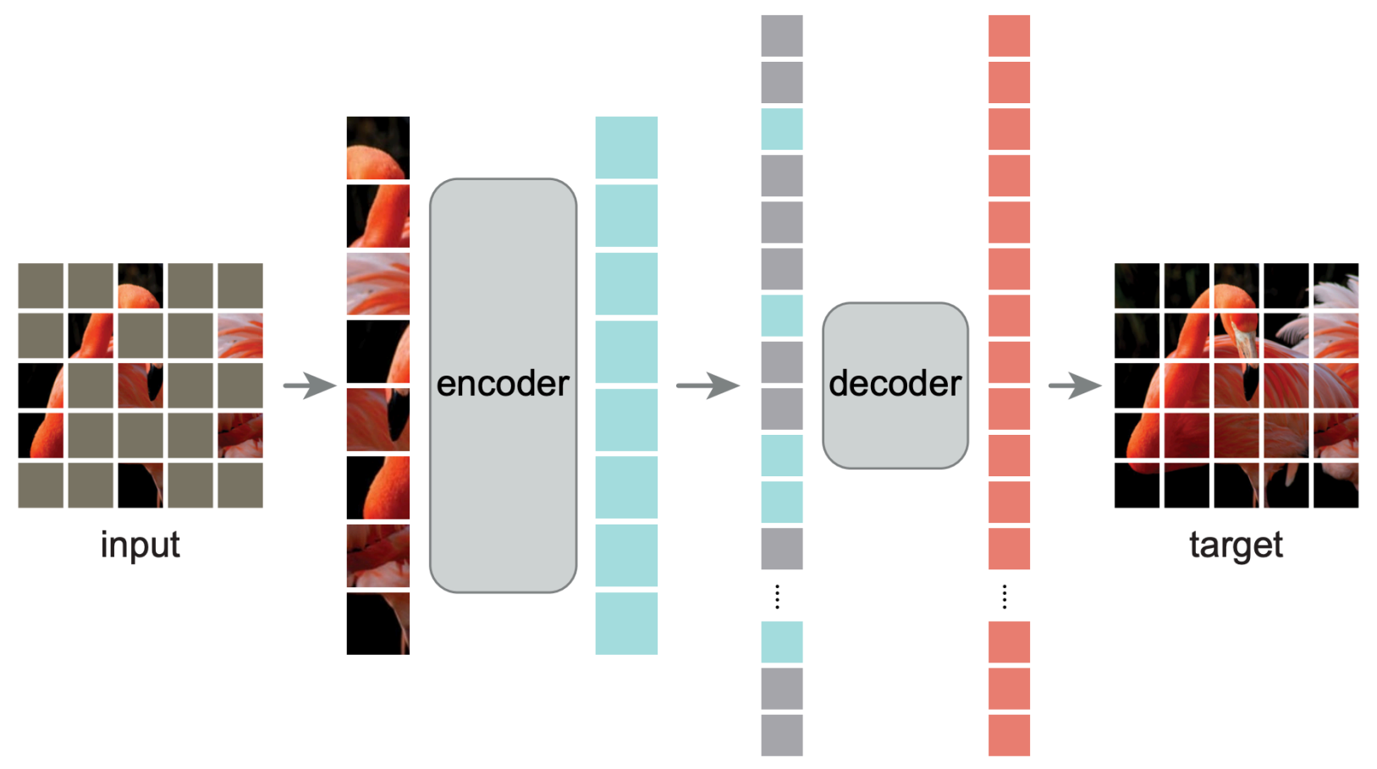
# **MultiMAE**: Multi-modal Multi-task Masked Autoencoders

Roman Bachmann* David Mizrahi* Andrei Atanov Amir Zamir

## Motivation

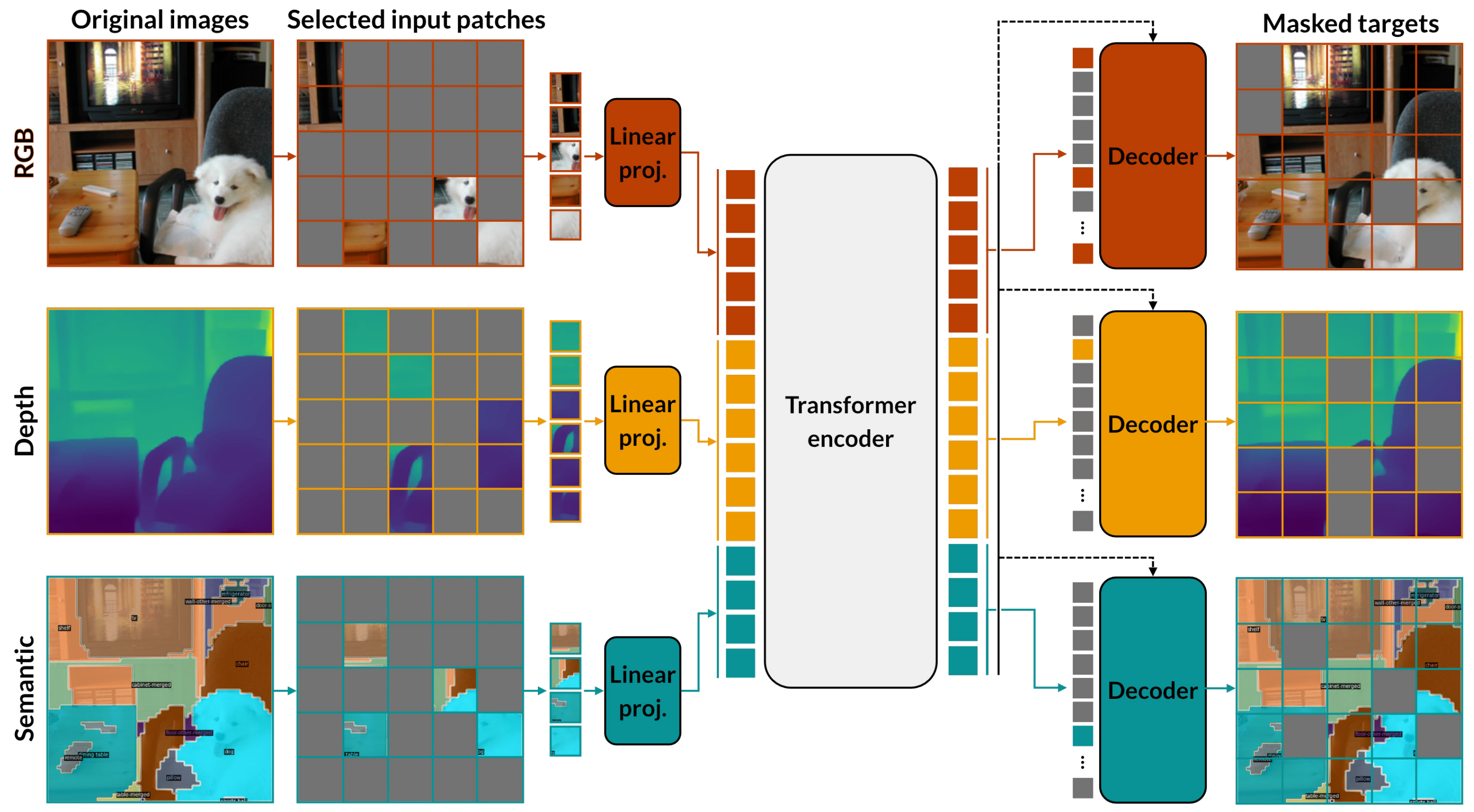
- We can process multiple **modalities** & solve many **tasks**. **Our machines should too!**
- Masked Autoencoders (MAE) [1] are a **simple** and **powerful** pre-training strategy, but limited to a **single** modality.



- We propose to use **multi-modal masking** to learn strong **cross-modal predictive coding** abilities and **shared scene representations**.

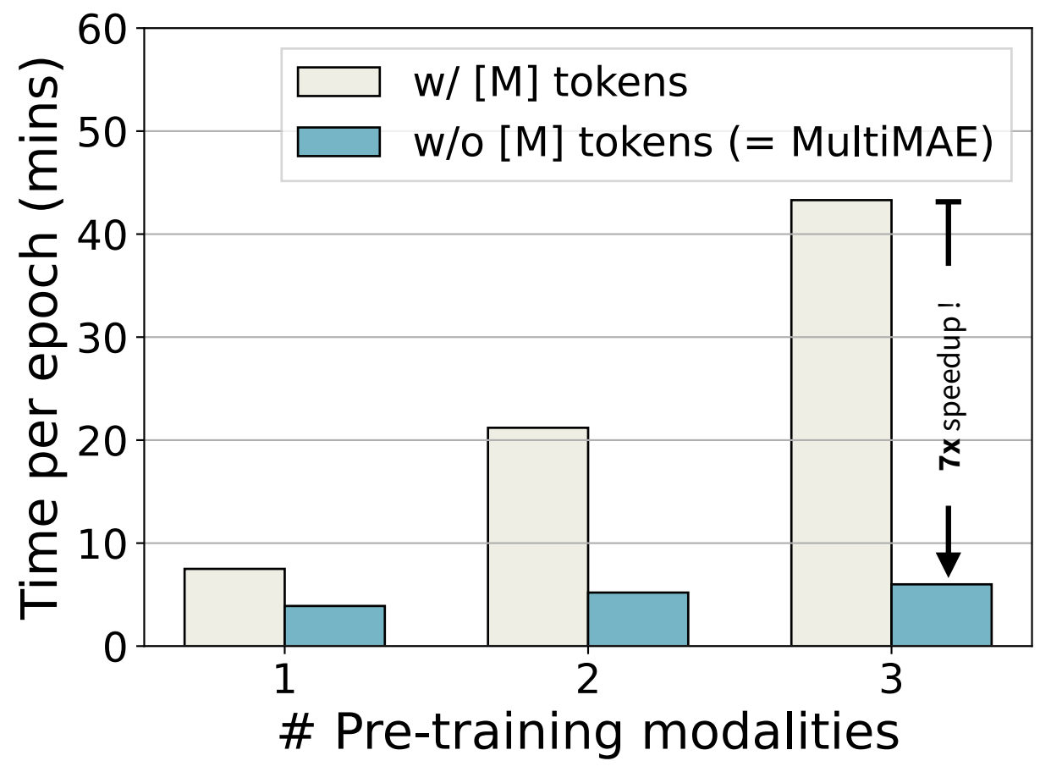
## MultiMAE pre-training

**Pre-training objective:** Reconstruct masked-out patches of multiple modalities

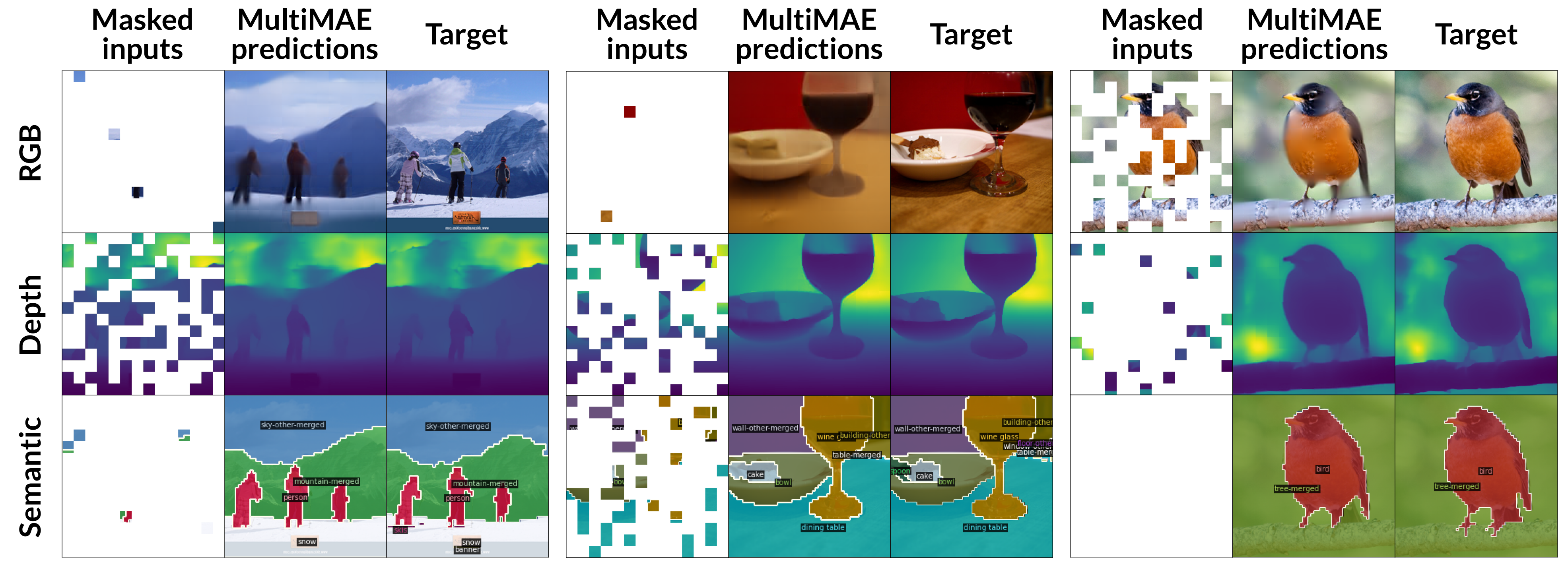


**Key properties:**

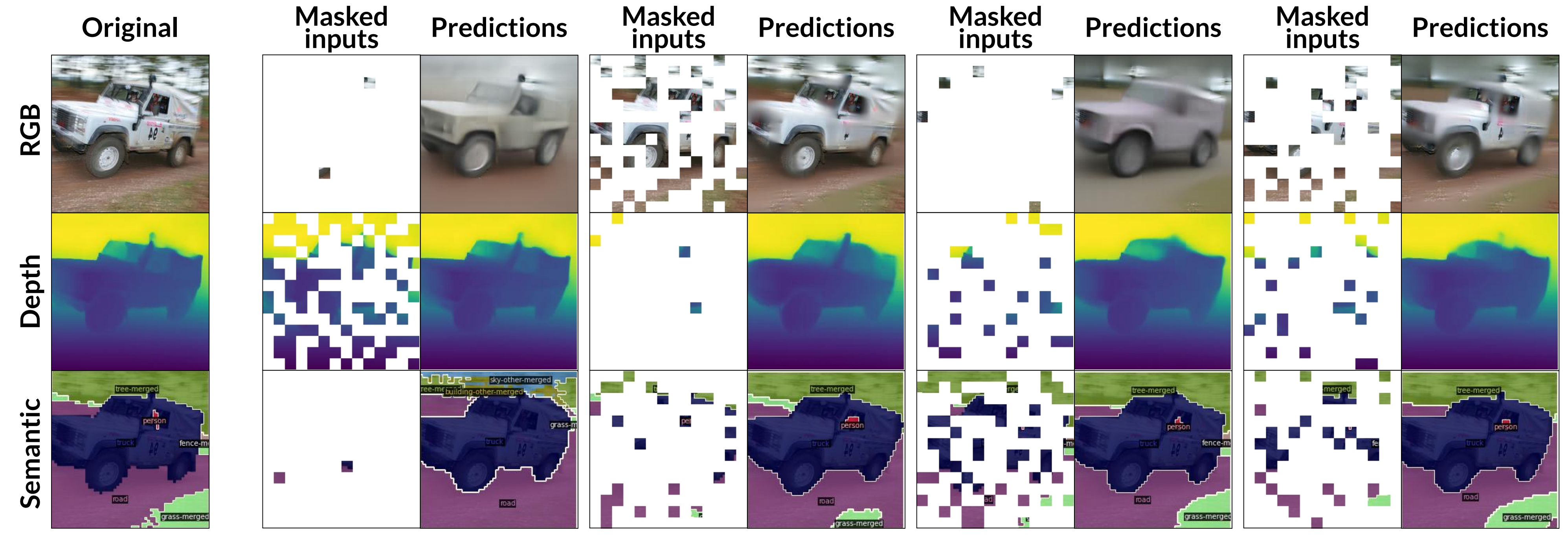
- **Applicable to any RGB dataset:** To avoid needing a large multi-task dataset, additional modalities are **entirely pseudo labeled**
- **Joint training:** Only a single pre-training run is needed to obtain a model that accepts **any combination of input modalities**
- **Efficient:** High masking ratio + **shared encoder with no mask tokens** (as in MAE) is especially beneficial in a multi-modal setting



## Masked multi-modal reconstructions

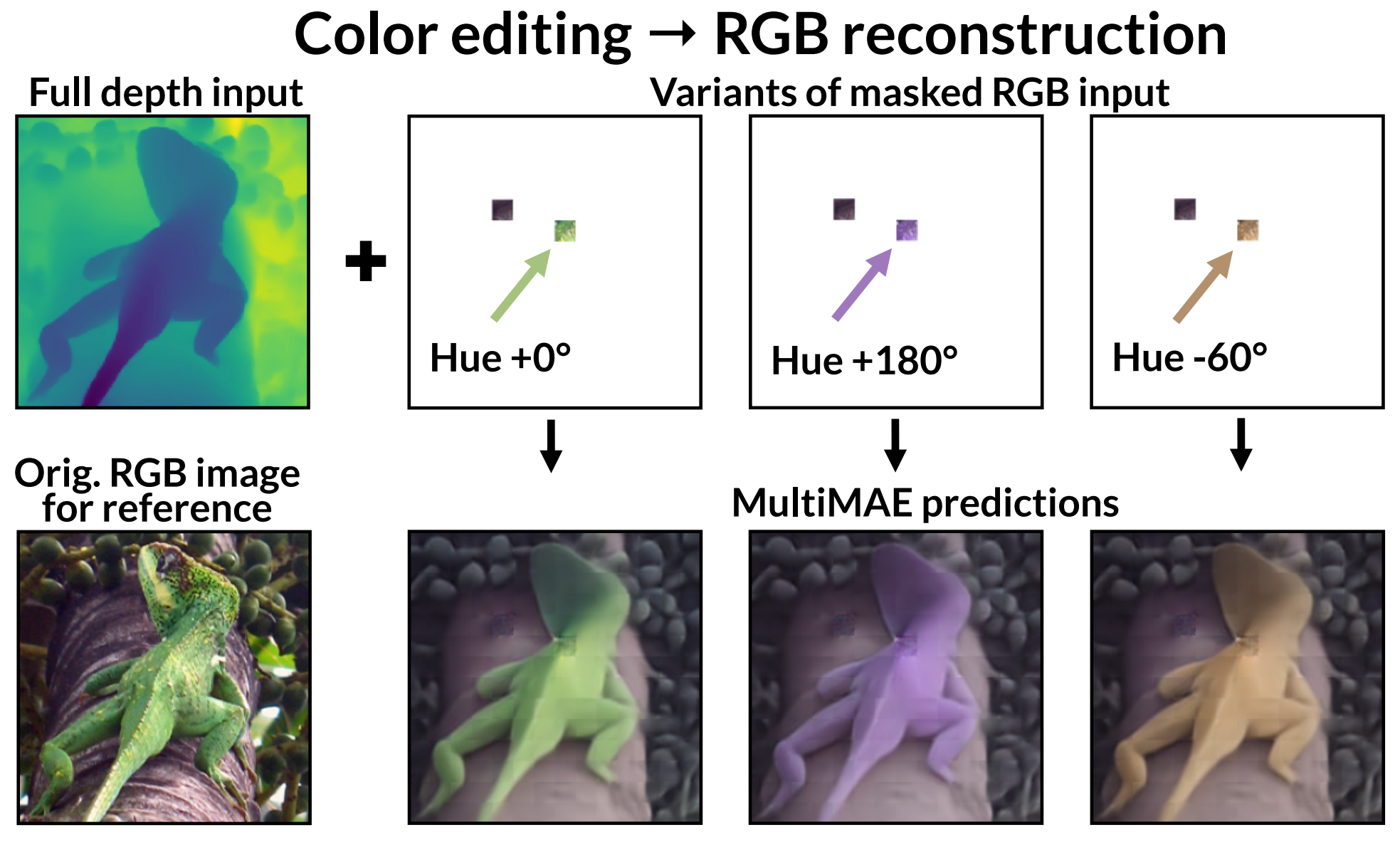


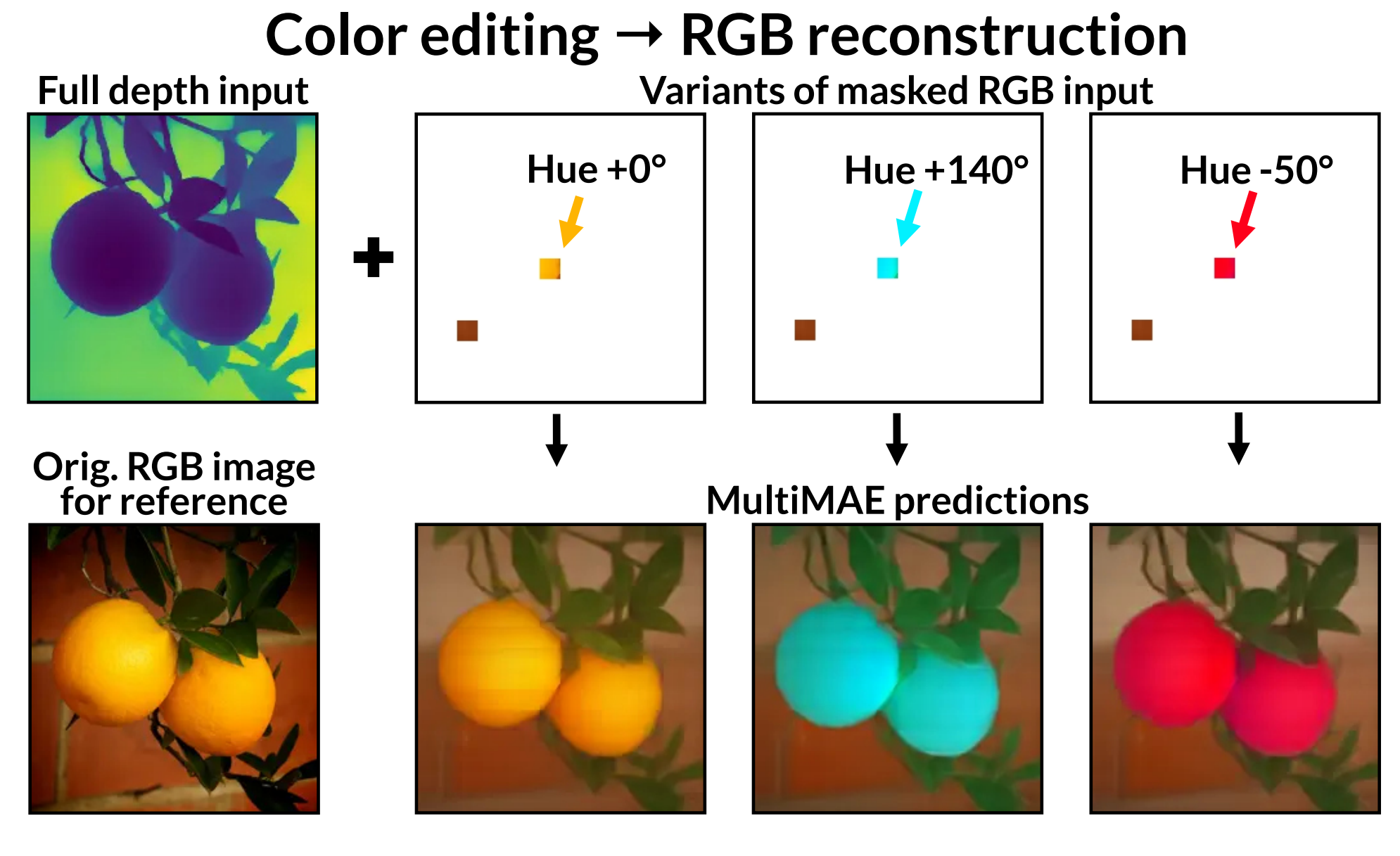
- Any-to-any cross-modal predictive coding learns **shared representations.** No matter the inputs given, predictions are semantically stable.

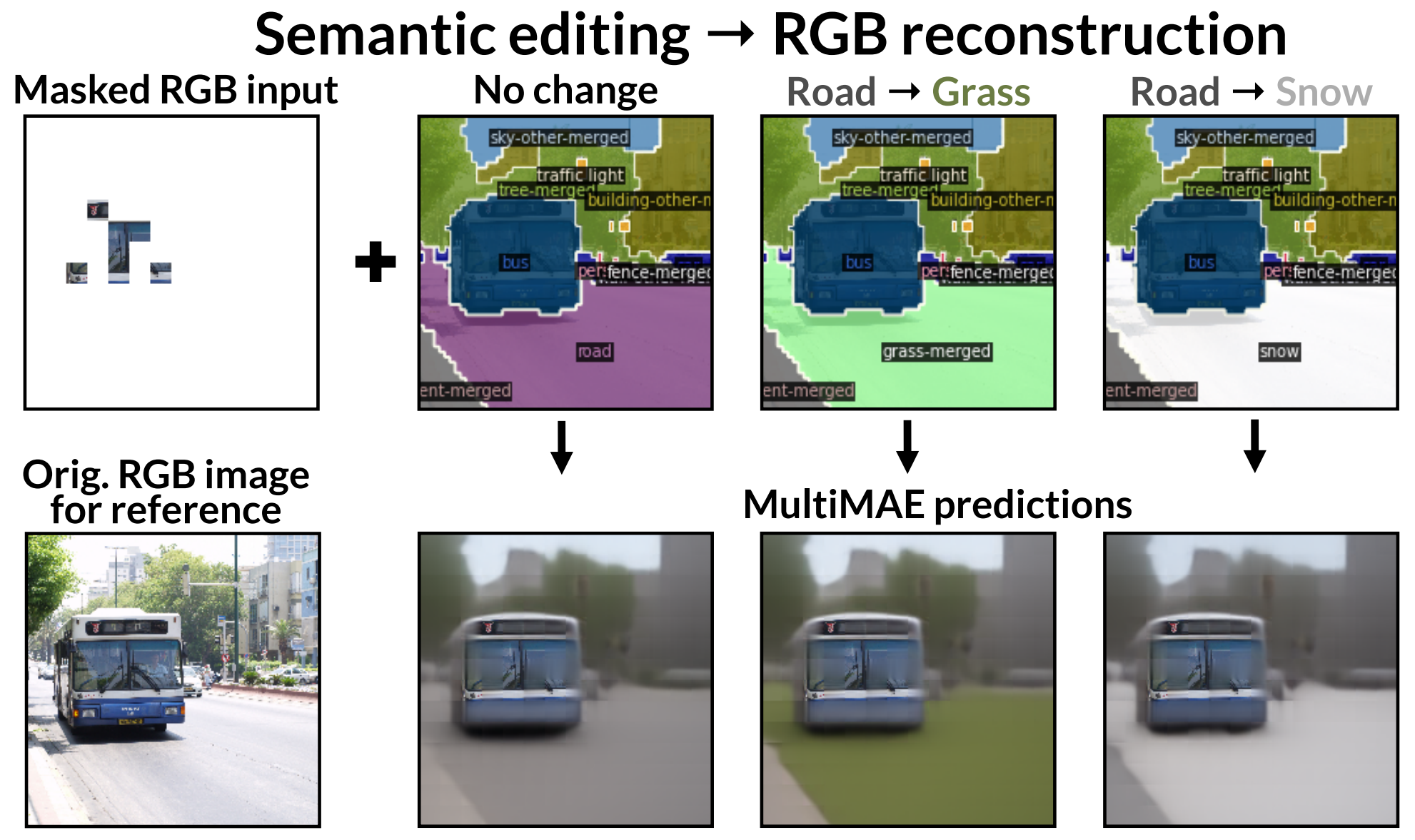


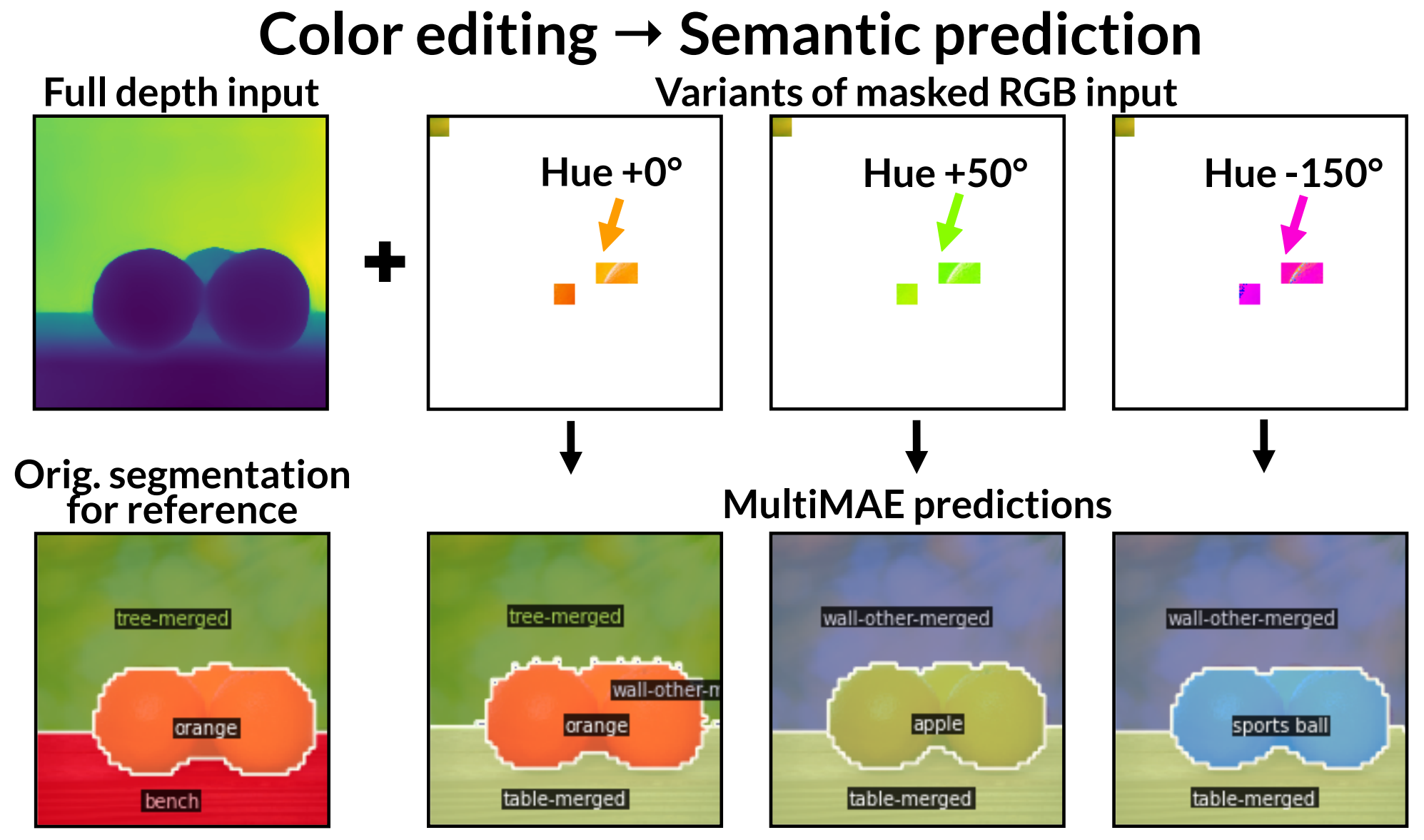
## Cross-modal predictive coding

- MultiMAE learns to effectively integrate information from different modalities, as shown here through through input modification.

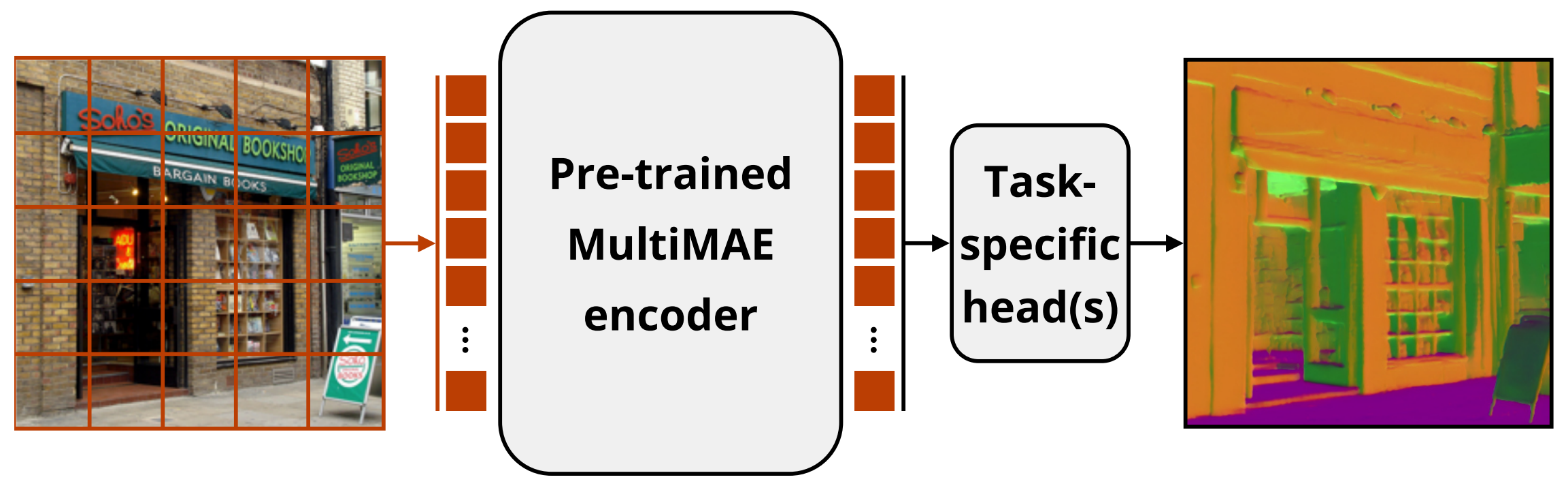

Color editing → RGB reconstruction
Full depth input
Variants of masked RGB input
Hue +0°
Hue +180°
Hue -60°
Orig. RGB image for reference
MultiMAE predictions


Color editing → RGB reconstruction
Full depth input
Variants of masked RGB input
Hue +0°
Hue +140°
Hue -50°
Orig. RGB image for reference
MultiMAE predictions





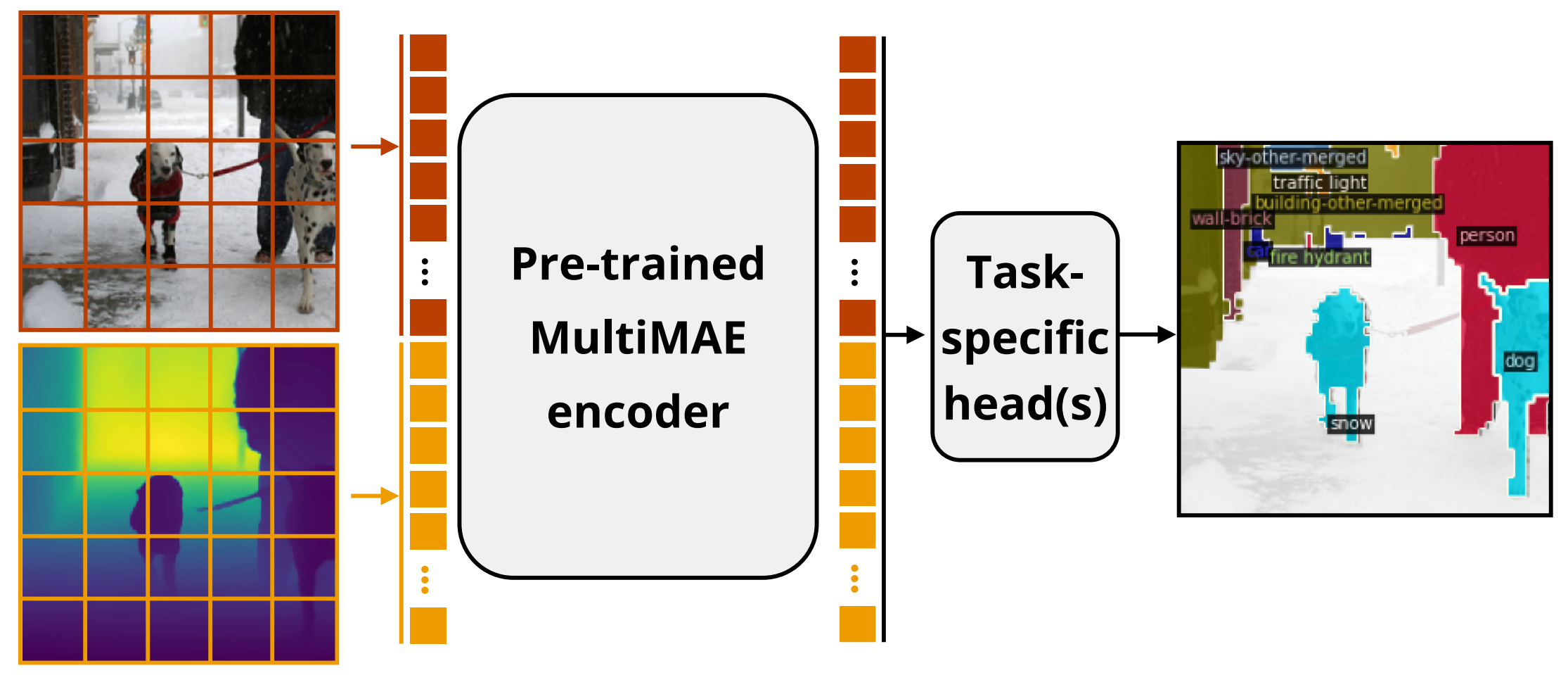
## RGB-only transfer



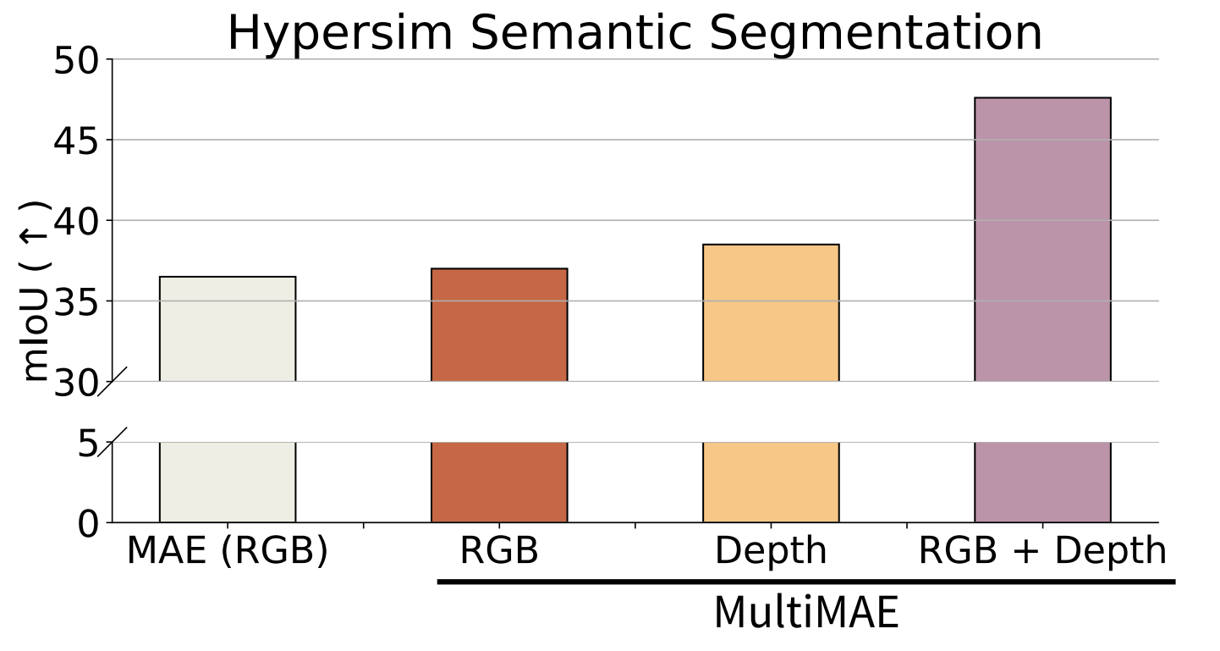
- Significantly outperforms ImageNet-supervised baseline (DeiT) & is competitive with MAE

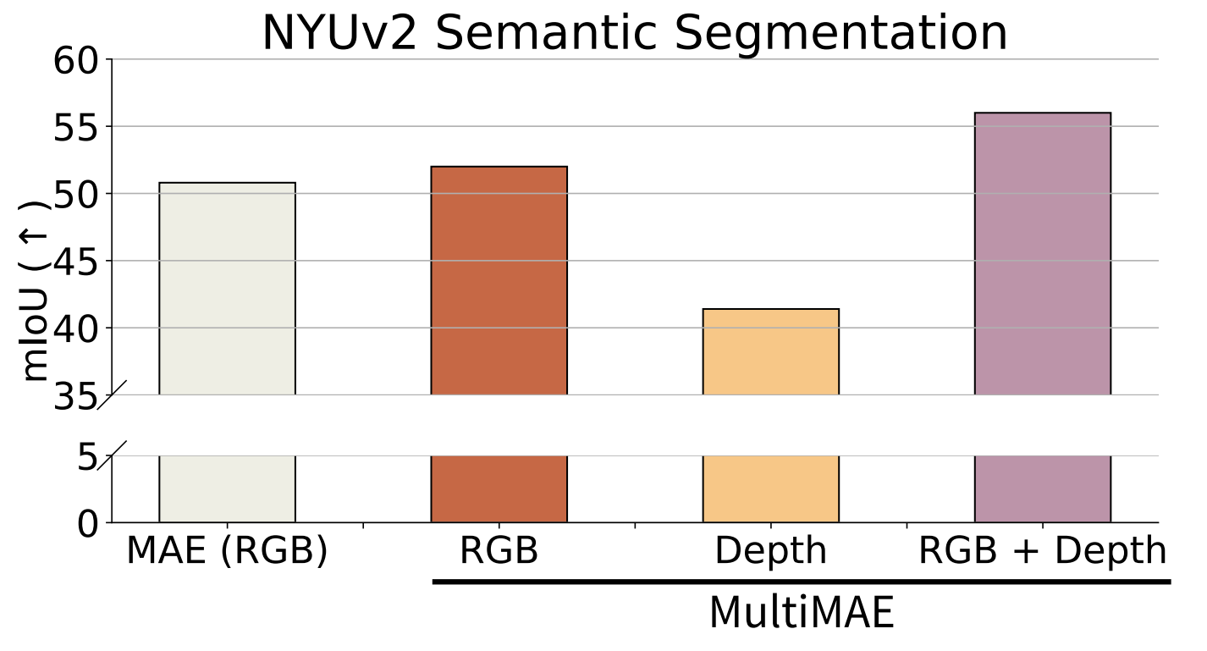
| Method | Arch. | Classification (Top 1 acc. ↑) | Semantic Segmentation (mIoU ↑) | | | Depth (δ1 ↑) |
|---|---|---|---|---|---|---|
| | | ImageNet-1K | ADE20K | Hypersim | NYUv2 | NYUv2 |
| Supervised (DeiT) | ViT-B | 81.8 | 45.8 | 33.9 | 50.1 | 80.7 |
| MAE | ViT-B | **83.3** | **46.2** | 36.5 | 50.8 | 85.1 |
| **MultiMAE** | ViT-B | **83.3** | **46.2** | **37.0** | **52.0** | **85.4** |

## Multi-modal transfer



- Supports **any subset of the modalities** used in pre-training
- If ground-truth modalities are unavailable, **can also accept pseudo labels** for improved performance over RGB-only





## Summary

**MultiMAE**: a **simple** and **efficient multi-modal pre-training** strategy for Vision Transformers

- Relies on **masking** to learn strong **cross-modal predictive coding** abilities
- **Retains the benefits of MAE** for RGB-only transfer
- Notable **performance gains** for multi-modal transfer

**multimae.epfl.ch**

**References:**

[1] Masked Autoencoders Are Scalable Vision Learners. *He et. al.* CVPR 2022