Meta FAIR Finds That Training Vision and Language Together From Scratch Beats Bolting Them Together Later

Kabui, Charles

2026-03-08

Read at ToKnow.ai

Meta FAIR: Training Vision and Language Together From Scratch

Unified multimodal pretraining beats bolt-on approaches

51x Vision needs more data than language at 1T params

50% MoE halves the scaling gap between modalities

1% Domain data needed for world modeling to emerge

Meta FAIR published “Beyond Language Modeling”, a large-scale study on how to build native multimodal foundation models from scratch, without starting from a language-pretrained backbone. The team used the Transfusion framework (next-token prediction for language, diffusion for vision) and trained on text, video, image-text pairs, and action-conditioned video simultaneously. Four findings stand out.

1. First, a Representation Autoencoder (RAE) built on SigLIP 2 outperforms both VAE-based encoders and raw pixels for visual understanding and generation, eliminating the need for separate encoders.
2. Second, adding vision data to language training does not hurt language performance; the two modalities are synergistic, and mixed-data training beats domain-specific training even with 5x less in-domain data.
3. Third, world-modeling capabilities (like navigation) emerge from general pretraining with as little as 1% domain-specific data.
4. Fourth, Mixture-of-Experts (MoE) architectures naturally learn modality specialization, with the model allocating more experts to text in early layers and more to vision in later layers, without human priors.

The most consequential number in the paper comes from its scaling law analysis: at 1 trillion parameters, vision's optimal data requirement is 51x larger than language's. This asymmetry explains why bolting a vision encoder onto a pretrained language model produces mediocre visual reasoning. MoE cuts this gap in half (exponent difference from 0.10 to 0.05), making a single model competitive with separate unimodal models on both modalities. For teams building multimodal products, the implication is direct: plan for dramatically more visual training data, or use MoE to compensate.

This work provides empirical backing for a shift already underway. Qwen3.5 ships a unified early-fusion multimodal model with MoE at consumer scale. Meta's paper now explains the theory behind why that architecture works, and where the remaining bottlenecks are.

Sources:

- Beyond Language Modeling: An Exploration of Multimodal Pretraining (arXiv)
- Paper Discussion on Hugging Face
- Transfusion: Predict the Next Token and Diffuse Images (arXiv)
- SigLIP 2: Multilingual Vision-Language Encoders (arXiv)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** ***/terms-of-service***