Vision-*and*-Language or Vision-*for*-Language? On Cross-Modal Influence in Multimodal Transformers

Stella Frank* **Emanuele Bugliarello*** **Desmond Elliott**

UNIVERSITÀ DI TRENTO — CiMeC Center for Mind/Brain Sciences

UNIVERSITY OF COPENHAGEN

Cross-Modal Input Ablation

RQ: Do V&L transformers use both input modalities equally?

We answer this question using *cross-modal input ablation:*

- Remove one modality at test time
- If performance changes, trained model expects both modalities & can recruit features cross-modally

Experiments

- Evaluation data: Flickr30k Entities val dataset
- Models: 5 V&L BERTs from VOLTA (Bugliarello+, 2021)

Findings

- All models use vision-for-language predictions effectively
- All models do *not* recruit language for vision tasks

Further exps show language-for-vision is not affected by:

- Architectures (e.g. single vs dual stream)
- MRC loss (cross-entropy vs KL divergence)
- Pretraining: initialisation, vision-first or V&L throughout
- Co-masking of detected objects

However, we find that Faster R-CNN object detector **predictions** often ***do not match*** human **descriptions**

Hard to learn link between language labels & visual categories!

Take-Away

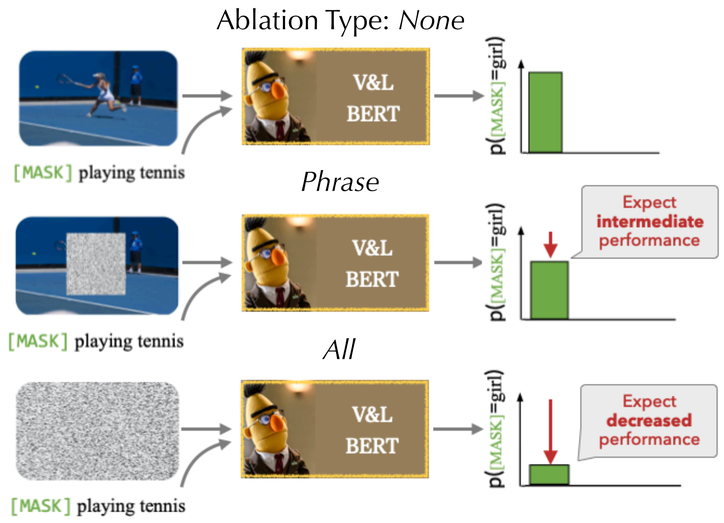
Cross-modal input ablation

- Straightforward check for cross-modal influence

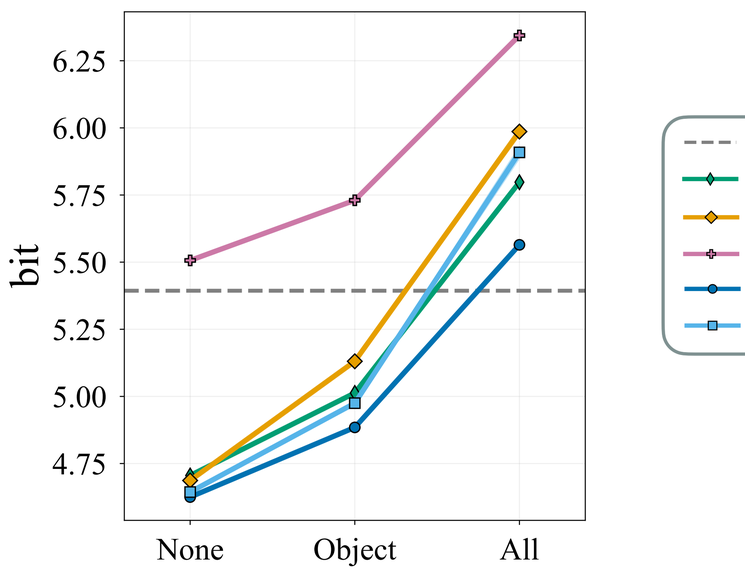
Future directions

- Better use of silver annotations from object detectors
- More downstream language-for-vision tasks needed

Vision-for-language Ablation

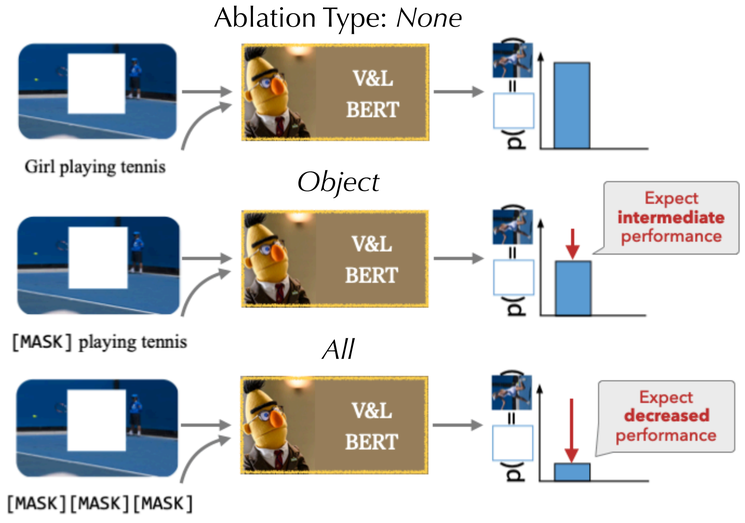


Performance on Masked Language Modelling

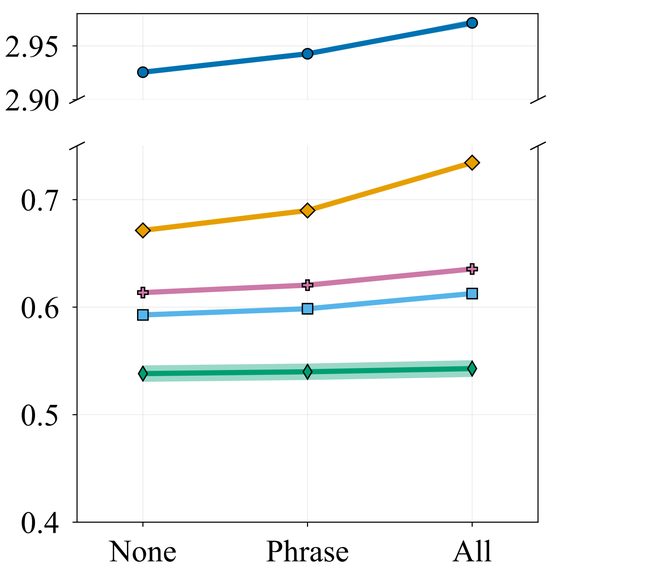


Lower performance without visual information

Language-for-Vision Ablation



Performance on Masked Region Classification



Similar performance with and without text