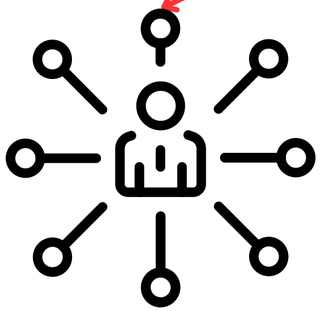
BEHAVIOUR SIMULATION CHALLENGE

TEAM 18

Background

Why is target and timely content valuable?

High user engagement

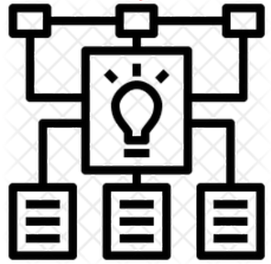


Brand Reach

Product Recall

Brand Credibility

High sales



Algorithmic Favorability

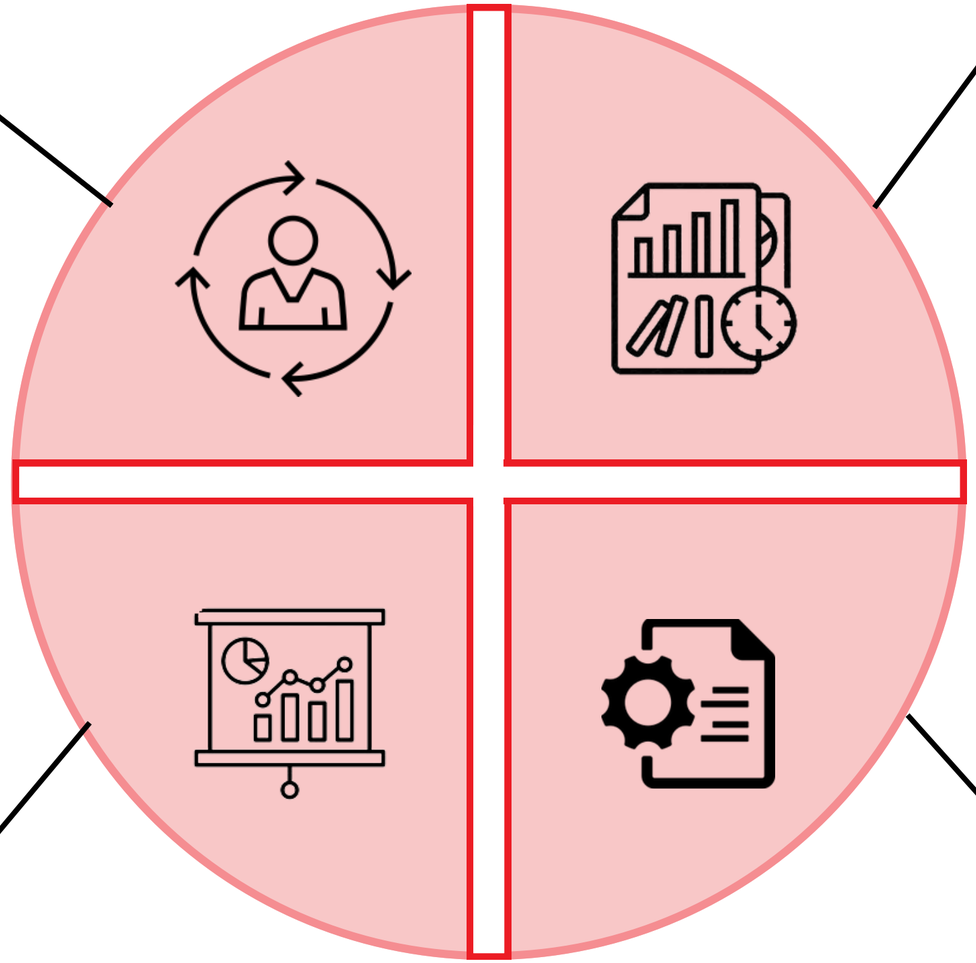
Background

PERSONALIZATION

Personalization to build strong connections between brand and consumers

PREDICTING FUTURE TRENDS

Predicting future trends to align with evolving customer needs



DATA-DRIVEN DECISIONS

Making data-driven decisions for building strategies which will yield desired outcomes

CONTENT VISIBILITY

Optimizing content visibility to increase the likelihood of content promoted

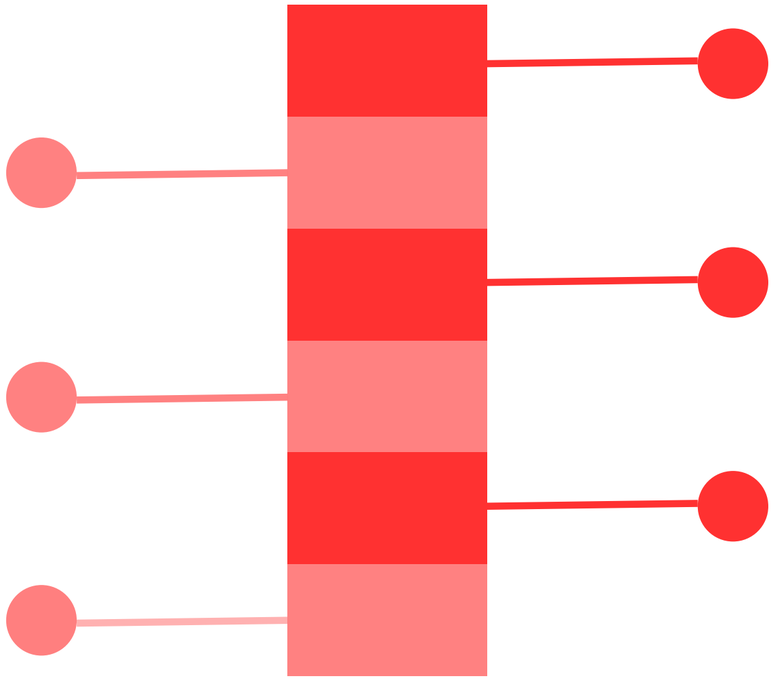
Multi-Modal Masked Learning M3L with InstructBLIP

Our Novel Approach: Multi-Modal Masked Learning (M3L)

Unified model to handle **static** and **motion media**

Captures human behavior

Scalable for **new modalities** such as time and images

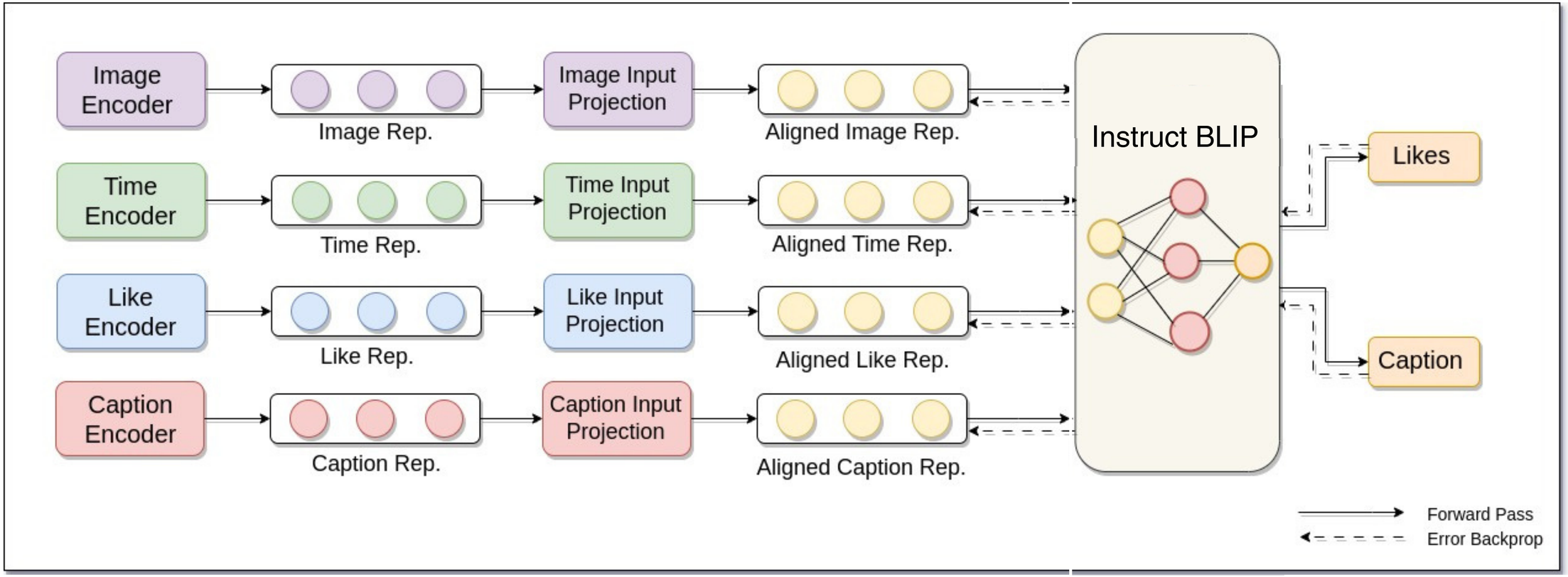


Unified model for both task: behavior and content simulation

Captures intricate relationships between **different modalities**

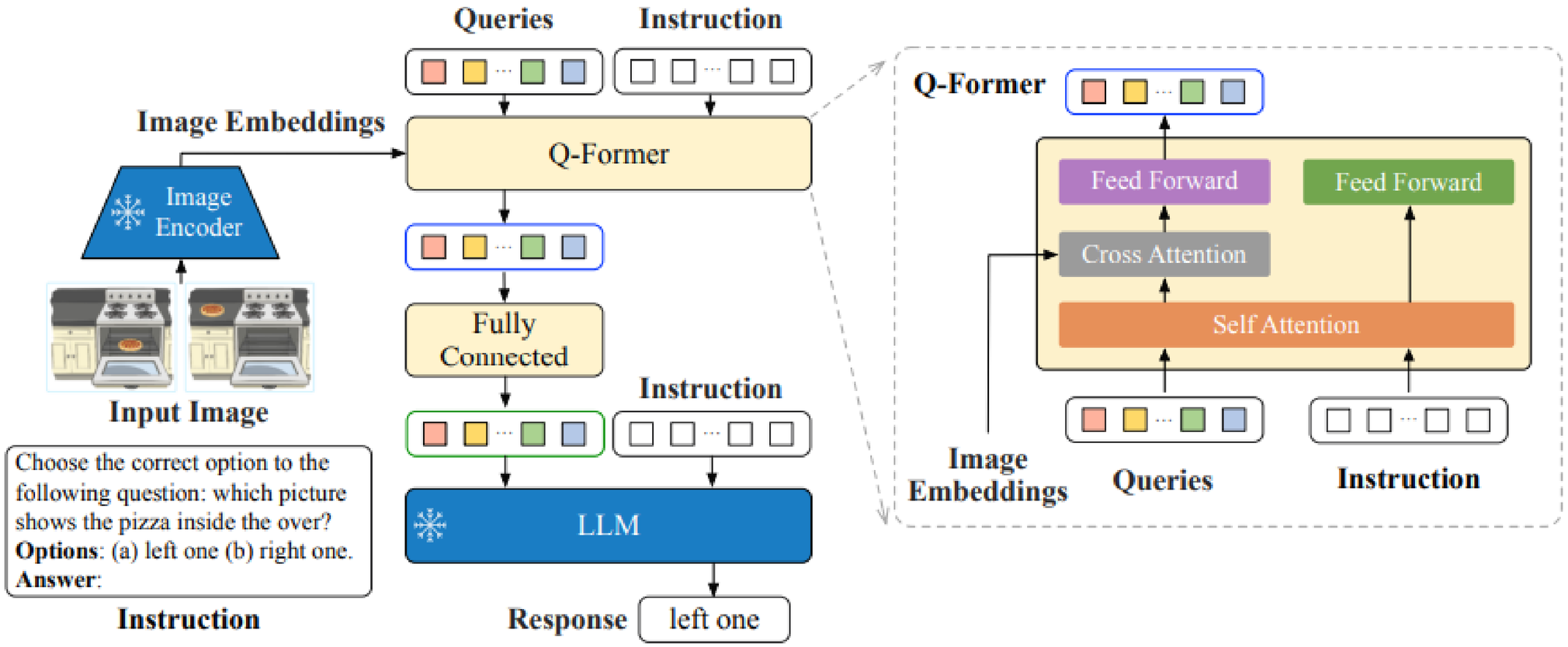
Contextual **cognizance**

Model Pipeline



End to end model pipeline proposed for finetuning the InstructBLIP model using M3L.

Why InstructBLIP?



InstructBLIP's standout feature lies in its pioneering **Instruction-Aware Visual Feature Extraction mechanism**, setting it on a trajectory for a more insightful understanding of user interactions and responses in social media.

arXiv:2305.06500 [cs.CV]

Video Frame Retriever

Duration :60 seconds

Frame rate : 30 frames per second

Total media volume = **1800 frames**

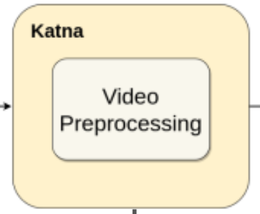
An 85-year-old primary school in Shanghai has been lifted off the ground — in its entirety — and relocated using new technology dubbed the "walking machine."

Sufficient to provide contextual information !

Video Frame Retriever

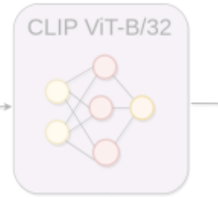
1 The pipeline consists of a **two-step distillation process**, that includes sampling frames at regular intervals

2 Then, we leverage the Katna library to extract the top **50 keyframes** from the sampled frames.

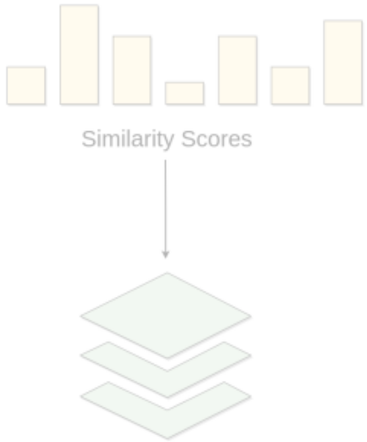


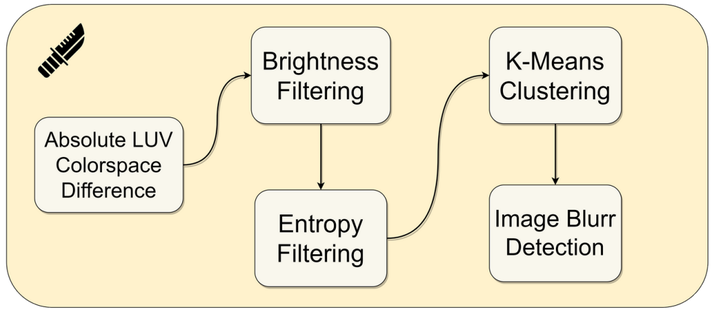
K-Means Clustering

Key Frames



Prompt



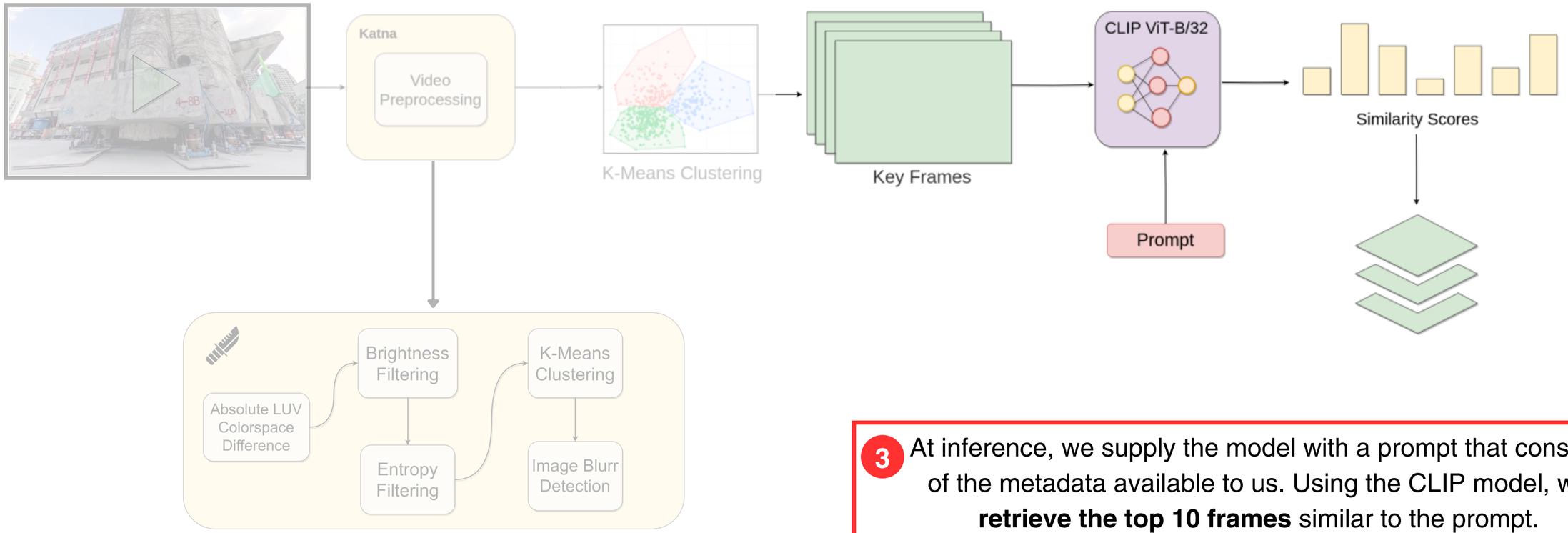


3 At inference, we supply the model with a prompt that consists of the metadata available to us. Using the CLIP model, we retrieve the top 10 frames similar to the prompt.

Video Frame Retriever

1 The pipeline consists of a two-step distillation process, that includes sampling frames at regular intervals

2 Then, we leverage the Katna library to extract the top 50 keyframes from the sampled frames.



3 At inference, we supply the model with a prompt that consists of the metadata available to us. Using the CLIP model, we **retrieve the top 10 frames** similar to the prompt.

Training Procedure

True prompt :- “An 85-year-old primary school in Shanghai has been lifted off the ground — in its entirety — and relocated using new technology dubbed the "walking machine.”

Video Preprocessing

Top 10 Retrieval

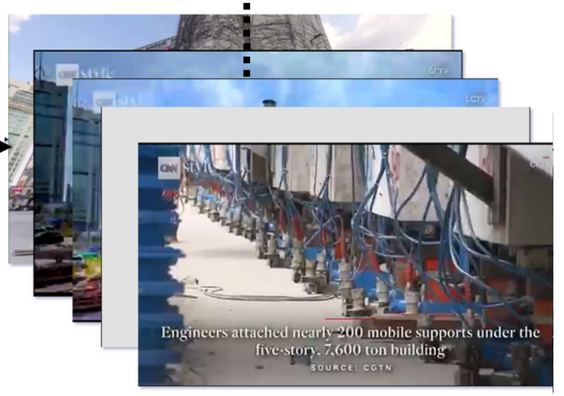
CLIP ViT-B/32

Prompt

"Ground Truth" frames

Backpropagation

Loss



"Predicted" frames

Meta prompt:- “Tweet date and time : 2020-11-15 16:01:18 , likes : 525, username : cnni. “

Video Frame Retriever

Training Procedure

Custom Loss Function **L for optimising model parameters**

$$DFRloss\left(q_i, fp_i^+, fn_{i,1}^-, \cdots, fn_{i,n}^-\right) = -\log \frac{e^{\mathrm{sim}\left(q_i, fp_i^+\right)}}{e^{\mathrm{sim}\left(q_i, fp_i^+\right)} + \sum_{j=1}^{n} e^{\mathrm{sim}\left(q_i, fn_{i,j}^-\right)}}$$

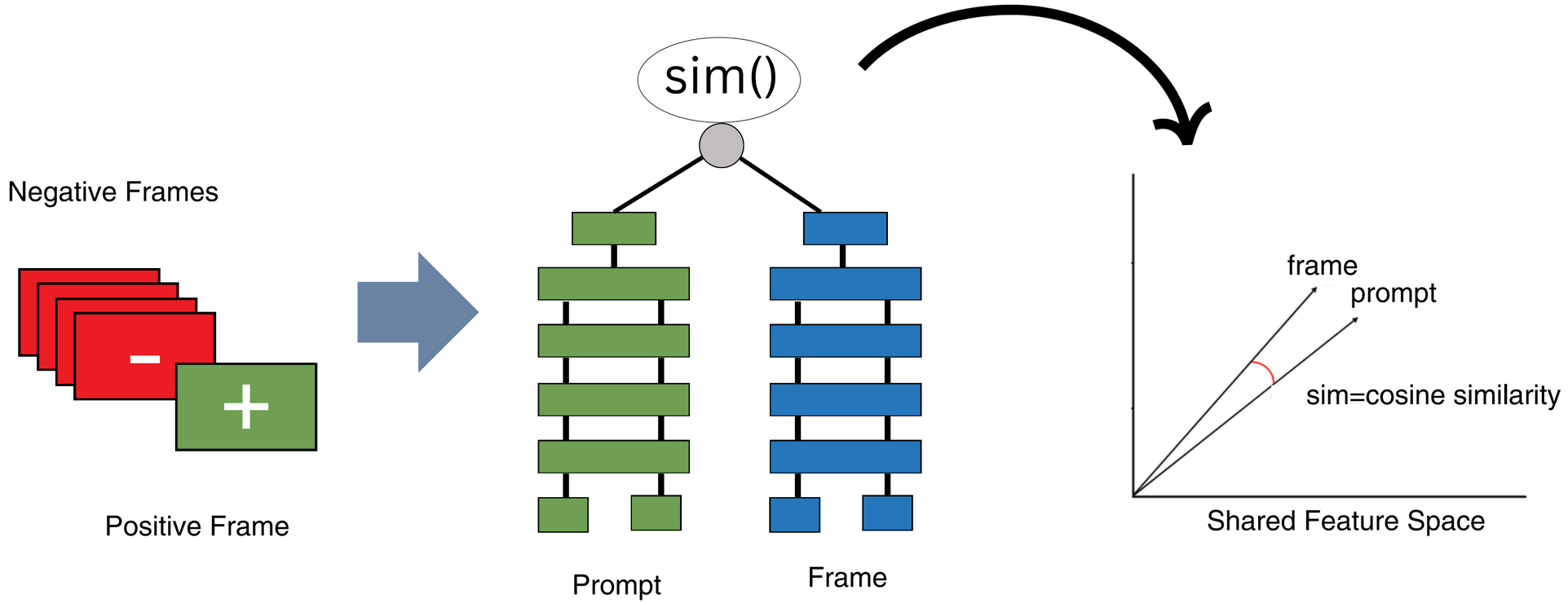
$$L = 0.1 \times BCEloss + 0.9 \times DFRloss$$

$$BCEloss = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right]$$

Training Procedure

$$DFRloss\left(q_i, fp_i^+, fn_{i,1}^-, \cdots, fn_{i,n}^-\right) = -\log \frac{e^{\mathrm{sim}\left(q_i, fp_i^+\right)}}{e^{\mathrm{sim}\left(q_i, fp_i^+\right)} + \sum_{j=1}^{n} e^{\mathrm{sim}\left(q_i, fn_{i,j}^-\right)}}$$

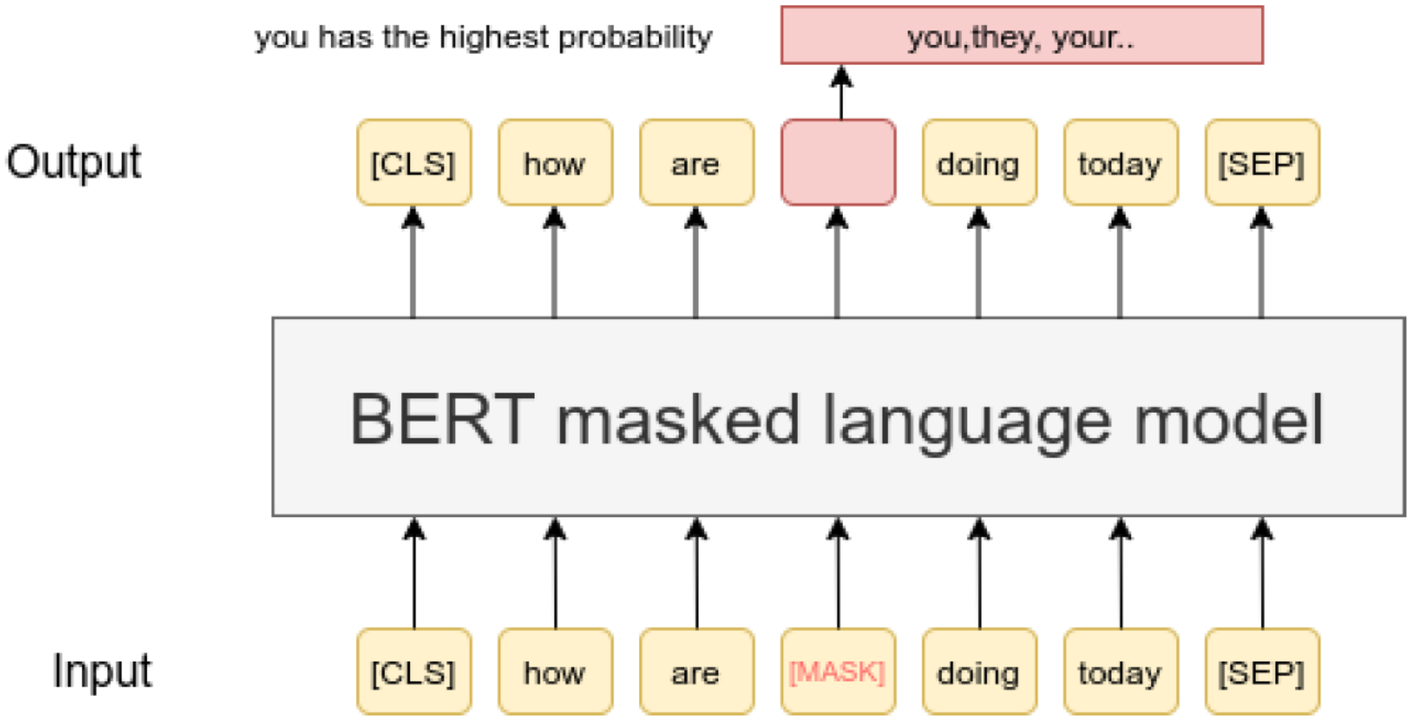
Inspired by the loss function presented in ***"Dense Passage Retrieval for Open-Domain Question Answering"***



arXiv:2004.04906 [cs.CL]

Masked Language Modelling

The "Masked Language Modelling" (MLM) technique involves randomly **masking** a small percentage of input tokens in sequences to predict these hidden tokens by feeding their final representations into an output softmax.



This approach thus concentrates solely on predicting masked words rather than reconstructing the entire input.

* arXiv:1810.04805

Training InstructBLIP using Multi-Modal Masked Learning

We carry out MLM on Multimodal data in our M3L approach for fine-tuning the InstructBLIP model for content and behaviour simulation to enhance the model's understanding of human behavior by providing it context-aware representations by predicting masked elements within the text.

Prompt for Content Simulation using M3L

You are content creating wizard. Today I challenge you to predict a caption for an post that was posted in the past. I'll provide you with the following information regarding the post:
Username, Date of post release, Time of post release and No of likes
Review the username and date as there could be some insights relevant for the caption. Identify the material/person/objects in the image and align the caption accordingly. This will help to predict a better caption.
Now given the image,
Username is MayoClinic.,
Release Date is 2020-07-19,
Release Time is 00:07:00,
Likes received is 78.
Predicted Caption: [MASK]

Training InstructBLIP using Multi-Modal Masked Learning

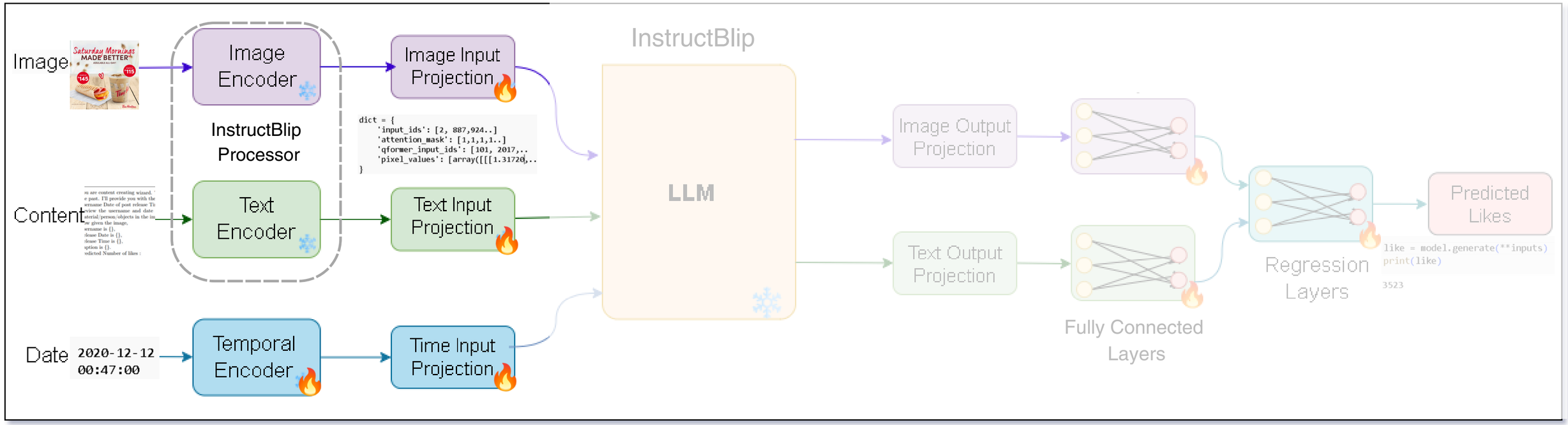
In the prompt for behavior simulation we mask the captions modality and ask the model to predict the same given the other modalities as depicted in the prompt below.

Prompt for Behavior Simulation task using M3L

You are content creating wizard. Today I challenge you to predict a the number of likes for a post that was posted in the past. I'll provide you with the following information regarding the post:
Username Date of post release Time of post release Caption
Review the username and date as there could be some insights relevant for the number of likes. Identify the material/person/objects in the image and obtain the number of likes. . This will help to predict the number of likes.
Now given the image,
Username is MayoClinic.,
Release Date is 2020-07-19,
Release Time is 00:07:00,
Caption is Many people are understandably anxious to get outdoors in light of quarantines. Do it safely with these tips for outdoor activities: <hyperlink> COVID19 <hyperlink>.
Predicted Number of likes : [MASK]

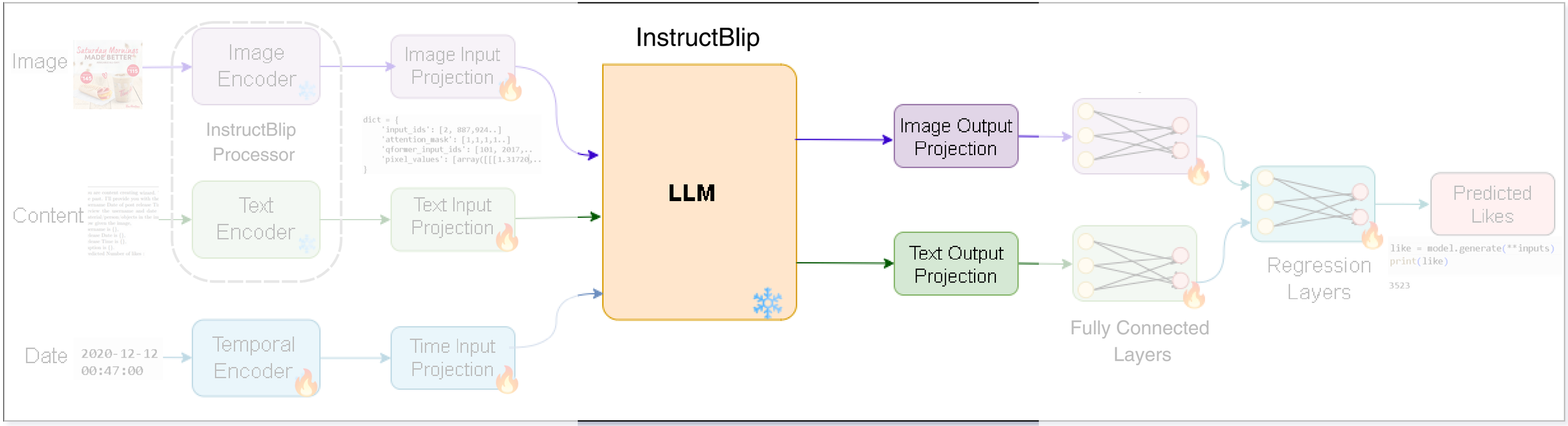
Training for the Behaviour Simulation Task

1. The text modality and image modality is processed by InstructBlip processor which consists of
   a. BLIP as image encoder and
   b. Llama/t-5 tokenizer as text encoder.



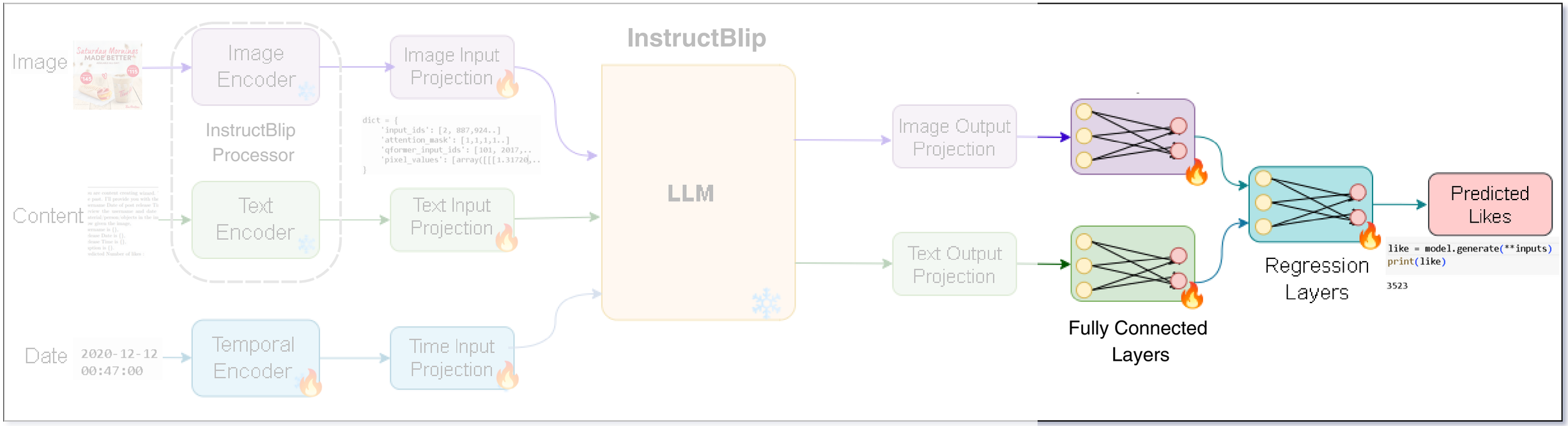
Training for the Behaviour Simulation Task

2 From the output of the model, last hidden states is extracted for both text and image encodings



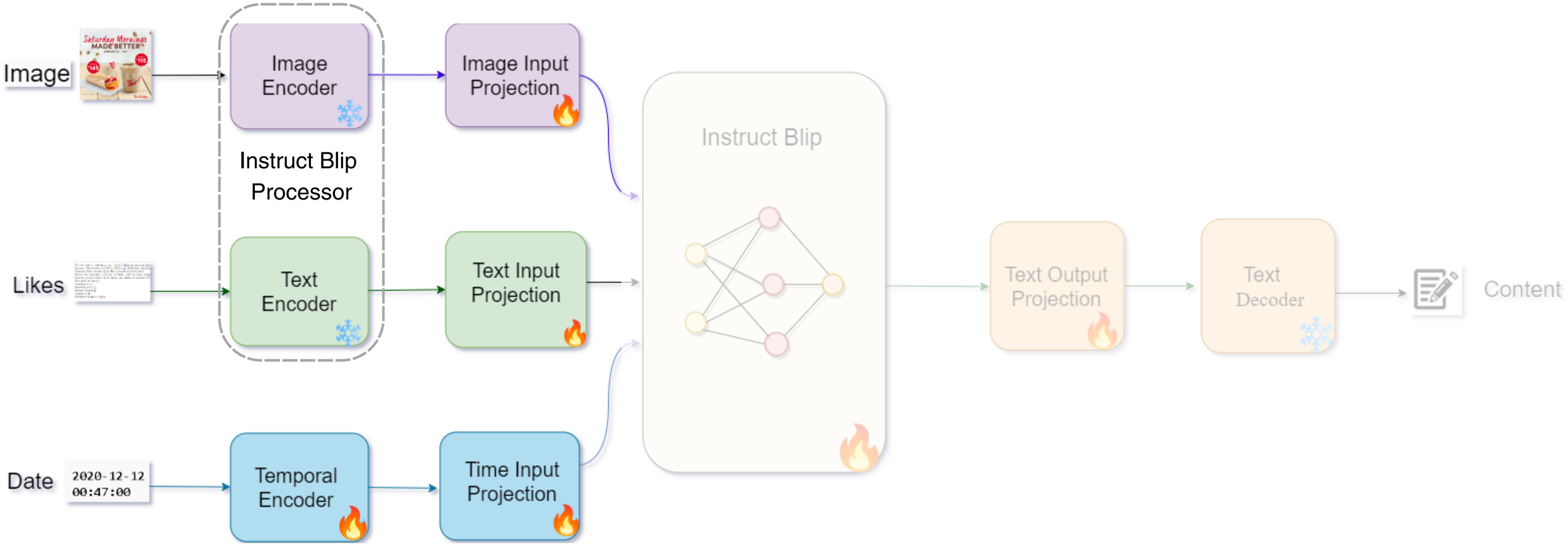
Training for the Behaviour Simulation Task

3 These encodings are passed through various connected and regressor layers before prediction of final output.



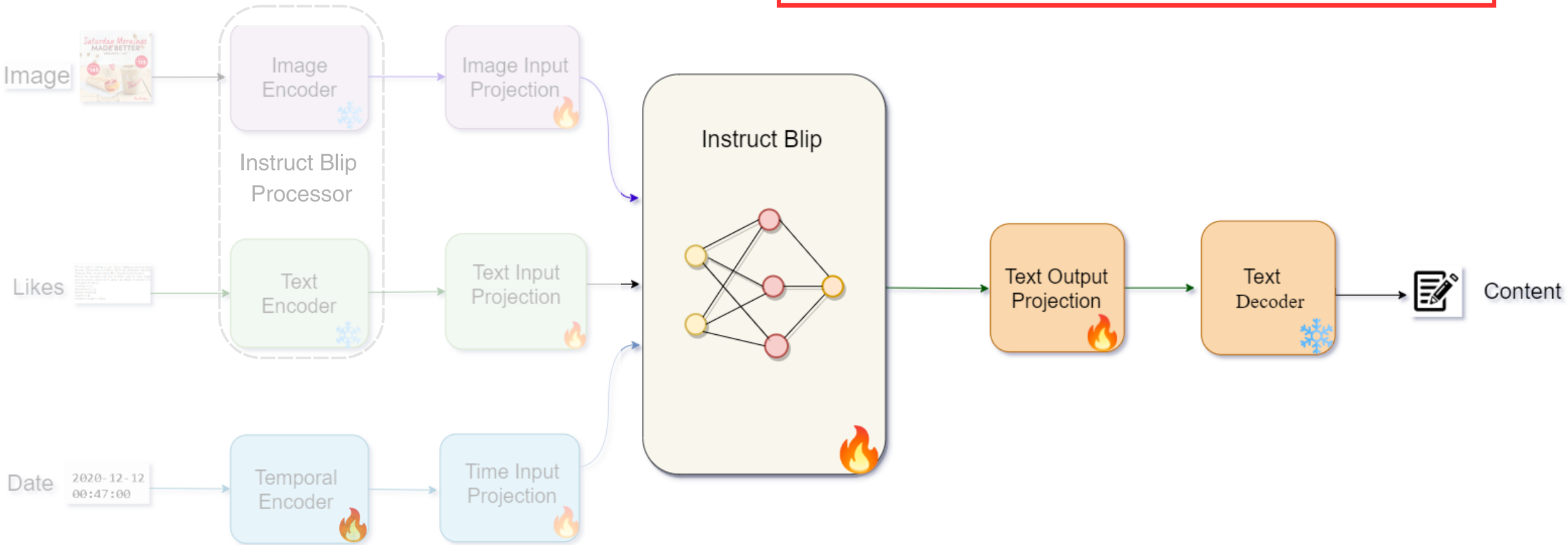
Training for the Content Simulation Task

1. The text modality and image modality is processed by InstructBlip processor which consists of
   a. BLIP as image encoder and
   b. Llama/t-5 tokenizer as text encoder.



Training for the Content Simulation Task

2 We derive outputs from the language output head of the InstructBLIP Vicuna model and pass it through a batch decoder to get the predicted captions



Training Setup

Algorithm 1 Masked Vision Language Modelling for Behaviour Simulation

InstructBLIP Model (M):
    `M` ← InstructBLIP Model
    `M_Q` ← Q-Former part of `M`
    `M_T` ← T-Former part of `M` processes temporal information.

Initialization:
`TWEET_METADATA` ← Extract tweet metadata from D_{train}
`LIKES_PROMPT` ← Initialize likes_prompt
`CAPTIONS_PROMPT` ← Initialize captions_prompt
`SHARED_FEATURES` ← Propagate batches through shared base layers
`T_FORMER` ← Create Temporal-Former based on `Q_FORMER`

Threshold Constraint: $0 \leq THRESHOLD \leq 100$

procedure MaskedVisionLM($D_{train}, EPOCHS$)
    for *epoch* ← 1 **to** $EPOCHS$ **do**
        for each batch in D_{train} **do**
            `BATCH_CONTENT` ← `LIKES_PROMPT`
            `BATCH_BEHAVIOR` ← `CAPTIONS_PROMPT`

            if THRESHOLD $>$ 50 **then**
                `LOSS` ← Negative log likelihood loss for caption generation
            else
                `LOSS` ← MSE loss for likes generation
            end if

            Backpropagate on the loss and update model parameters θ
        end for
    end for
end procedure

$$MSEloss = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$NLLloss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i)]$$

How does InstructBLIP learn to model behaviour?

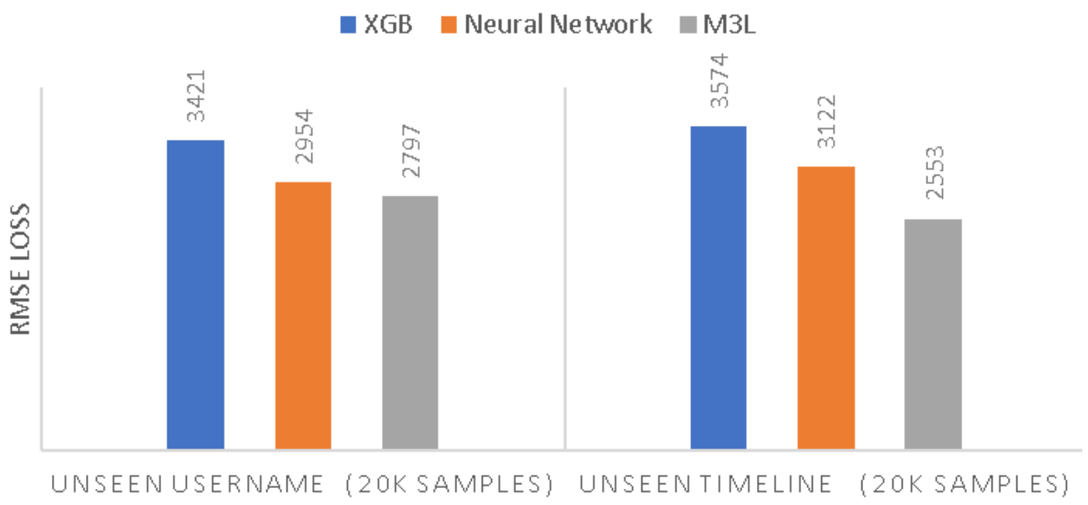
Training for Behaviour Simulation + **Training for Content Simulation**

Behaviour Optimized InstructBLIP

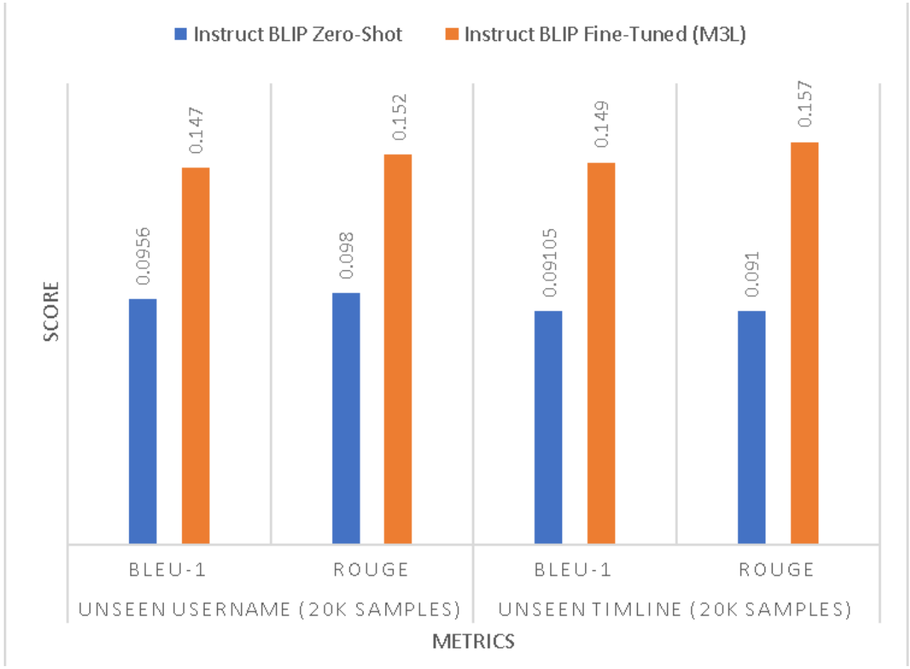
The multitasking fine-tuning with a custom loss in InstructBLIP refines caption generation aligned with human preferences and predicts engagement through nuanced cues.

Model Performance

Behaviour Simulation



Content Simulation



*Results evaluated on training data provided by Adobe for the Behaviour Simulation Challenge

Demo Video of Inference

We can further extend our model and train it for multiple tasks to help it inculcate a deeper knowledge of human behaviour and equip it to predict more output modalities.