At Meta, nearly every post related to commerce is multimodal.

Influencers upload their fashion posts to Instagram with captions and hashtags.

Cross-modal pairs:
1) Masked Language Modeling (MLM)
2) Masked Image Modeling – Feature Regression (MIM-fr)
3) Masked Image Modeling – KL-Divergence (MIM-kl)

Multimodal pre-training model:

Trade-offs:
- MLM: single-modal pre-training
- MIM-kl: extends single-modal pre-training with an additional contrastive loss
- MIM-fr: multi-modal pre-training

McNemar’s test is not significant, which works like a gate function to turn on/off

Cross-modal retrieval:
- Cross-modal retrieval
- Multi-modal retrieval

Attention Visualization:
- Visualization of text-to-image attention from different pre-trained models.