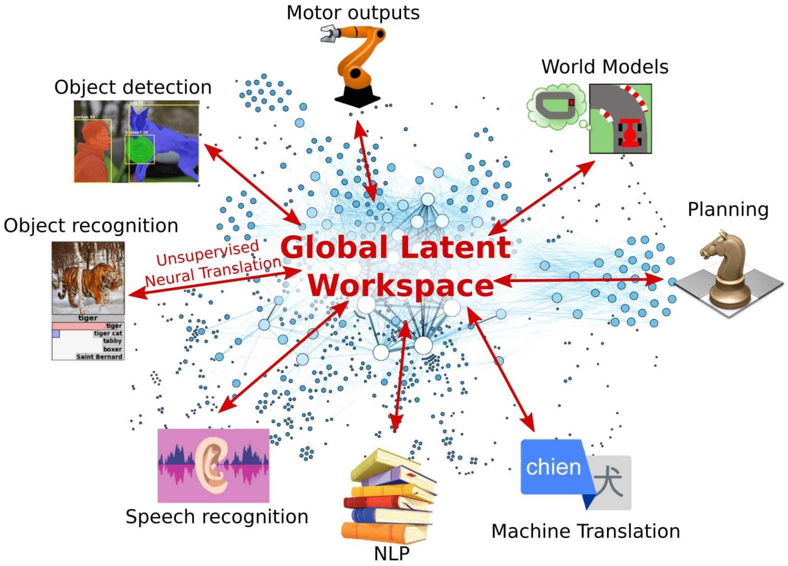
Semi-supervised multimodal representation learning through a global workspace

Heterogeneous Data and Large Representation Models in Science Workshop

Introduction

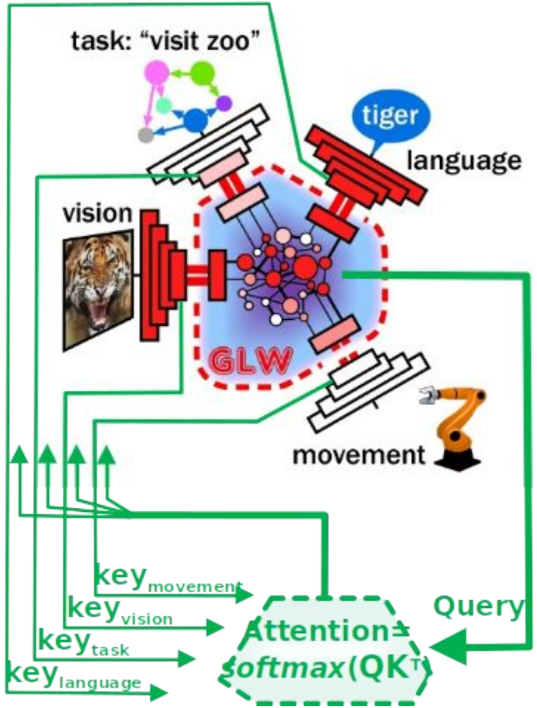


?

Introduction

First work focuses on semi-supervision and how to reduce the amount of multimodal data in training

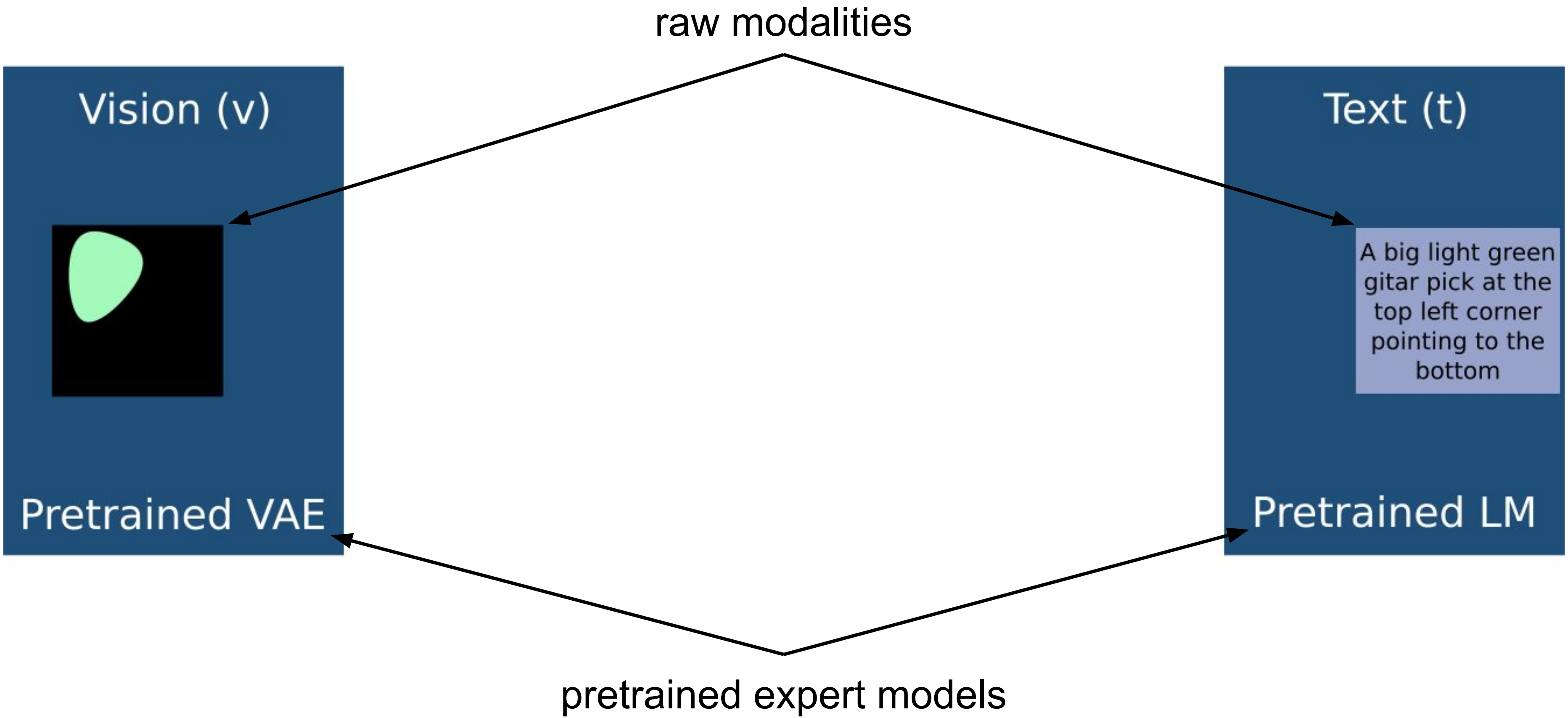
Second work focuses on how to improve this model and make it closer to the Global Workspace Theory



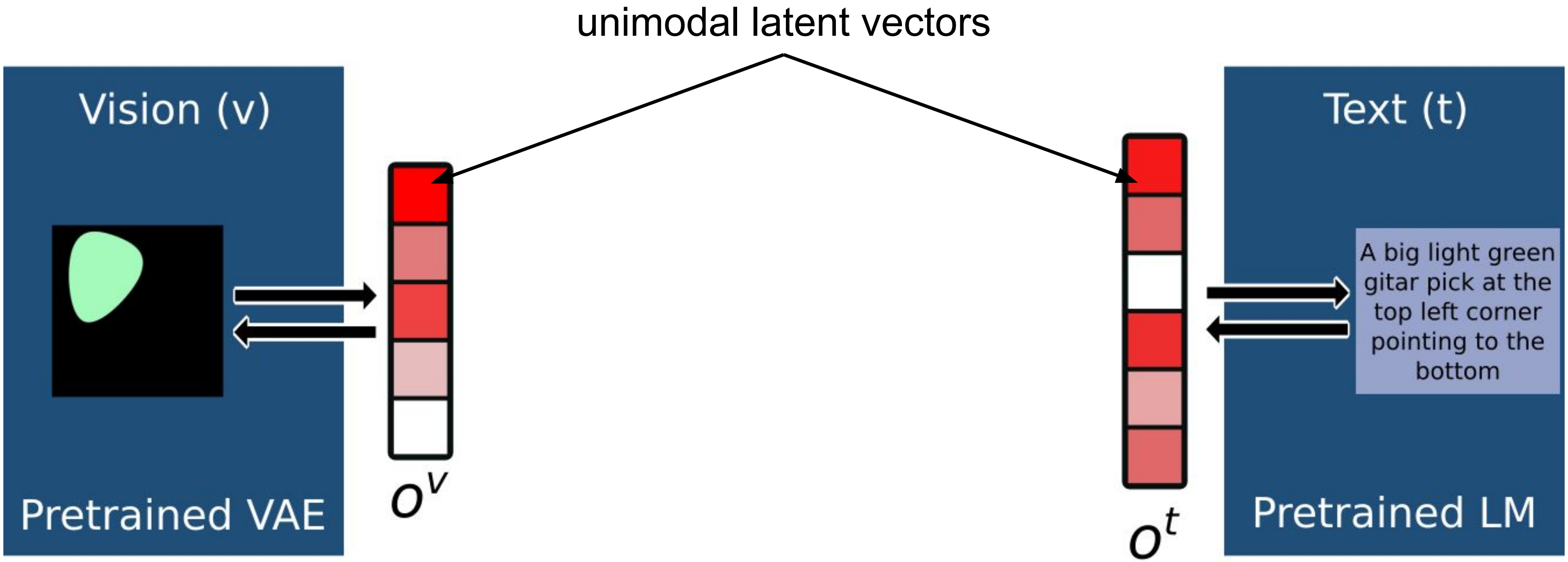
Semi-Supervised Multimodal Representation Learning

Semi-supervised Multimodal Representation Learning through a Global Workspace,
Devillers; Maytié; VanRullen

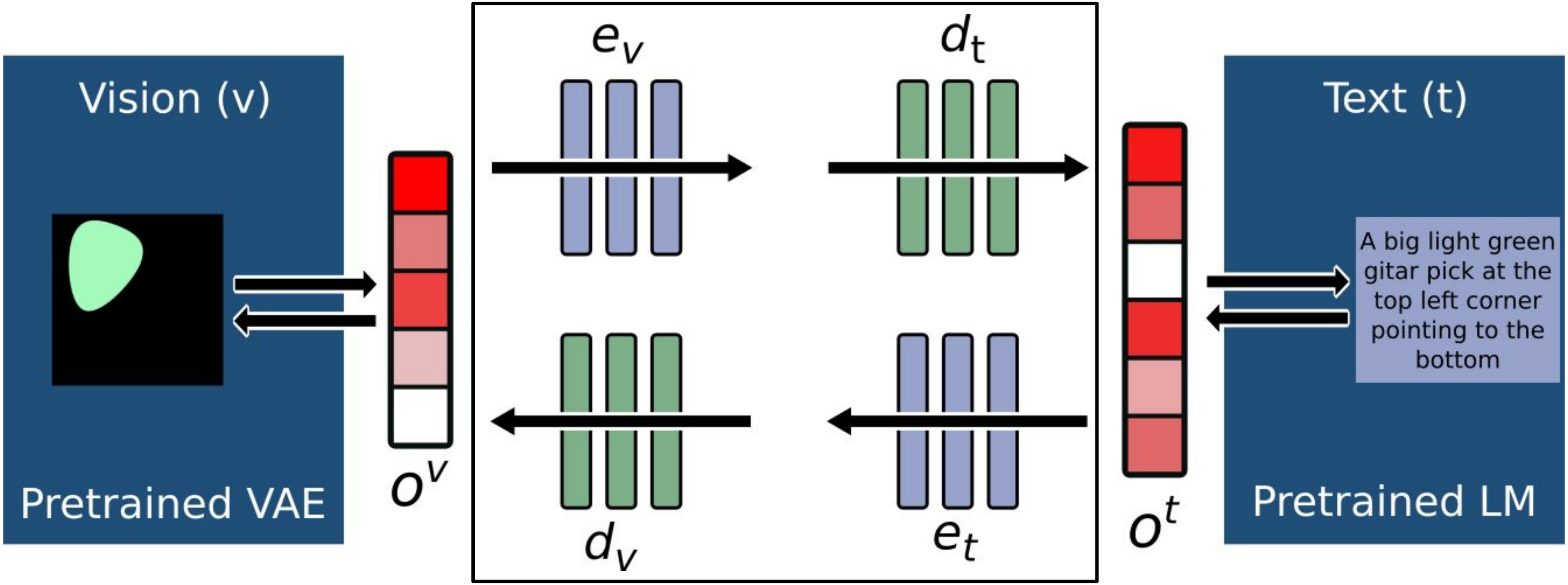
Semi-Supervised Multimodal Representation Learning



Semi-Supervised Multimodal Representation Learning

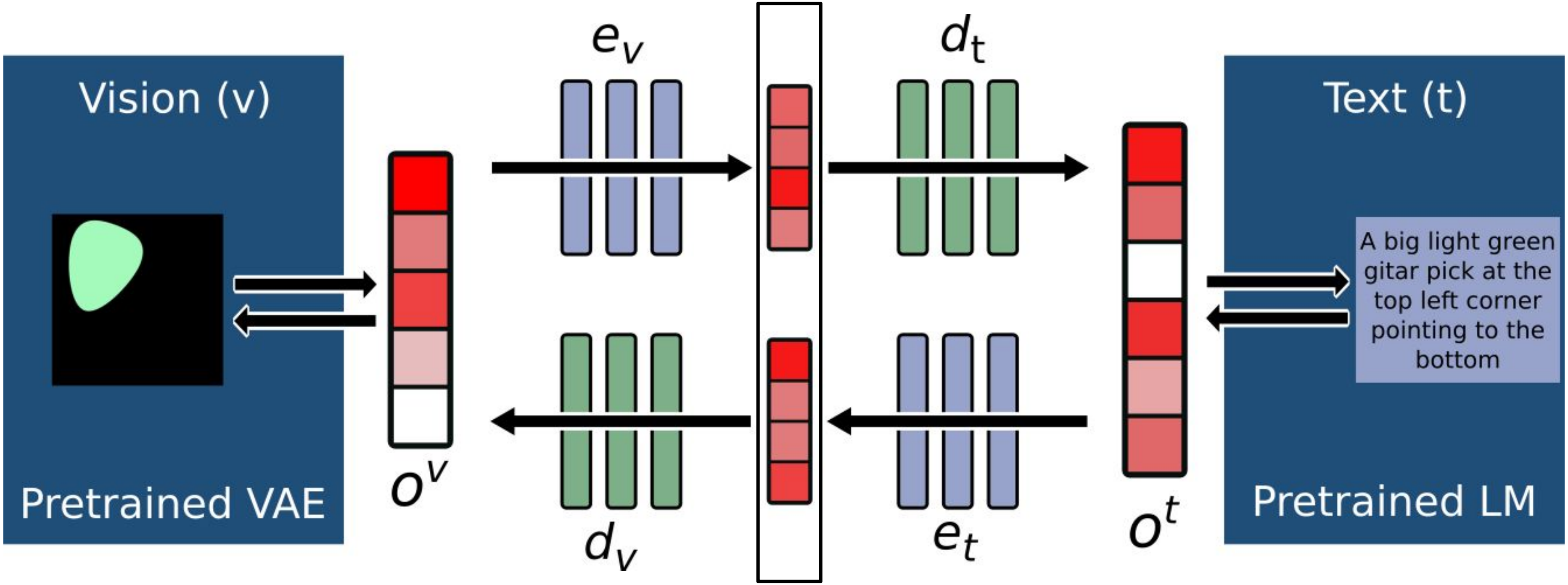


Semi-Supervised Multimodal Representation Learning



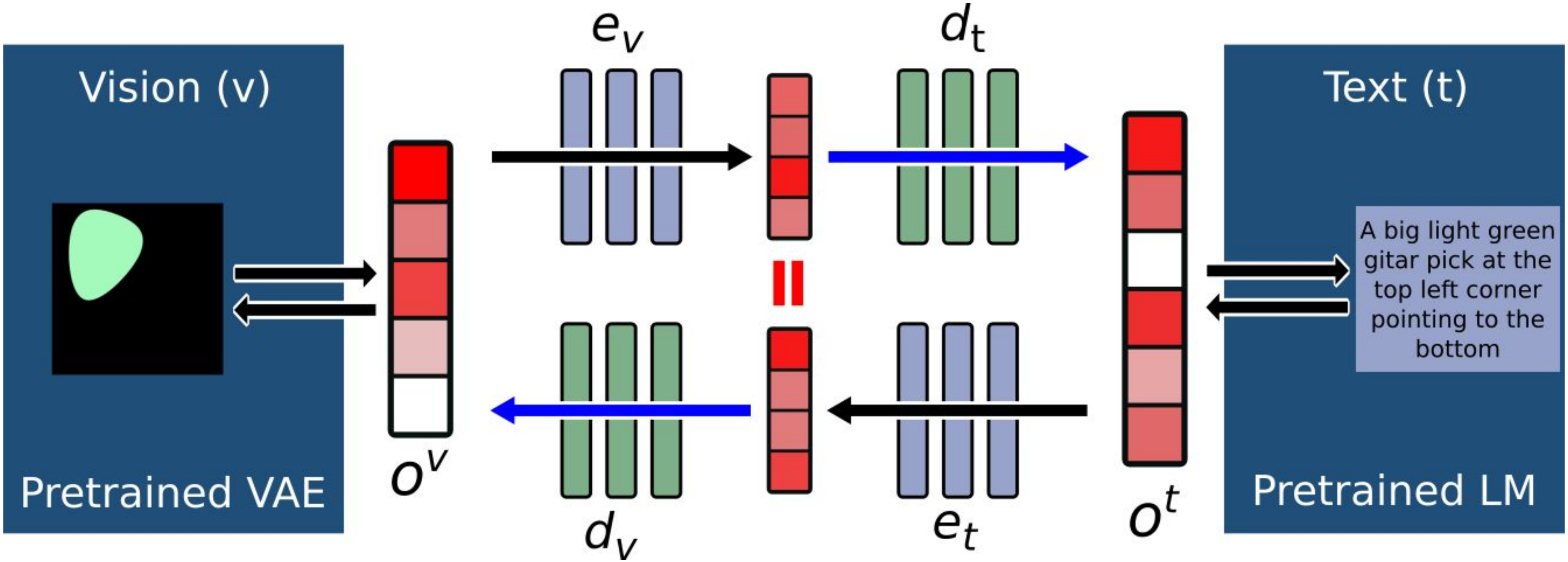
Set of encoders and decoders to …

Semi-Supervised Multimodal Representation Learning



… link unimodal vectors to the Global Workspace

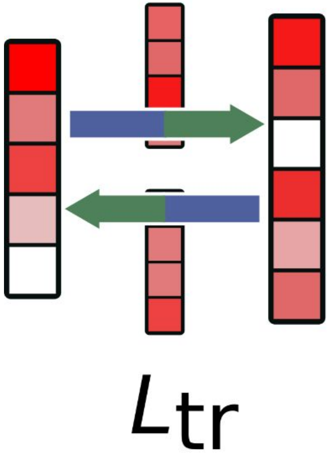
Semi-Supervised Multimodal Representation Learning



2 important properties:

- **alignment** : align representations from both modalities
- **broadcast** : capable to translate information from the GW back to each modality

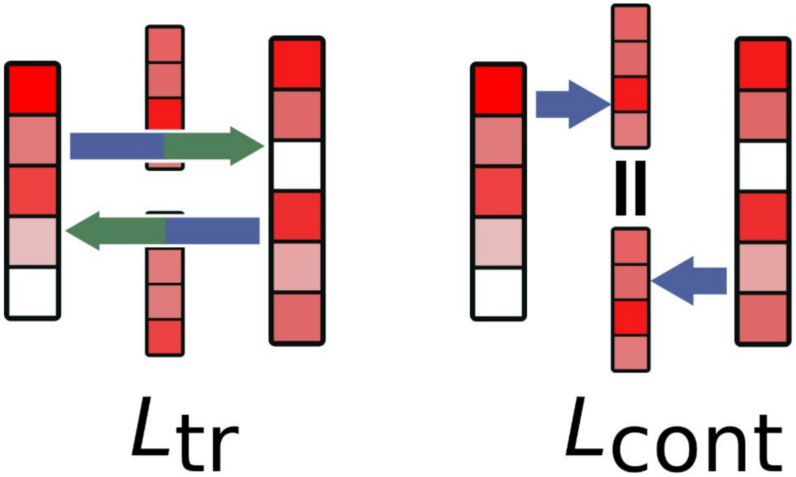
Semi-Supervised Multimodal Representation Learning



$$total\,loss = \boxed{\alpha \cdot translation\,loss}$$

supervised loss

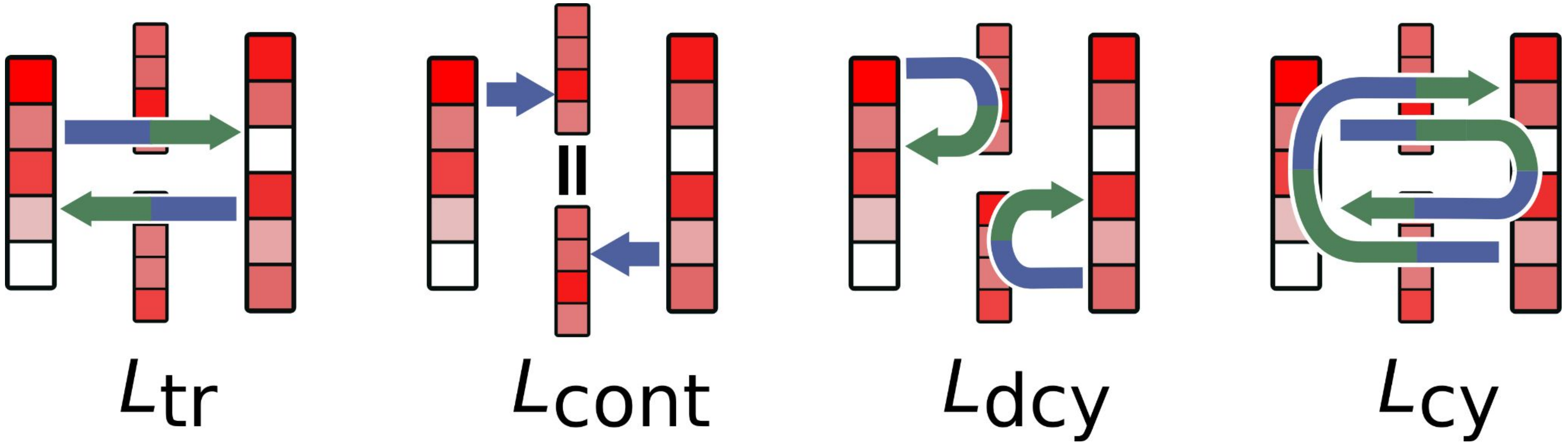
Semi-Supervised Multimodal Representation Learning



$$total\,loss = \boxed{\alpha \cdot translation\,loss + \beta \cdot contrastive\,loss}$$

supervised loss

Semi-Supervised Multimodal Representation Learning



$$total\,loss = \underbrace{\alpha \cdot translation\,loss + \beta \cdot contrastive\,loss}_{\text{supervised loss}} + \underbrace{\gamma \cdot demi-cycle\,loss + \theta \cdot cycle\,loss}_{\text{unsupervised loss}}$$

Semi-Supervised Multimodal Representation Learning

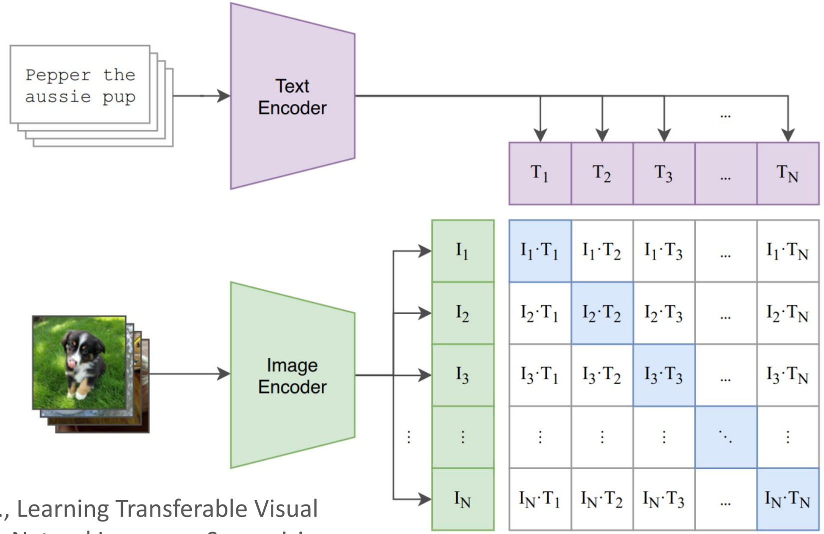
Using different combination of losses → different models

Model Properties	Semi-supervision	"Global Workspace" (Alignment + Broadcast)
L_{tr}	$-$	$-$
$L_{\mathrm{tr}} + L_{\mathrm{cont}}$	$-$	$++$
$L_{\mathrm{tr}} + L_{\mathrm{cy}}$	$++$	$-$
$L_{\mathrm{tr}} + L_{\mathrm{cont}} + L_{\mathrm{dcy}} + L_{\mathrm{cy}}$	$+++$	$+++$
L_{cont}	$-$	$\pm$ (no broadcast)
$L_{\mathrm{tr}} + L_{\mathrm{dcy}}$	$+$	$+$

Semi-Supervised Multimodal Representation Learning

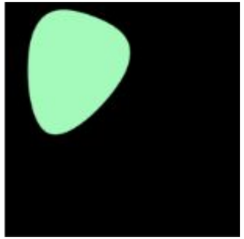
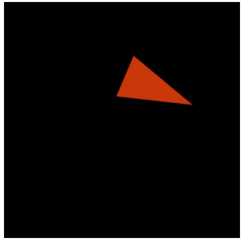
Model Properties	Semi-supervision	"Global Workspace" (Alignment + Broadcast)
L_{tr}	$-$	$-$
$L_{\text{tr}} + L_{\text{cont}}$	$-$	$++$
$L_{\text{tr}} + L_{\text{cy}}$	$++$	$-$
$L_{\text{tr}} + L_{\text{cont}} + L_{\text{dcy}} + L_{\text{cy}}$	$+++$	$+++$
L_{cont}	$-$	$\pm$ (no broadcast)
$L_{\text{tr}} + L_{\text{dcy}}$	$+$	$+$

- Contrastive Global Workspace ≃ CLIP
- Contrastive Global Workspace ⟶ CLIP-like
- Used as a baseline



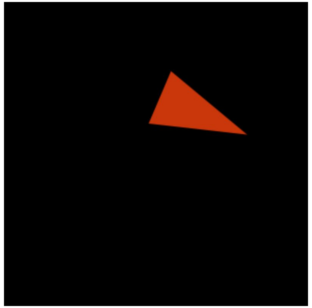
Radford & al., Learning Transferable Visual Models From Natural Language Supervision

Semi-Supervised Multimodal Representation Learning

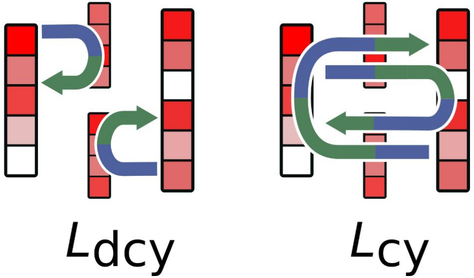
dataset / modalities	Simple Shapes 3 object type (triangle, egg, diamond) with 4 attributes (position, rotation, size, color)		Factory Table in a virtual environment with 3 different attributes (position, rotation, color)	
vision				
proto-language	1 (6, 6) 3.21 8 (0, 1, 0.3)	0 (18, 6) 2.89 5 (1, 0.2, 0.2)	(2,3) 3.12 (1, 0.1, 0.3)	(0,1) 3.35 (0.8, 0.7, 0.1)
language	A big light green egg shape pointing to the bottom at the top left corner	A medium red triangle at the middle right, pointing to the East	A red table close to the cones at the top pointing to the top right	An orange table at the middle of the scene pointing to the West

Semi-Supervised Multimodal Representation Learning

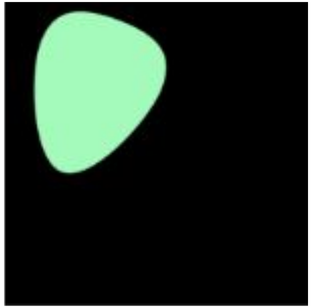
Unpaired data :



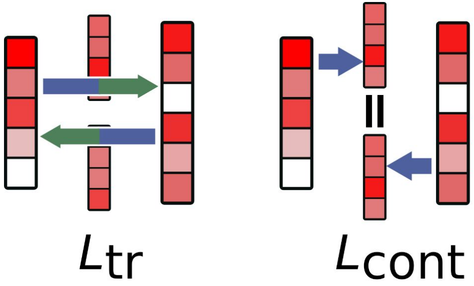
A medium red egg shape at the top right pointing to the East



Paired data :



A big light green egg shape pointing to the bottom at the top left corner

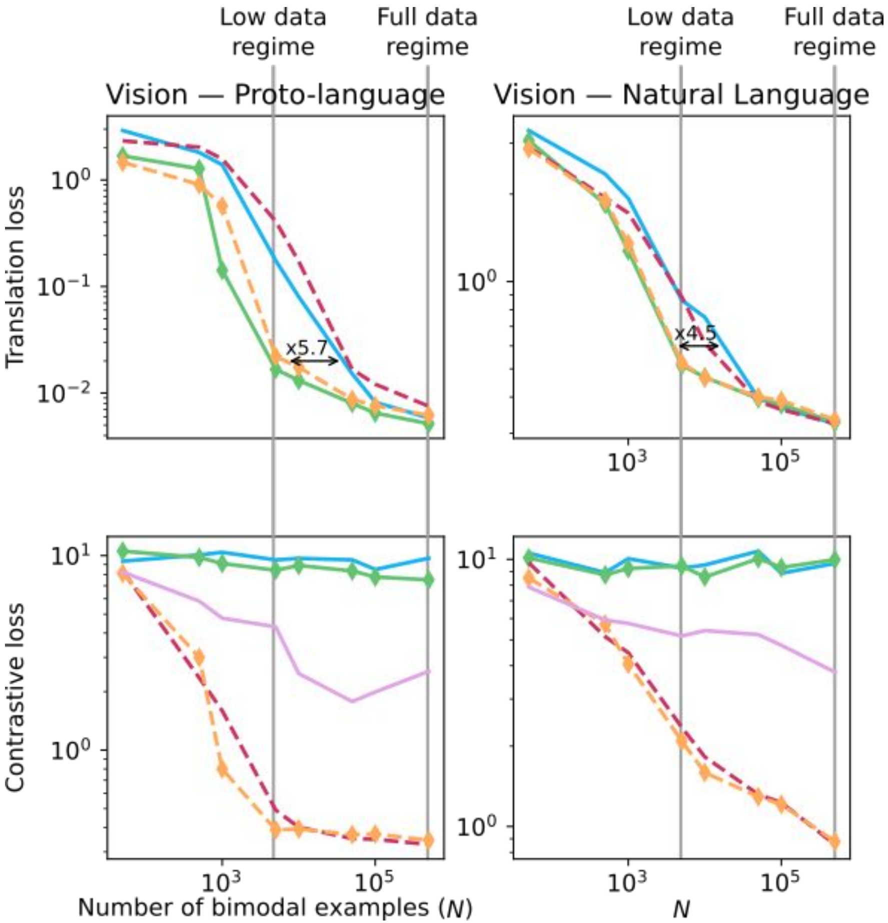


Semi-Supervised Multimodal Representation Learning

Simple Shapes Dataset

Influence of the number of paired data on translation and alignment performances

Model Properties	Semi-supervision	"Global Workspace" (Alignment + Broadcast)
L_{tr}	$-$	$-$
$L_{\text{tr}} + L_{\text{cont}}$	$-$	$++$
$L_{\text{tr}} + L_{\text{cy}}$	$++$	$-$
$L_{\text{tr}} + L_{\text{cont}} + L_{\text{dcy}} + L_{\text{cy}}$	$+++$	$+++$
L_{cont}	$-$	$\pm$ (no broadcast)
$L_{\text{tr}} + L_{\text{dcy}}$	$+$	$+$



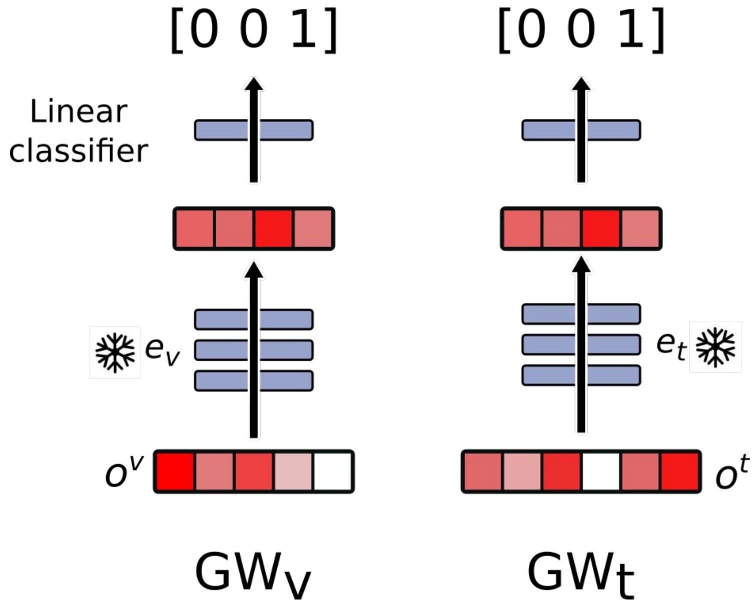
Semi-supervised Multimodal Representation Learning through a Global Workspace, Devillers & al.

Semi-Supervised Multimodal Representation Learning

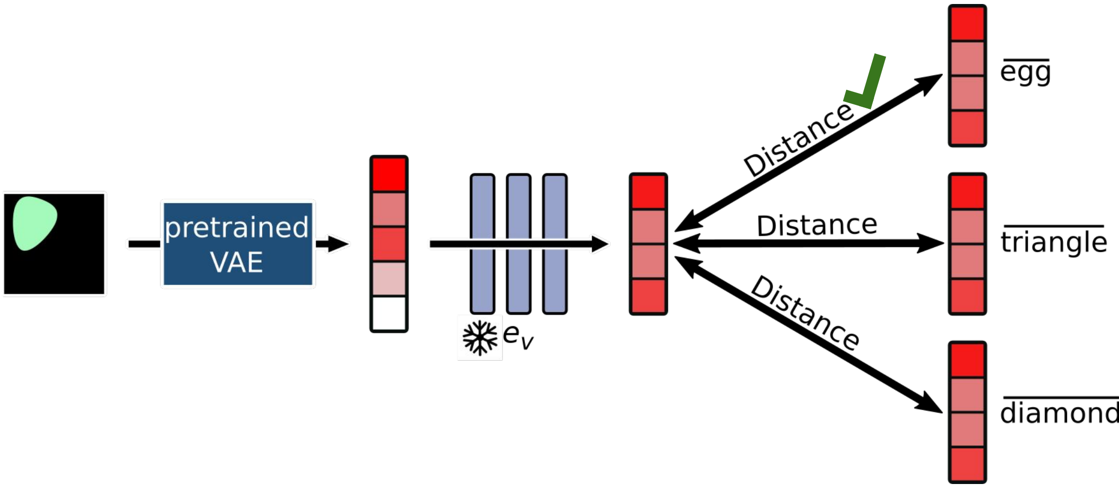
Downstream task : shape classification

Two different setup:

Linear probe using visual or textual representation



Zero-shot classification : comparing images representation to averaged vectors of each class

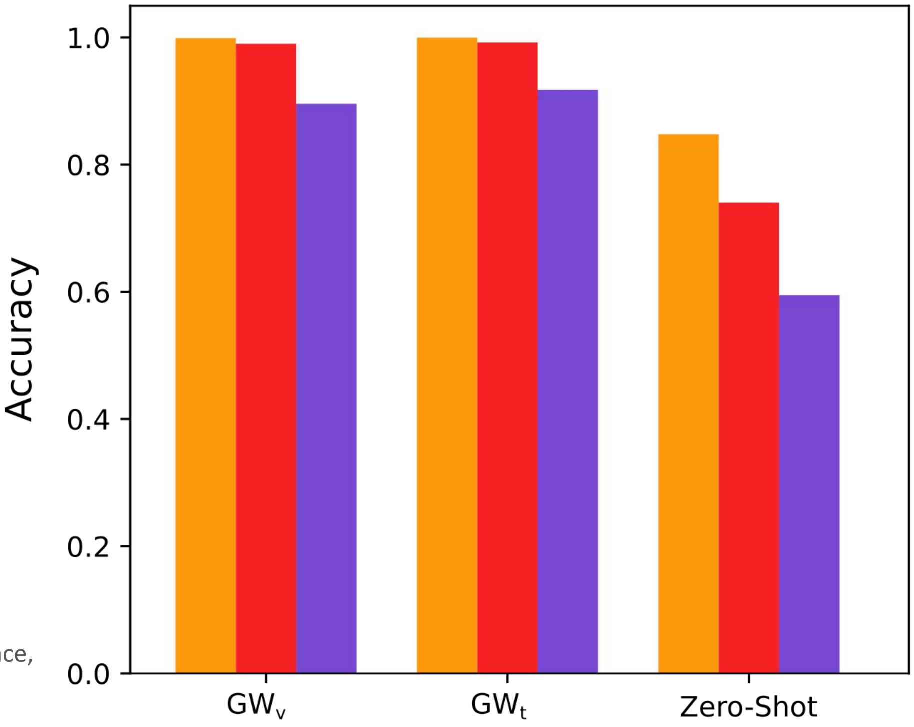


Semi-Supervised Multimodal Representation Learning

Downstream task : shape classification

Model Properties	Semi-supervision	"Global Workspace" (Alignment + Broadcast)
L_{tr}	$-$	$-$
$L_{\text{tr}} + L_{\text{cont}}$	$-$	$++$
$L_{\text{tr}} + L_{\text{cy}}$	$++$	$-$
$L_{\text{tr}} + L_{\text{cont}} + L_{\text{dcy}} + L_{\text{cy}}$	$+++$	$+++$
L_{cont}	$-$	$\pm$ (no broadcast)
$L_{\text{tr}} + L_{\text{dcy}}$	$+$	$+$

Linear probe and zero-shot performance on three-way shape classification



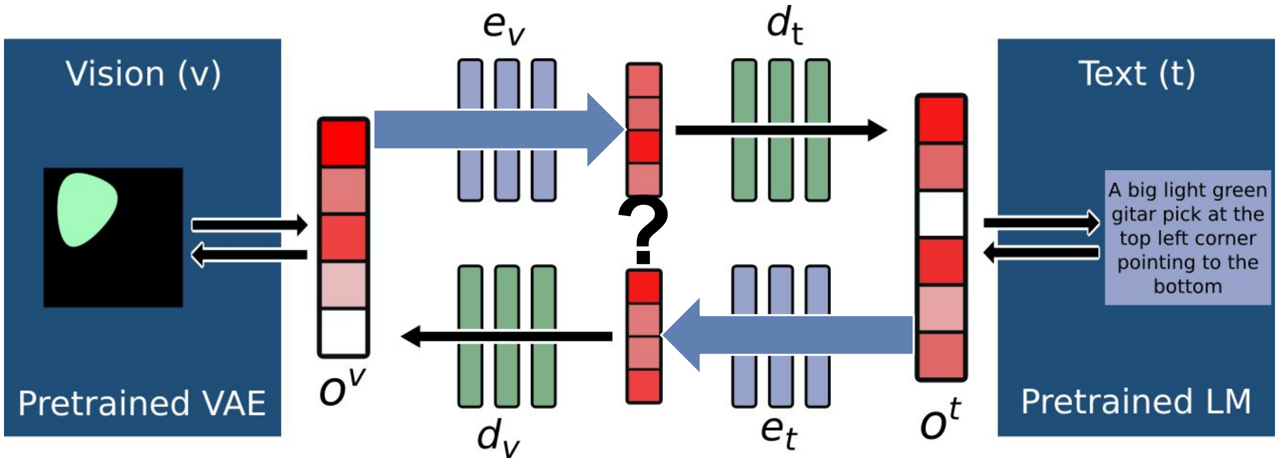
CLIP-like model (contrastive GW) performs always worse than Global Workspace models with broadcast in addition to alignment

Semi-supervised Multimodal Representation Learning through a Global Workspace, Devillers & al.

A Global Workspace with fusion

A Global Workspace with fusion

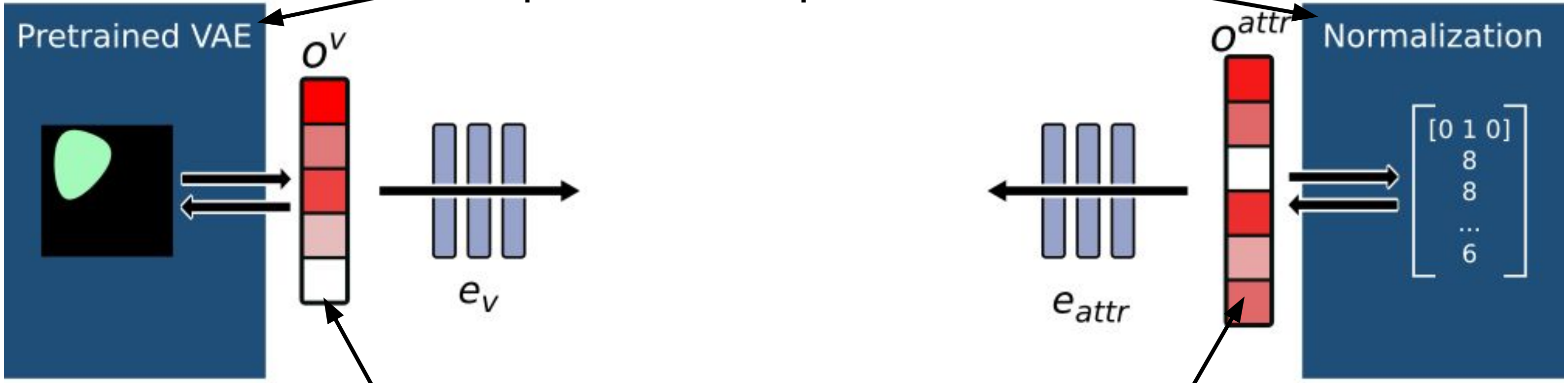
Current model: only one modality at a time entering inside the GW



Modify the model to encode multiple modalities at the same time → **Fusion**

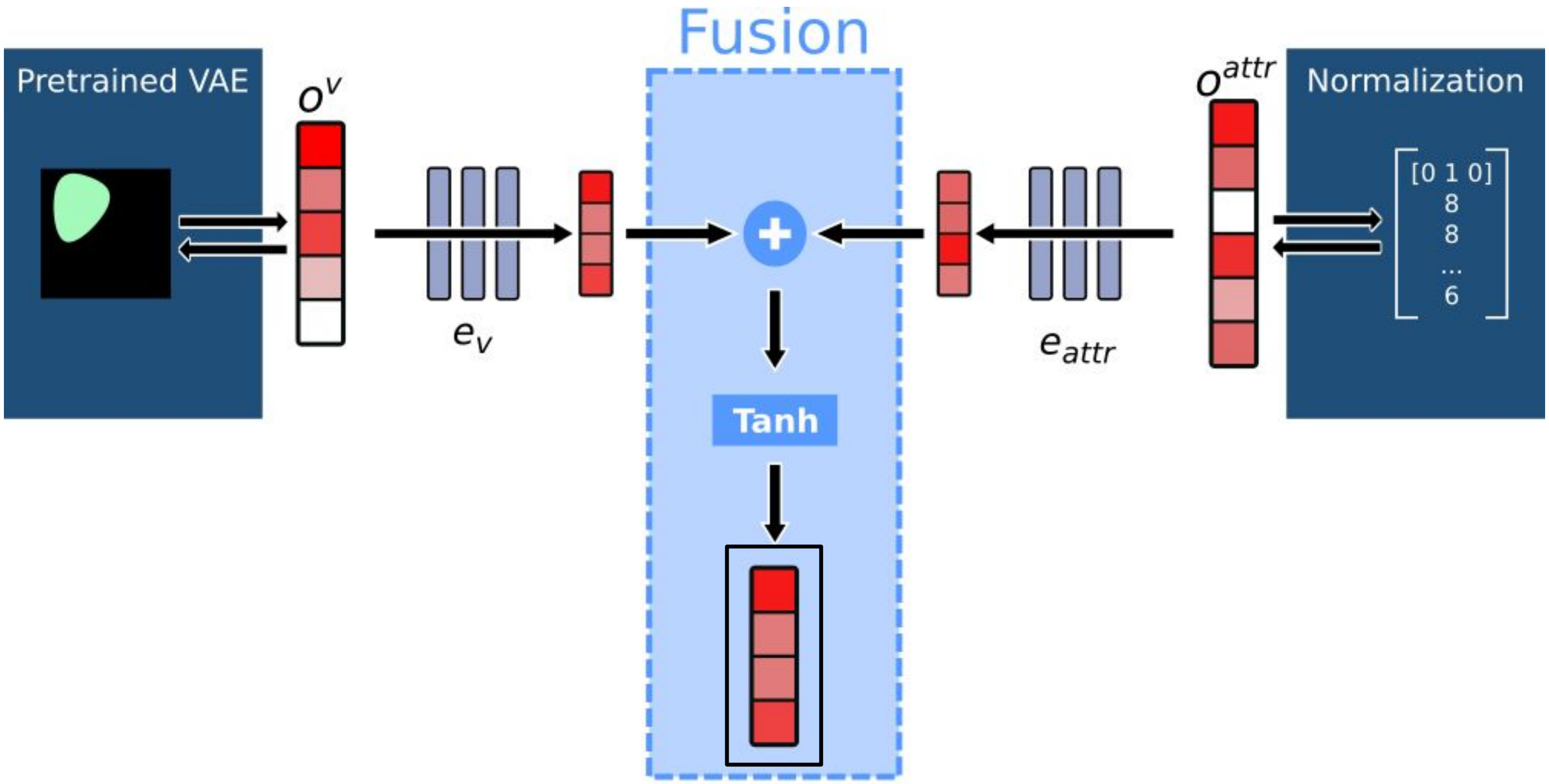
A Global Workspace with fusion

pretrained expert models



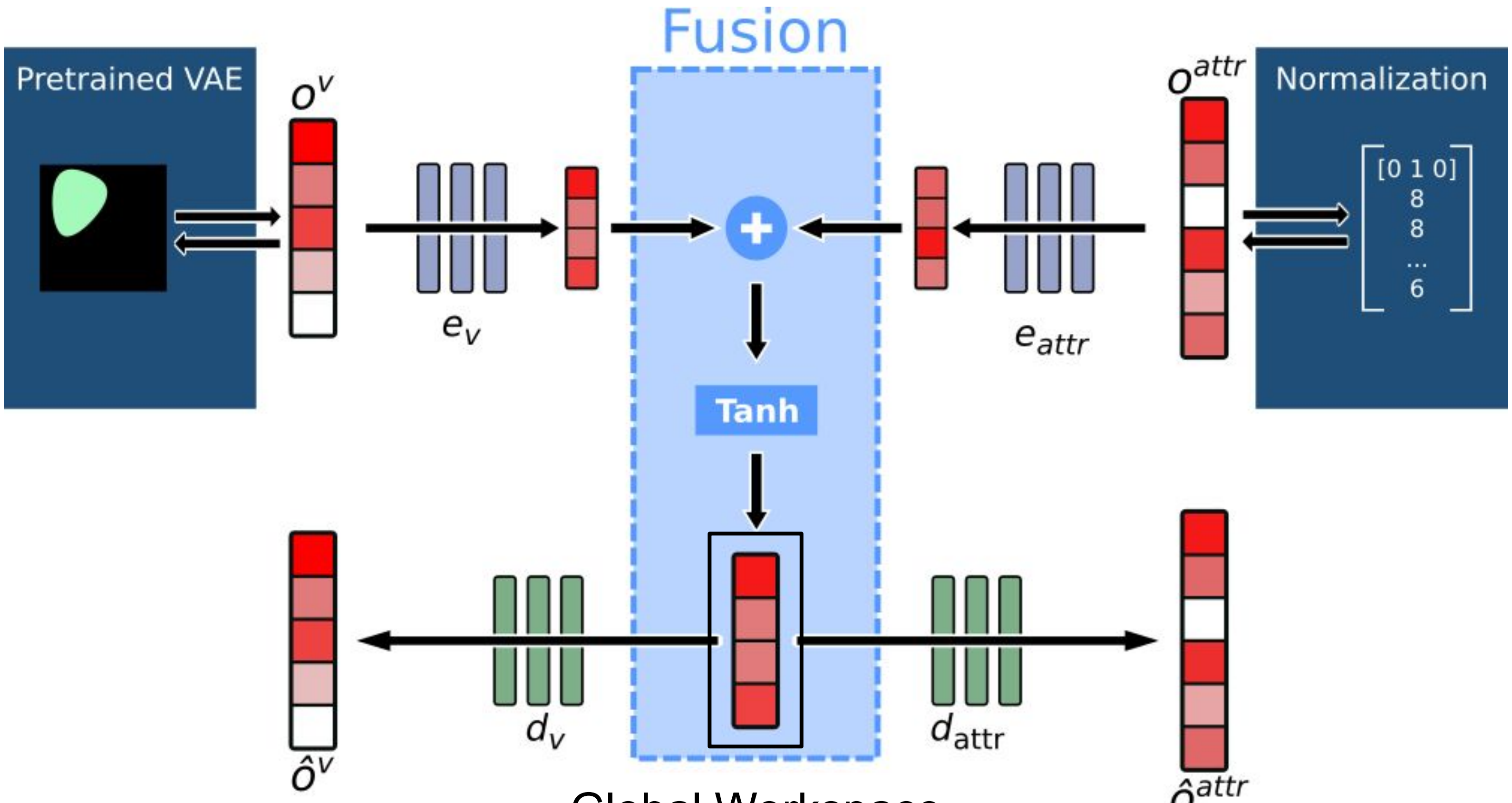
unimodal latent vectors

A Global Workspace with fusion



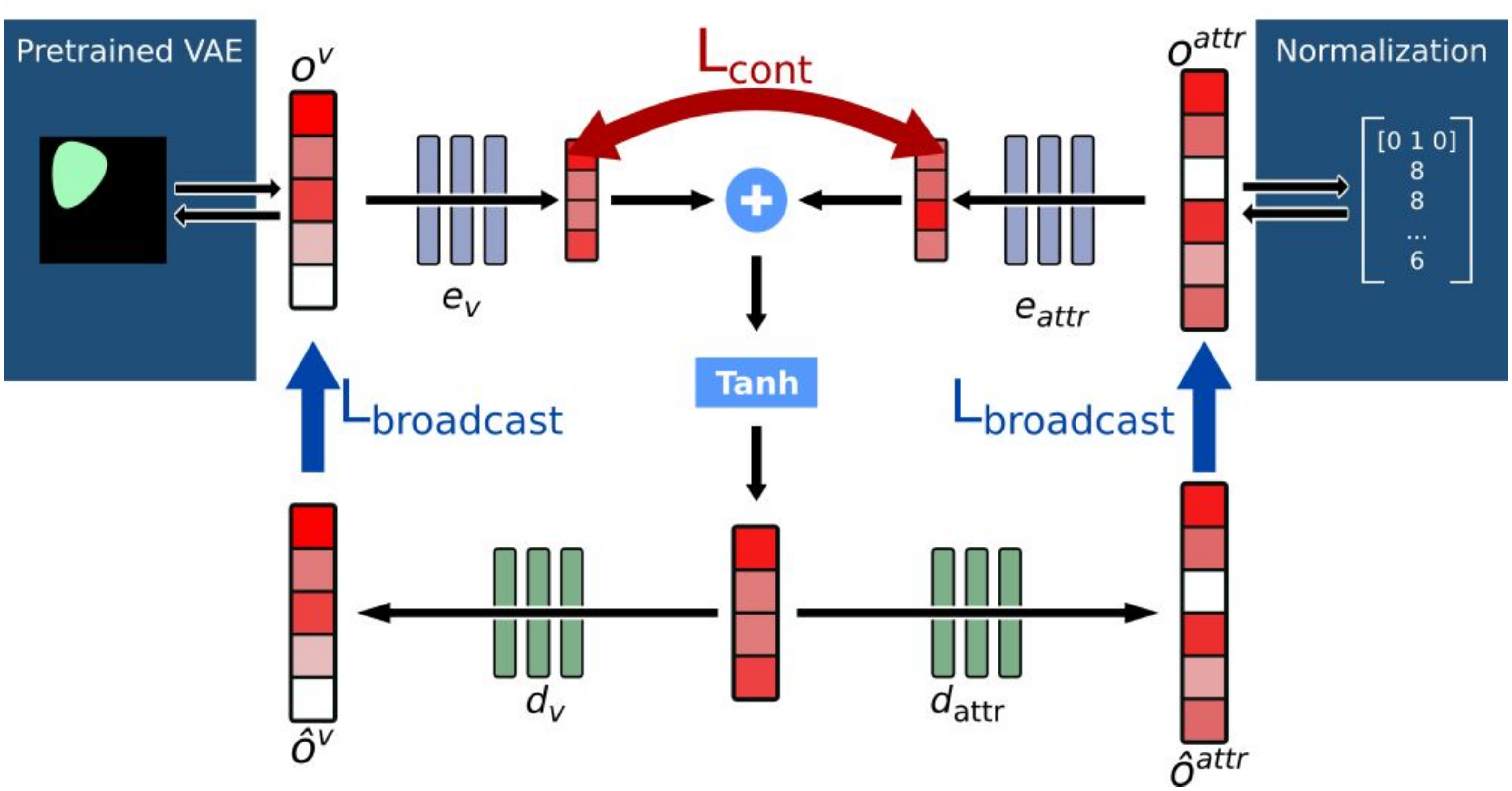
Global Workspace

A Global Workspace with fusion

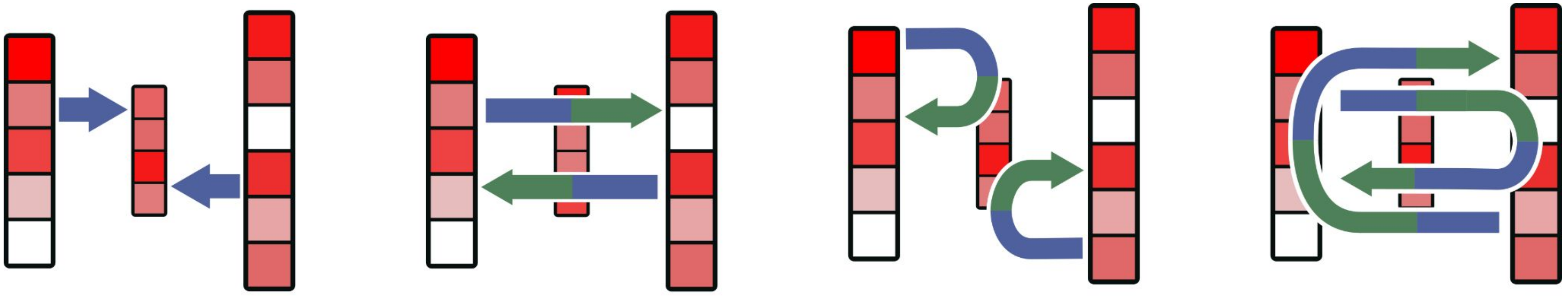


Global Workspace

A Global Workspace with fusion



A Global Workspace with fusion



L_{cont} L_{tr} L_{dcy} L_{cy}

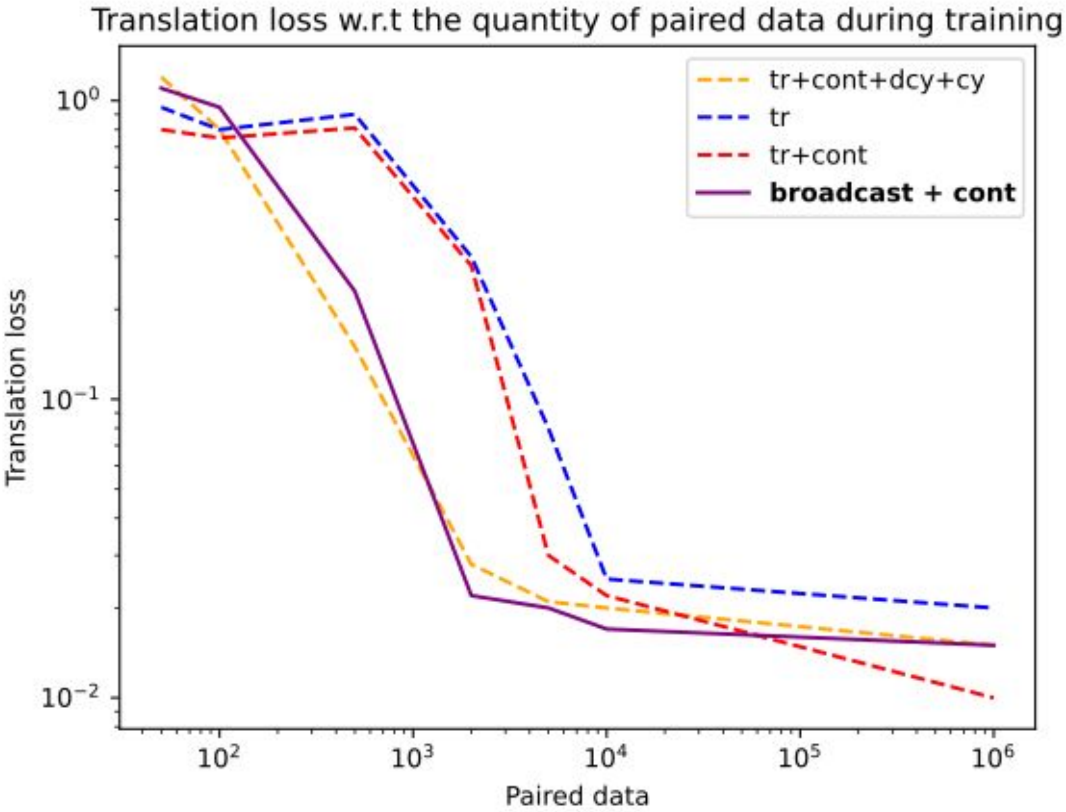
Contrastive loss

Broadcast loss

A Global Workspace with fusion

Simple Shapes Dataset

Influence of the number of paired data on translation performances



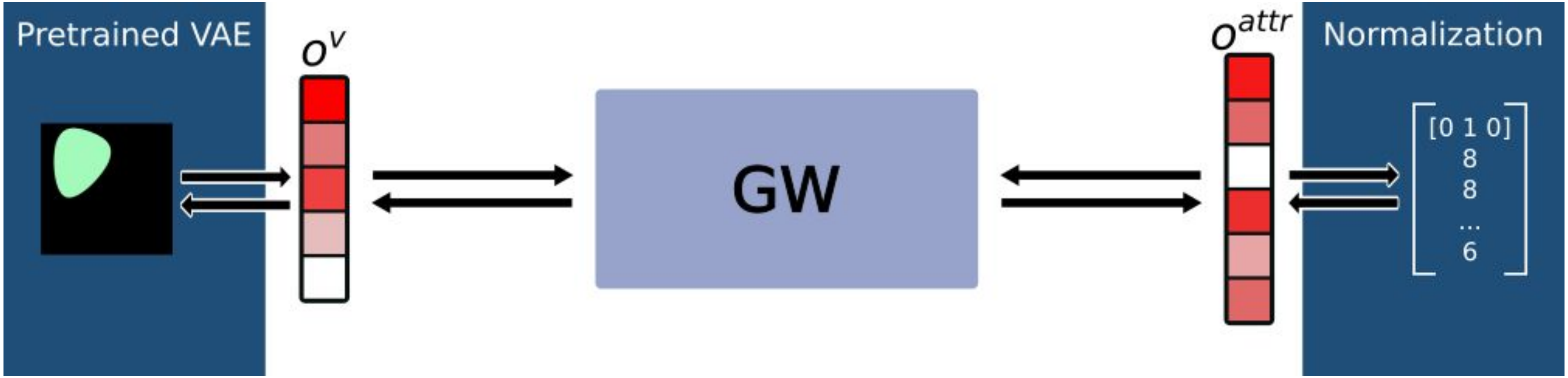
With the fusion, we find the same semi-supervision results than the model trained with all the losses

Adding attention to the Global Workspace

Adding attention to the Global Workspace

Adding an attention system on top to select which modality enters in the GW

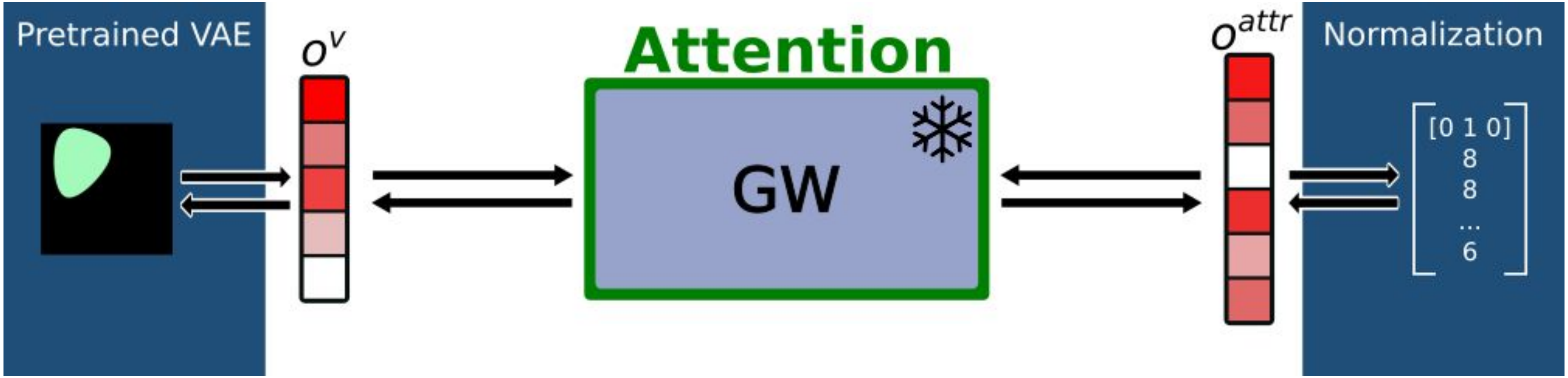
Two steps training:
1- Train the Global Workspace with the two losses



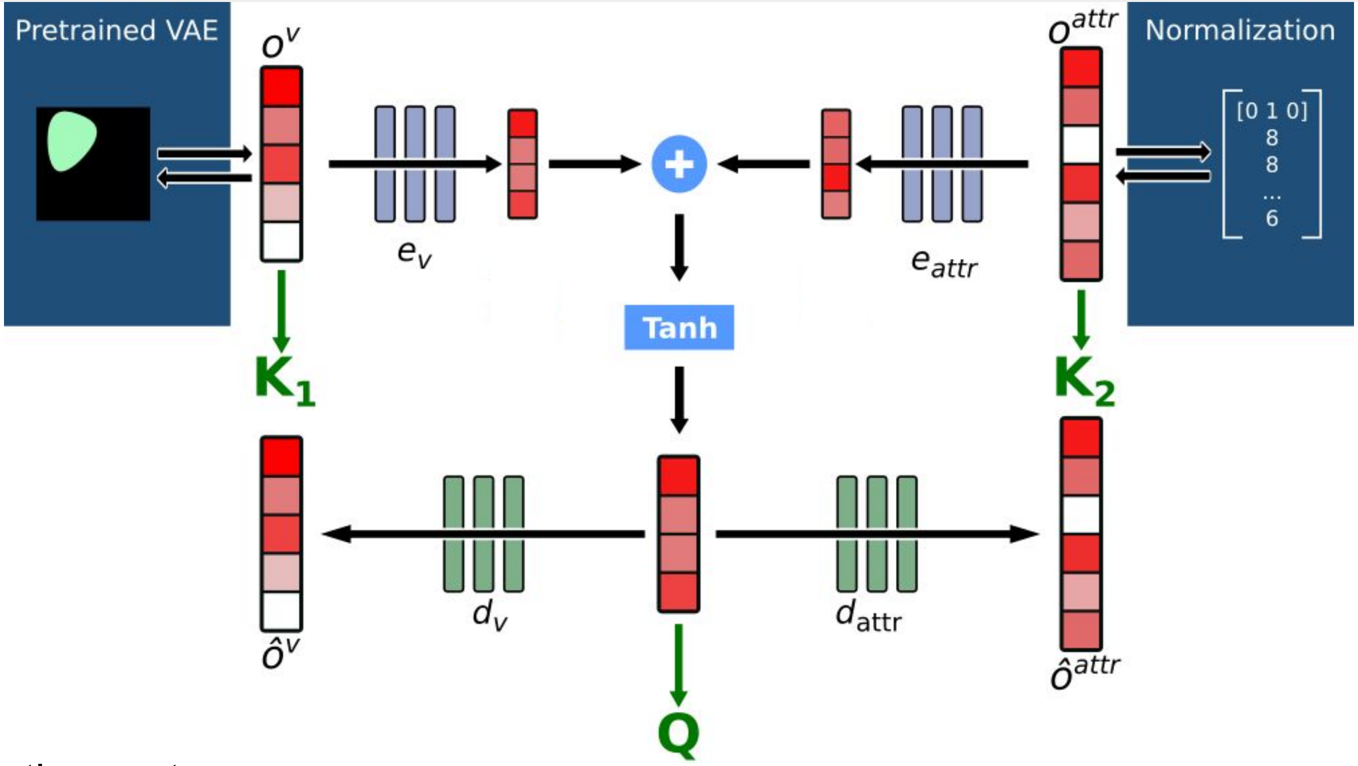
Adding attention to the Global Workspace

Adding an attention system on top to select which modality enters in the GW

Two steps training:
2- Train the attention system on top of the pretrained model



Adding attention to the Global Workspace



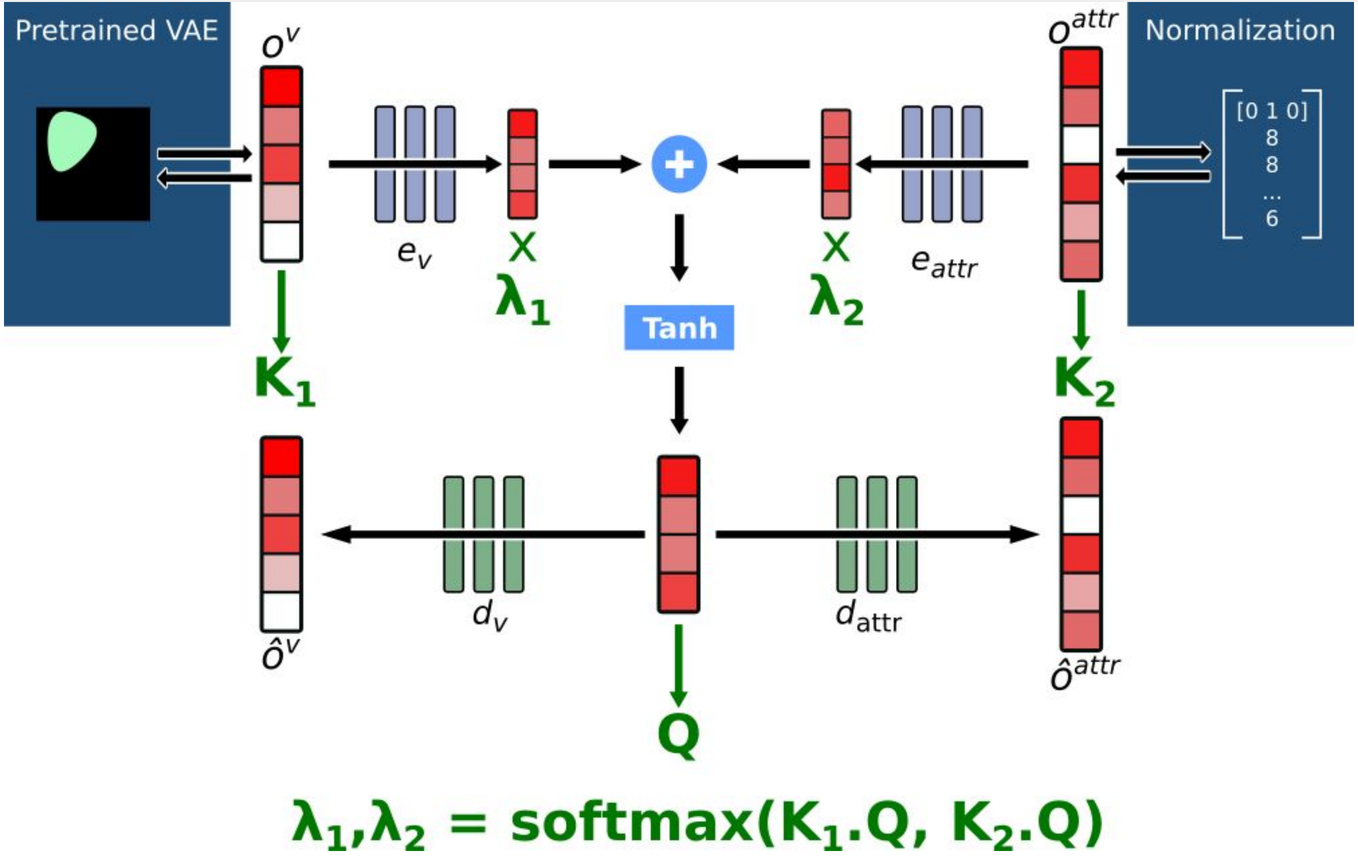
Key Query attention system:

- Keys: coming from unimodal latent
- Query: coming from the Global Workspace vector

$K_1 = W_1.o^v \quad K_2 = W_2.o^{attr}$

$Q = W_q.z$

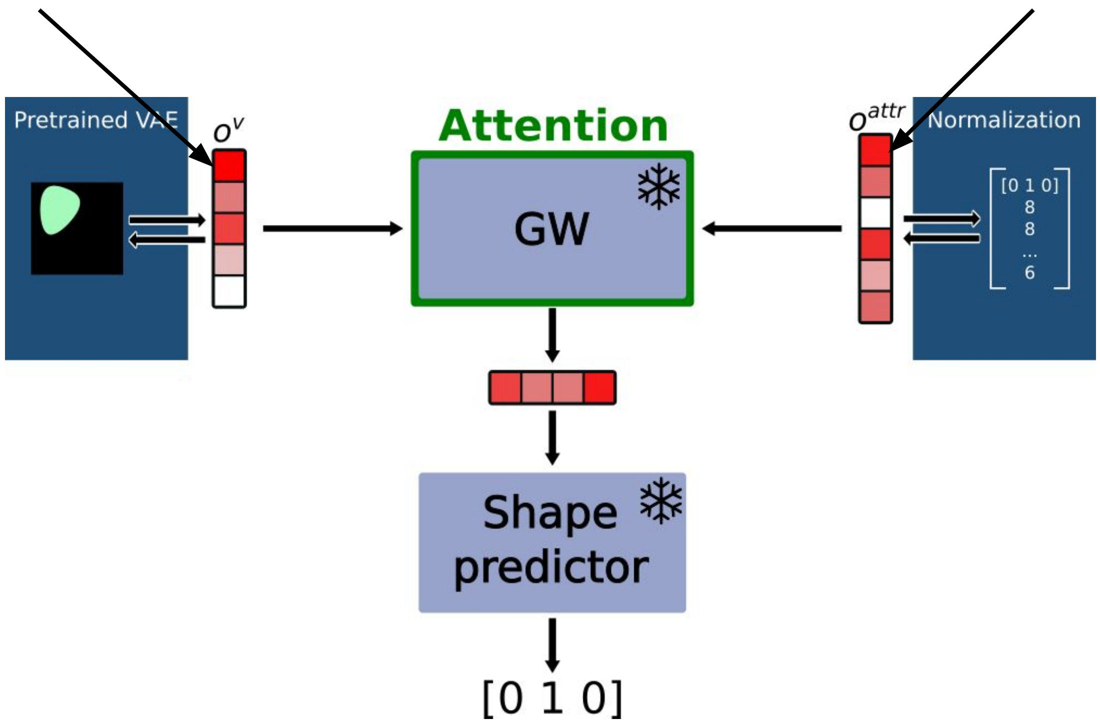
Adding attention to the Global Workspace



Adding attention to the Global Workspace

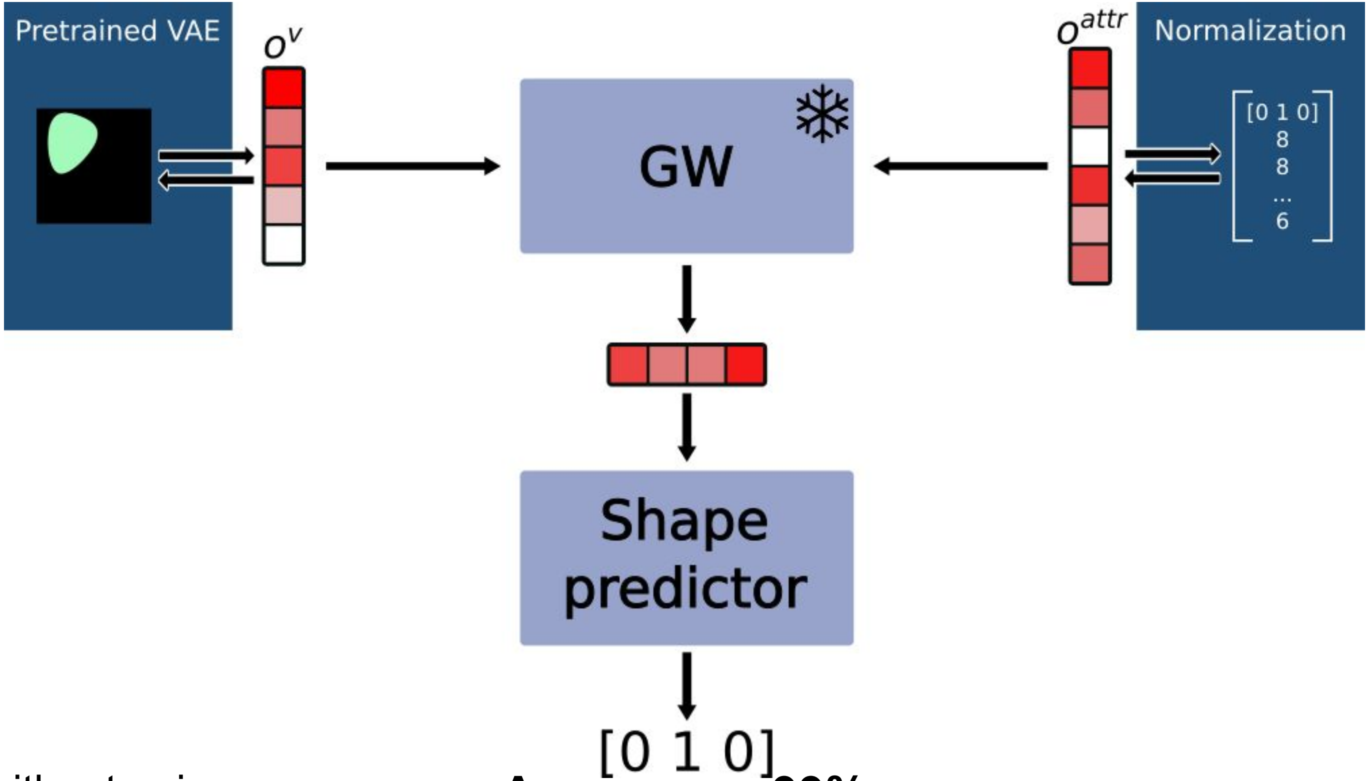
Train the attention on shape classification from the Global Workspace with corruption on one side

Add corruption here … … or here



Adding attention to the Global Workspace

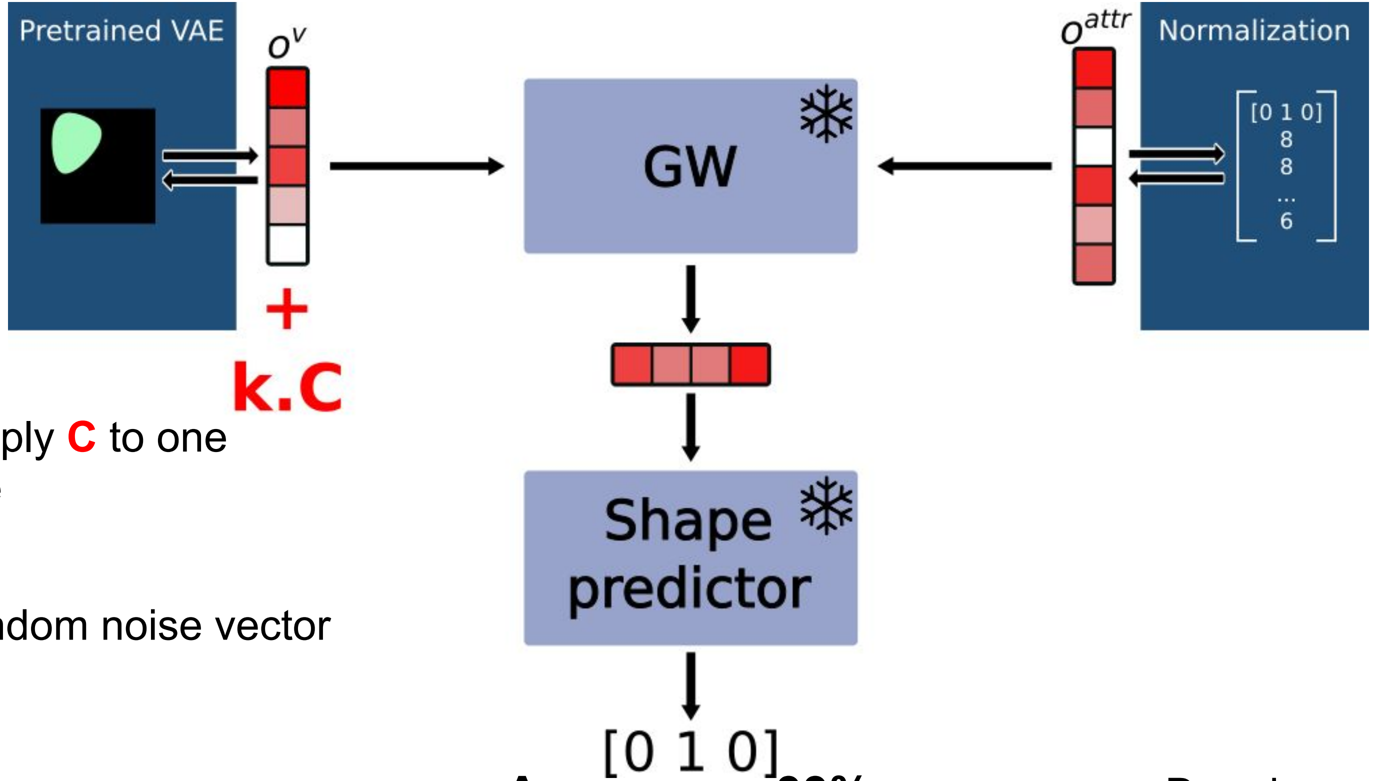
First: Train shape predictor to classify the shape from the Global Workspace (GW)



Works well without noise → **Accuracy = 99%**

Adding attention to the Global Workspace

What is the robustness of this model to a fixed noise



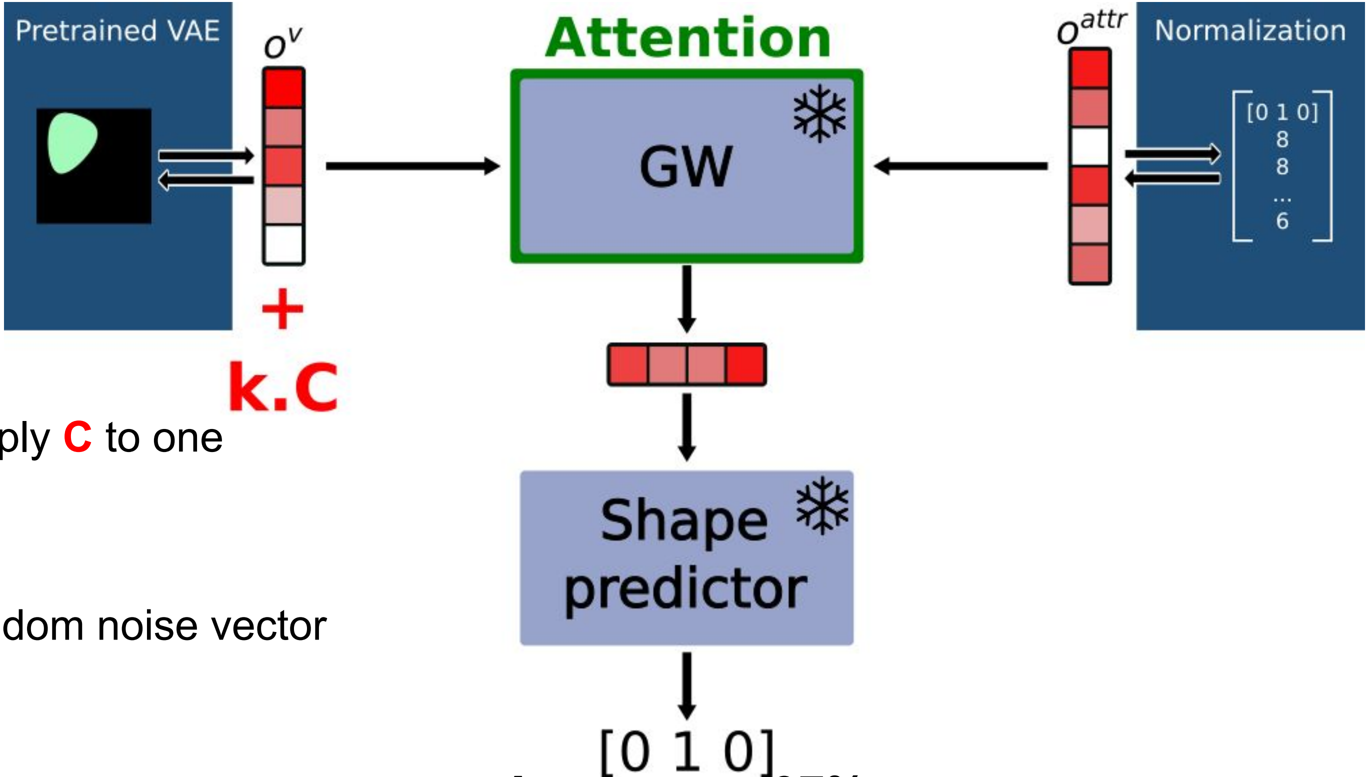
Randomly apply **C** to one side at a time

C = Fixed random noise vector
k ~ N(0,σ2)

Accuracy = 33% ← Drop in performances

Adding attention to the Global Workspace

Train attention system to select the modality entering in the GW by maximizing the accuracy



Randomly apply **C** to one side at a time

C = Fixed random noise vector
k ~ N(0,σ2)

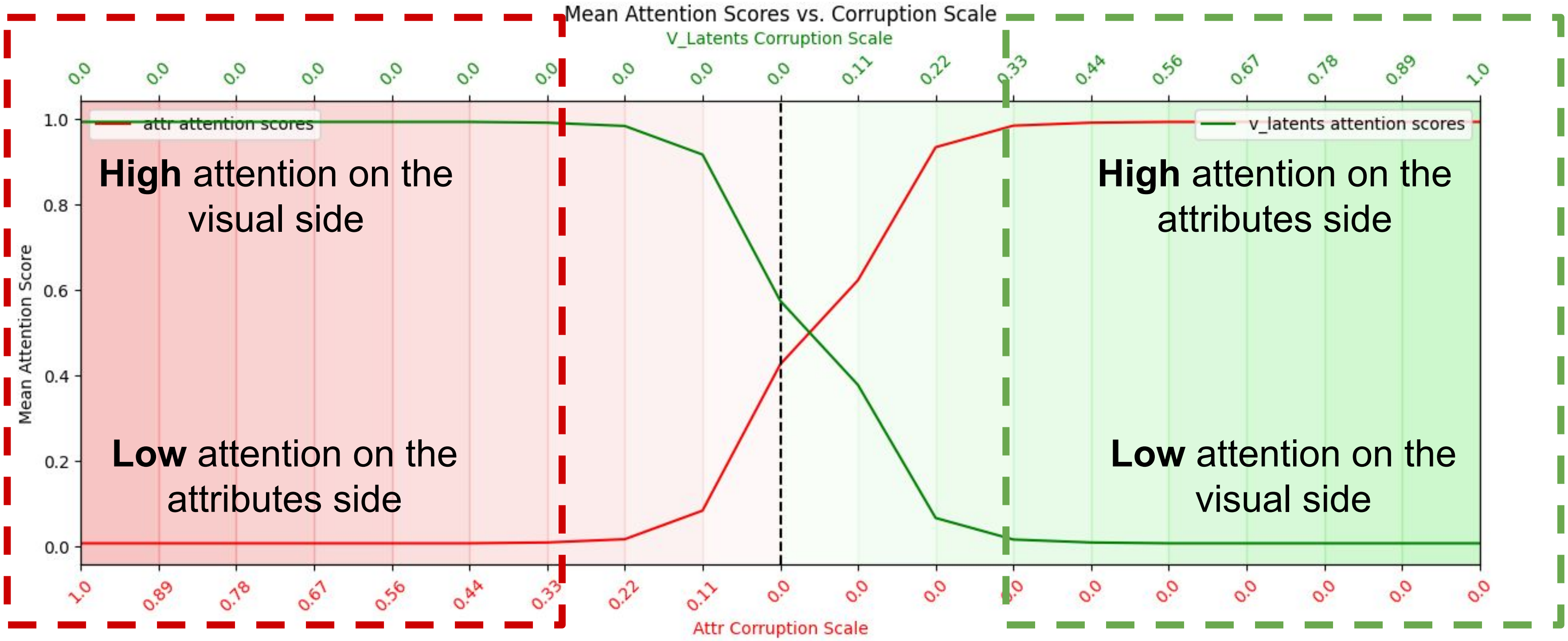
Accuracy = 97% ⟵ Good performances

Adding attention to the Global Workspace

How much the model is paying attention to each modality according to the noise level ?

Very noisy attributes

Very noisy images



Conclusion

Conclusion

Multiple specialized pre-trained models are able to **collaborate** through the Global Workspace by sharing information

Training relies on multiple losses leading to **semi-supervised** setting and decreasing need of multimodal data

Fusion allows to combine multiple modalities entering in the Global Workspace at the same time

Attention can select the right combination of input modalities to achieve a goal (e.g. noise robustness)

CerCo
UMR5549

ANITI
Université Fédérale
Toulouse Midi-Pyrénées

Funded by the European Union

erc
European Research Council
Established by the European Commission

Thank you for your attention


R. VanRullen


B. Devillers


H. Chateau-Laurent

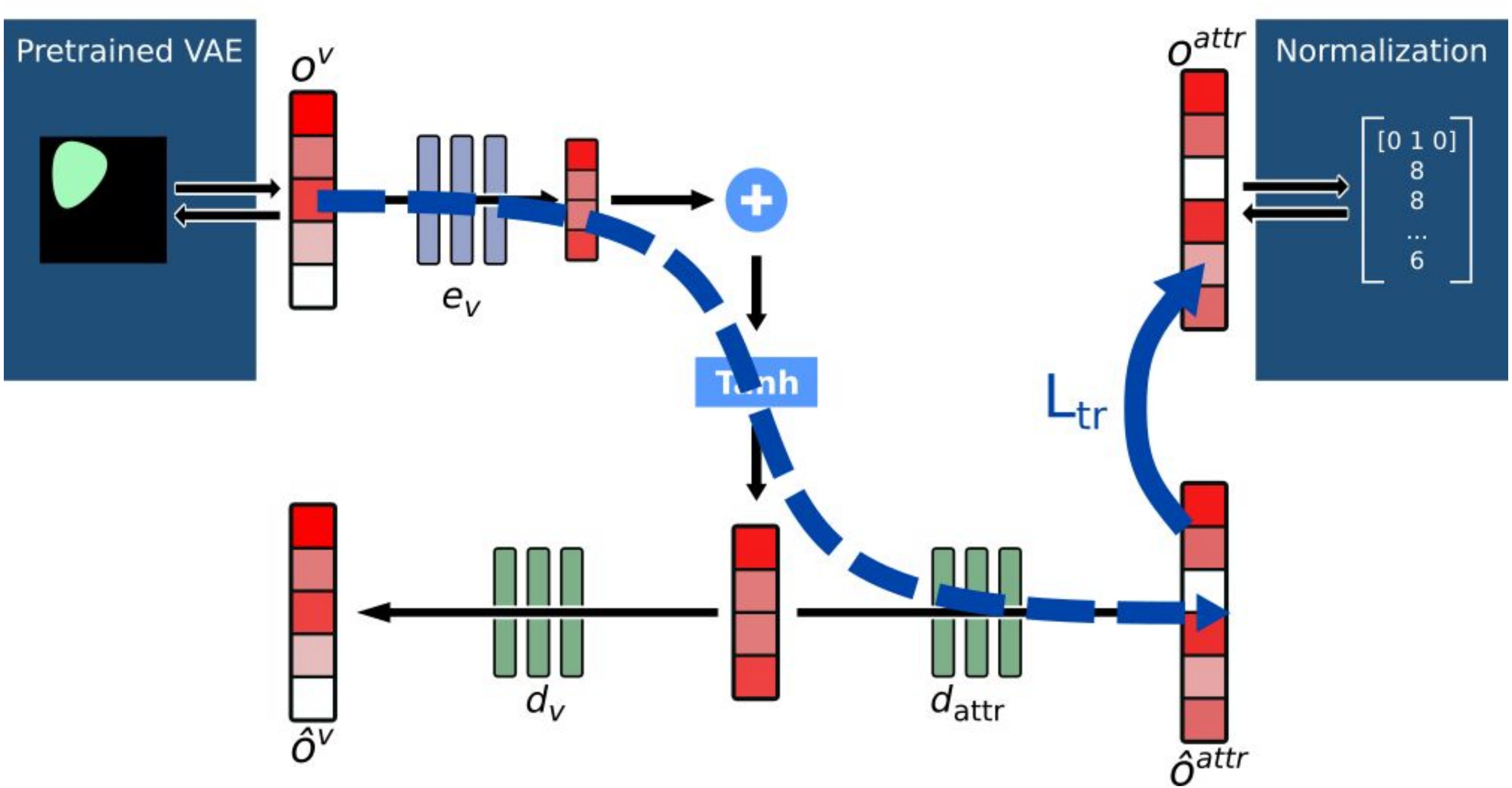

M. Nikolaus


R. Bertin-Johannet

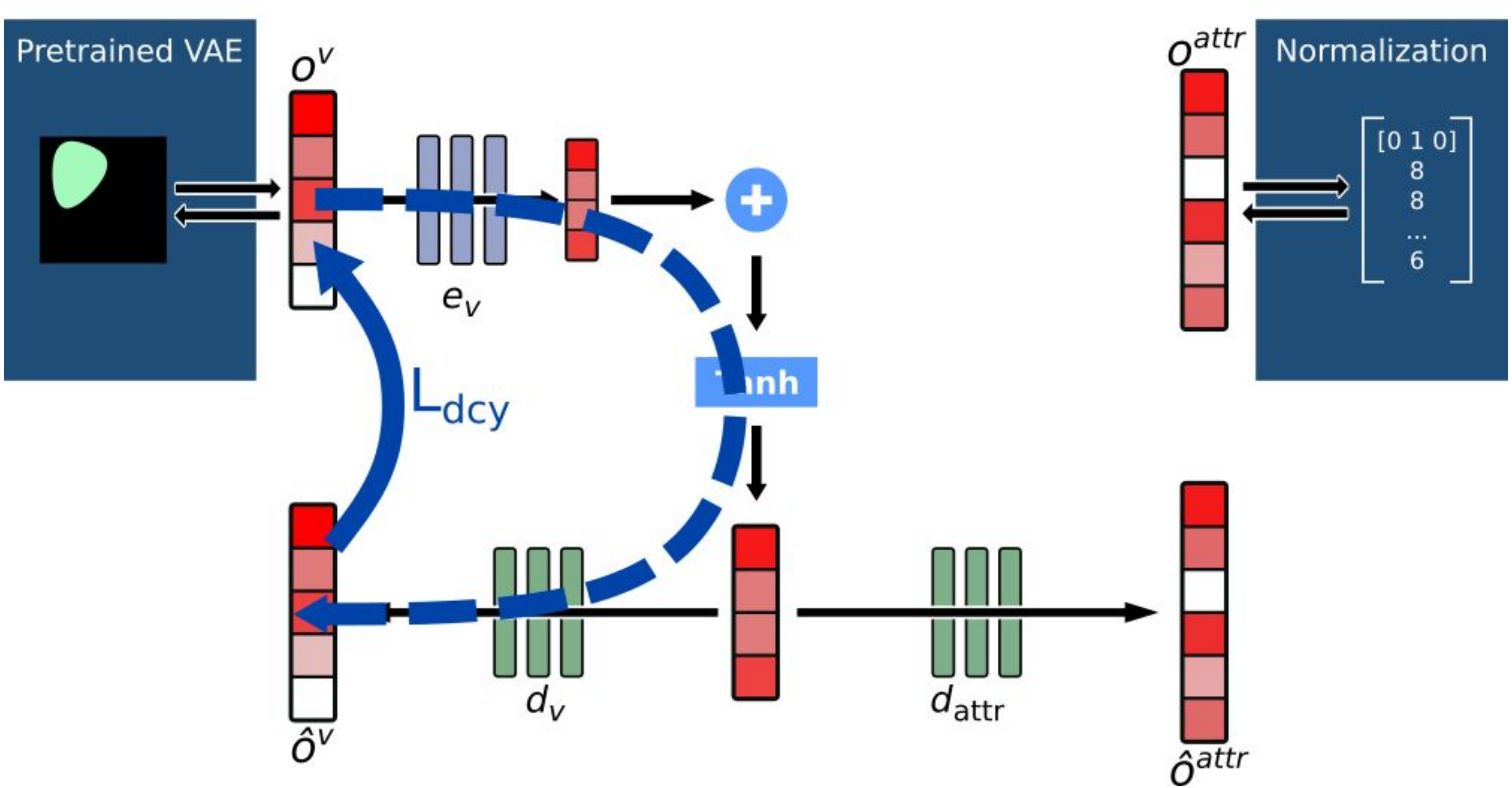

L. Maytié


L. Scipio

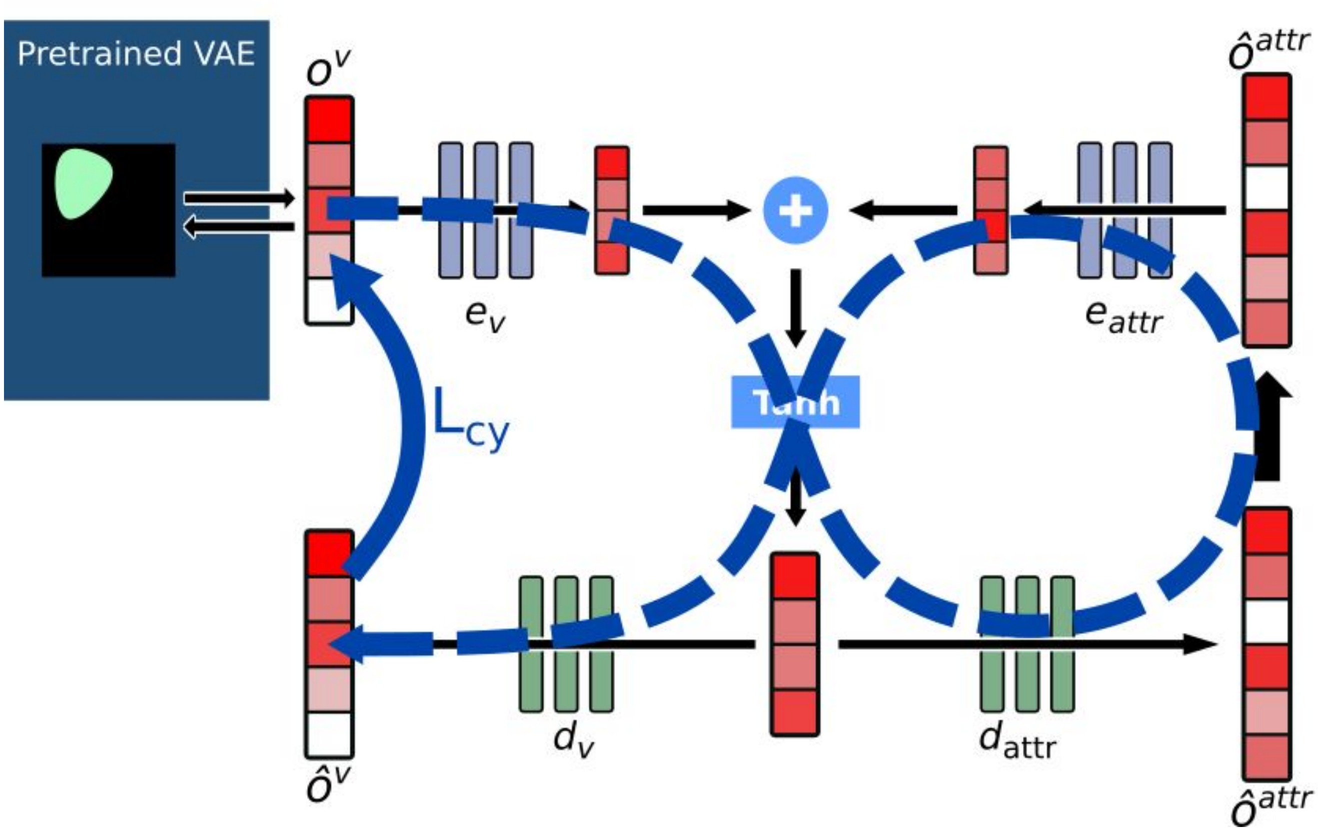
A Global Workspace with fusion



A Global Workspace with fusion



A Global Workspace with fusion



A Global Workspace with fusion

Ground Truth text

Associated images

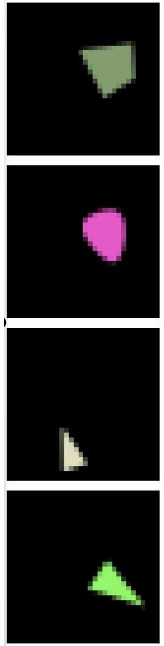
Translated images

The image is a large object pointing towards the top-left corner, and it is at the middle right, and it is olive green colored, and is looks like a kite

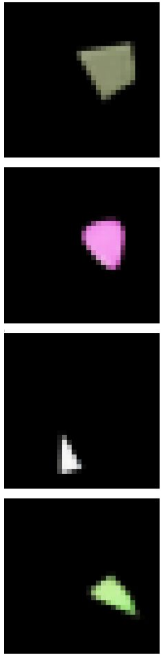
It is a guitar pick, medium size and in pink color, pointing to the bottom, at the center right

The image is a tiny triangle pointing towards the upper-left corner, it is at the bottom with a beige color

It is a medium size lime green triangle pointing to the bottom right at the lower right side of the image

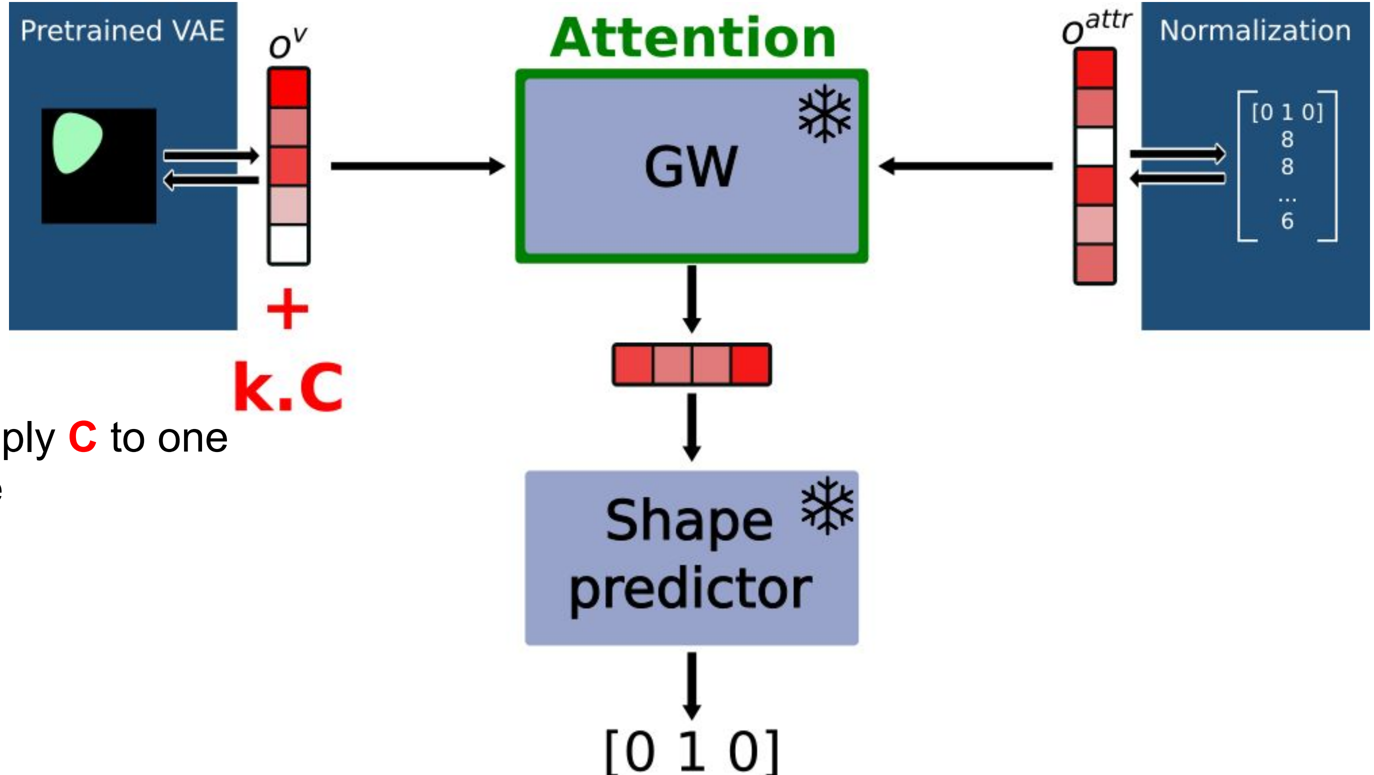


Translation through the Global Workspace



Adding attention to the Global Workspace

What is the robustness of this model with attention to a random noise



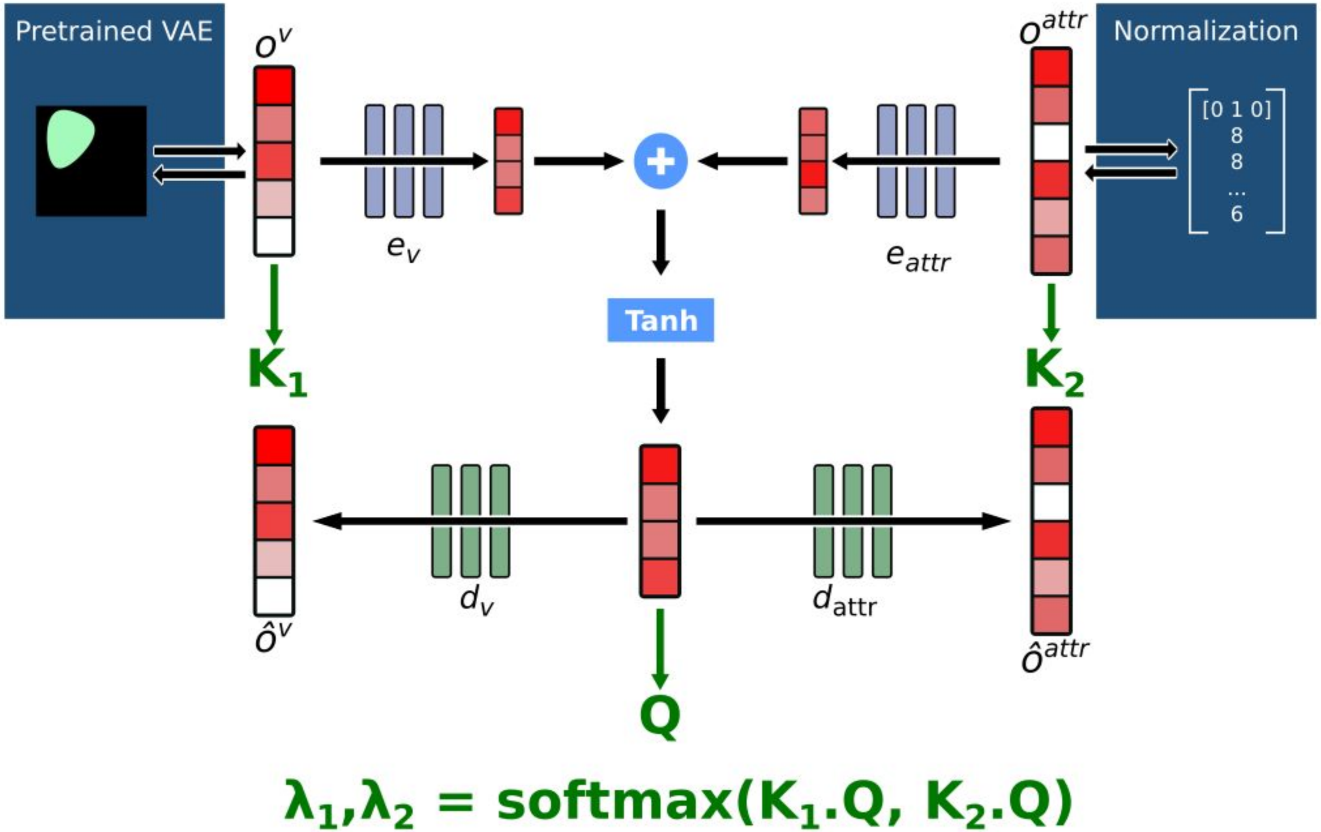
Randomly apply **C** to one side at a time

C ~ N(0,σ1)
k ~ N(0,σ2)

Accuracy = 37% ← Very bad

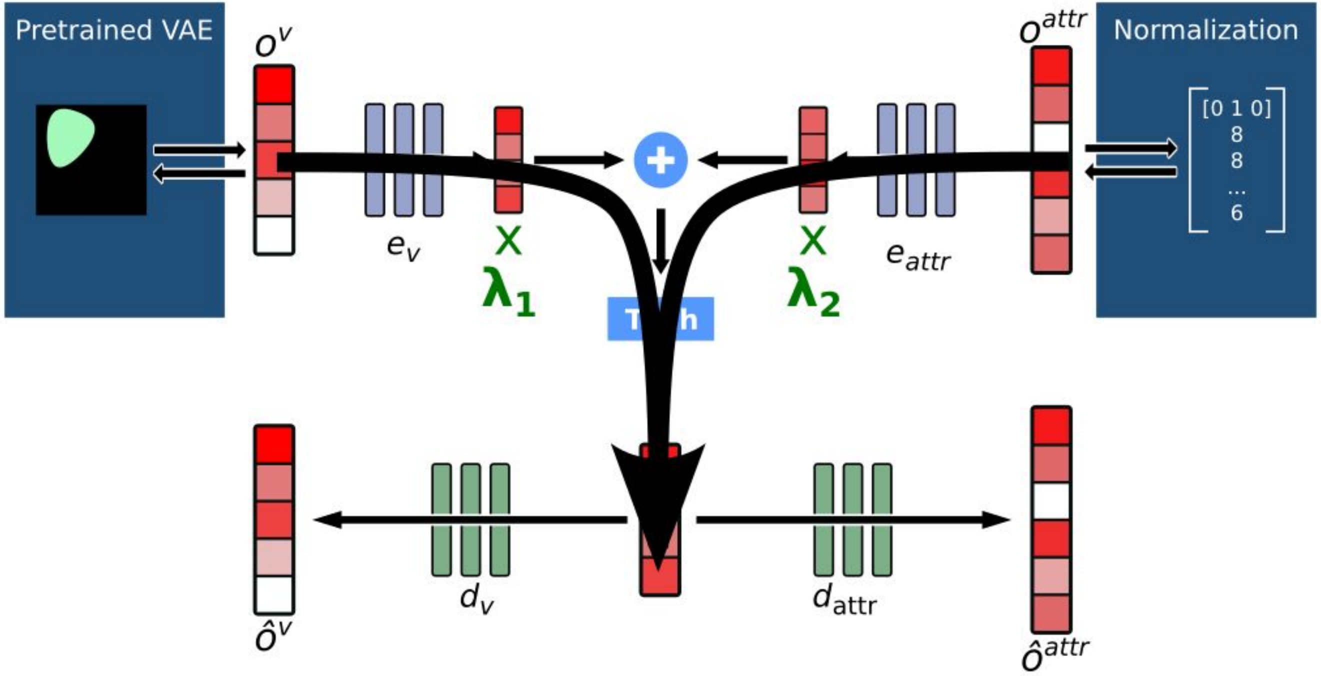
Adding attention to the Global Workspace

Make a 2 step attention system to adapt the Query to a non random GW



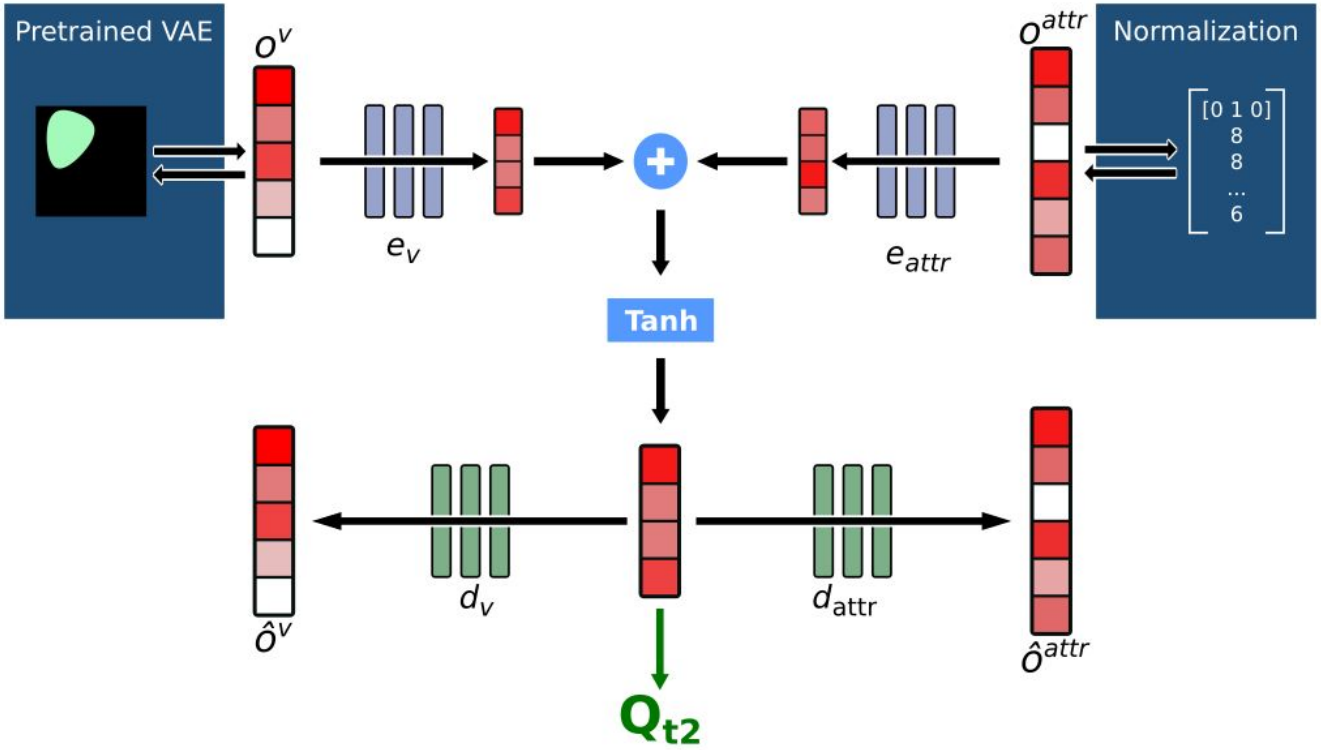
Adding attention to the Global Workspace

For this, encode inputs through GW using 1st step attention (random Query)



Adding attention to the Global Workspace

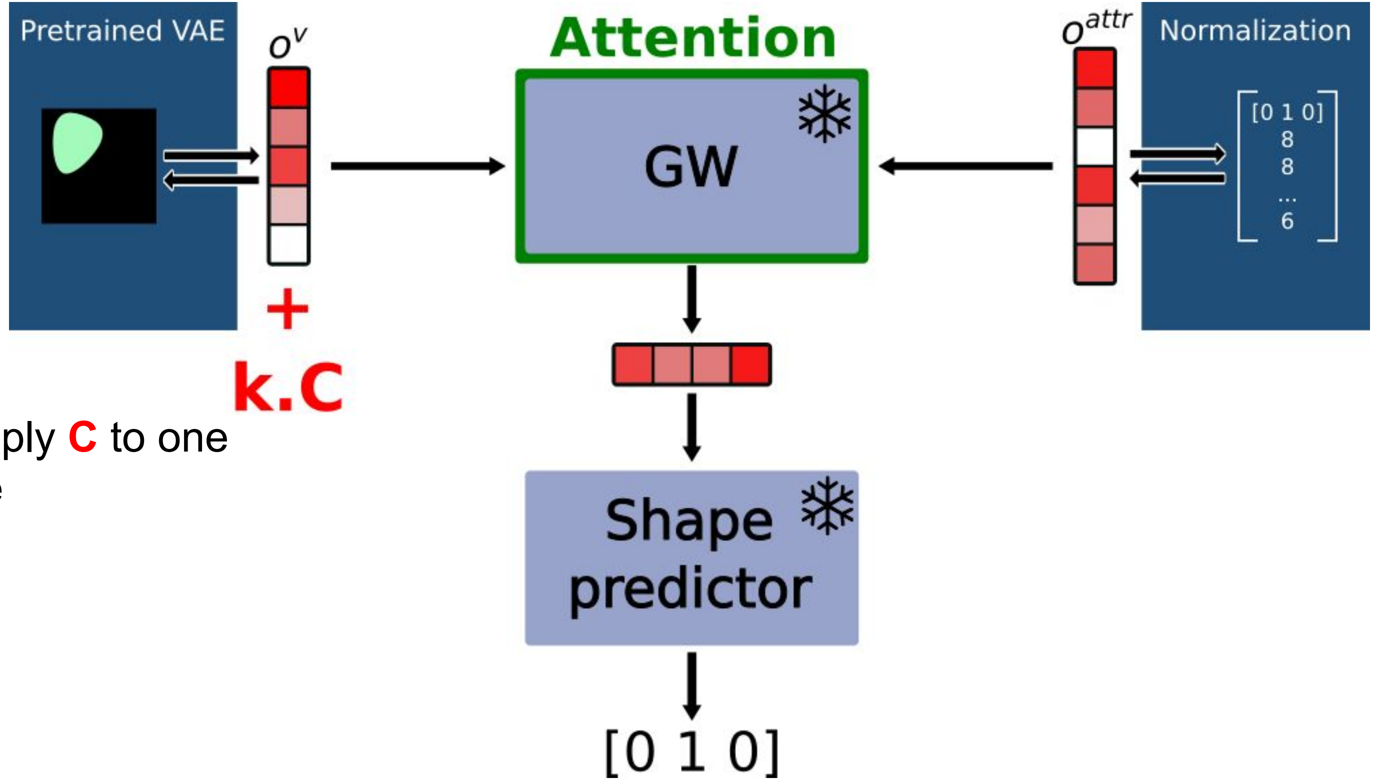
Generate a new Query from the obtain GW



$$\lambda_1,\lambda_2 = \mathrm{softmax}(K_1.Q_{t2},\ K_2.Q_{t2})$$

Adding attention to the Global Workspace

To use it in the second pass of the input through the GW



Randomly apply **C** to one side at a time

C ~ N(0,σ1)
k ~ N(0,σ2)

Accuracy = 89% ← Way better

Adding attention to the Global Workspace

How much the model is paying attention to each modality according to the noise level ?

Very noisy attributes

Very noisy images