OSKAR: Omnimodal Self-supervised Knowledge Abstraction and Representation

Mohamed Abdelfattah*†, Kaouther Messaoud*, and Alexandre Alahi

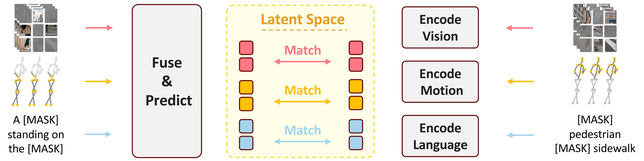
EPFL VITA

NEURAL INFORMATION PROCESSING SYSTEMS

1 Overview

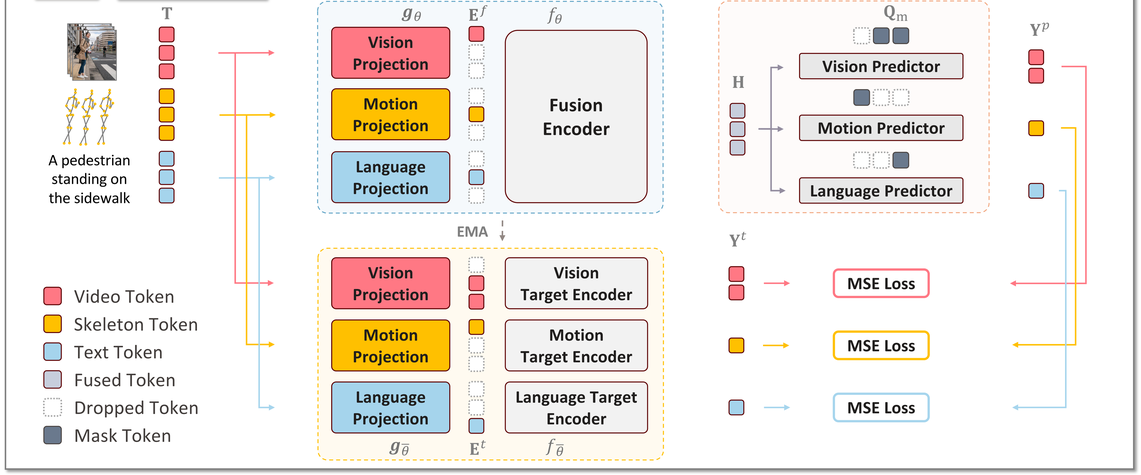
In one sentence:
OSKAR is a self-supervised multimodal foundation model that learns in the latent space.

1) Pretraining Stage

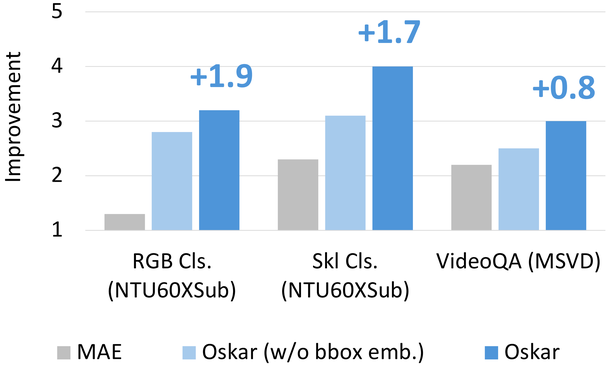


2) Downstream Tasks

2 Method



3 Features vs Pixels

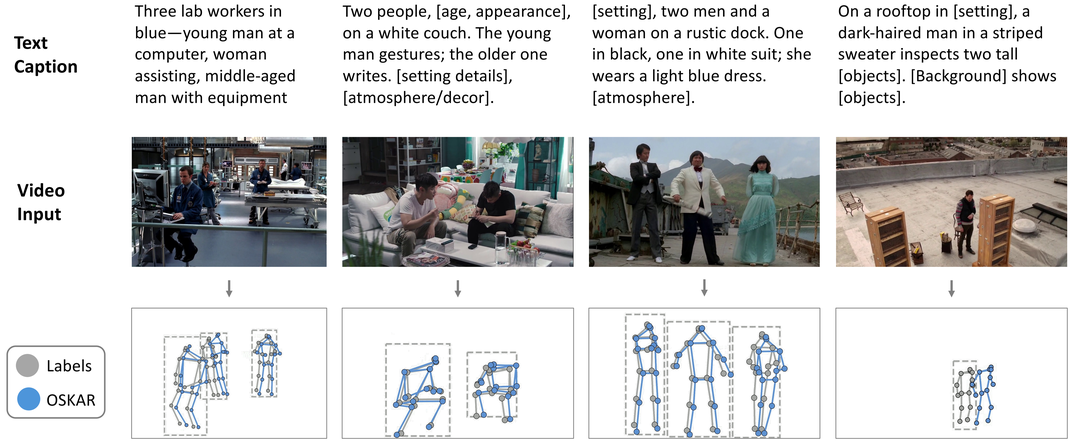


Predicting the **latent features** results in superior downstream performance compared to predicting in the pixel space.

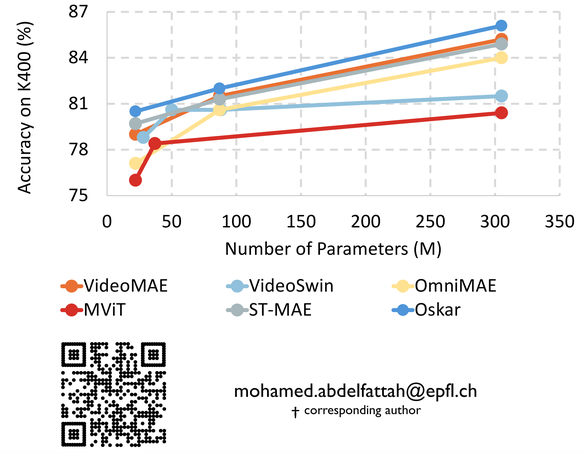
4 Action Recognition

Method	Resolution	GFLOPs	Acc.
Small Models (<80M parameters)			
VideoMAE-S	16×224²	57	79.0
SlowFast+NL	80×224²	234	79.8
MViTv1-B	32×224²	170	80.2
OSKAR-S	**16×224²**	**57**	**80.5**
Medium Models (80–150M parameters)			
OmniMAE-B	16×224²	180	80.6
TimeSformer-B	96×224²	2380	80.7
BEVT-B	32×224²	282	81.1
ST-MAE-B	16×224²	180	81.3
VideoMAE-B	16×224²	180	81.5
OSKAR-B	**16×224²**	**180**	**82.0**
Large Models (150–700M parameters)			
VideoSwin-L	32×224²	604	83.1
OmniMAE-L	16×224²	597	84.0
VideoMAE-L	16×224²	597	85.2
V-JEPA-L	16×224²	597	85.6
OSKAR-L	**16×224²**	**596**	**86.1**

5 Feature visualization



6 Scalability



mohamed.abdelfattah@epfl.ch
† corresponding author