

**Next Text Token Prediction**

**Next Visual State Prediction**

**Autoregressive Model**

**Text Tokenizer**

**Vision Encoder**

**Visual  
Representation**

**Data**

**World Model**

**Architecture**

**Scaling  
Behavior**

