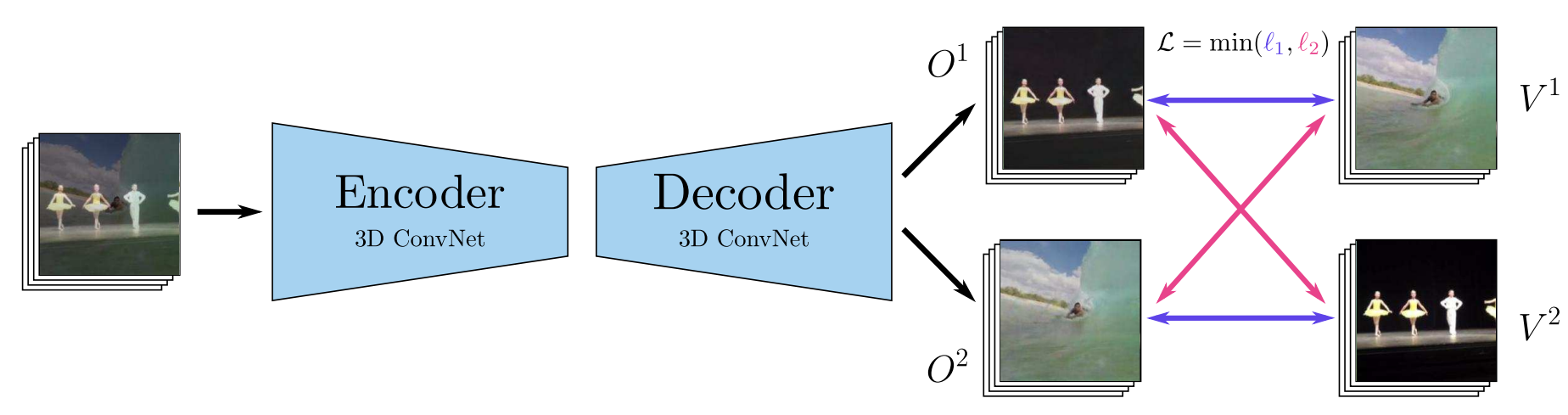
Controllable Attention for Structured Layered Video Decomposition

Jean-Baptiste Alayrac*,[1], Joao Carreira*,[1], Relja Arandjelović[1] and Andrew Zisserman[1,2]

*equal contribution, [1]Deepmind, [2]University of Oxford

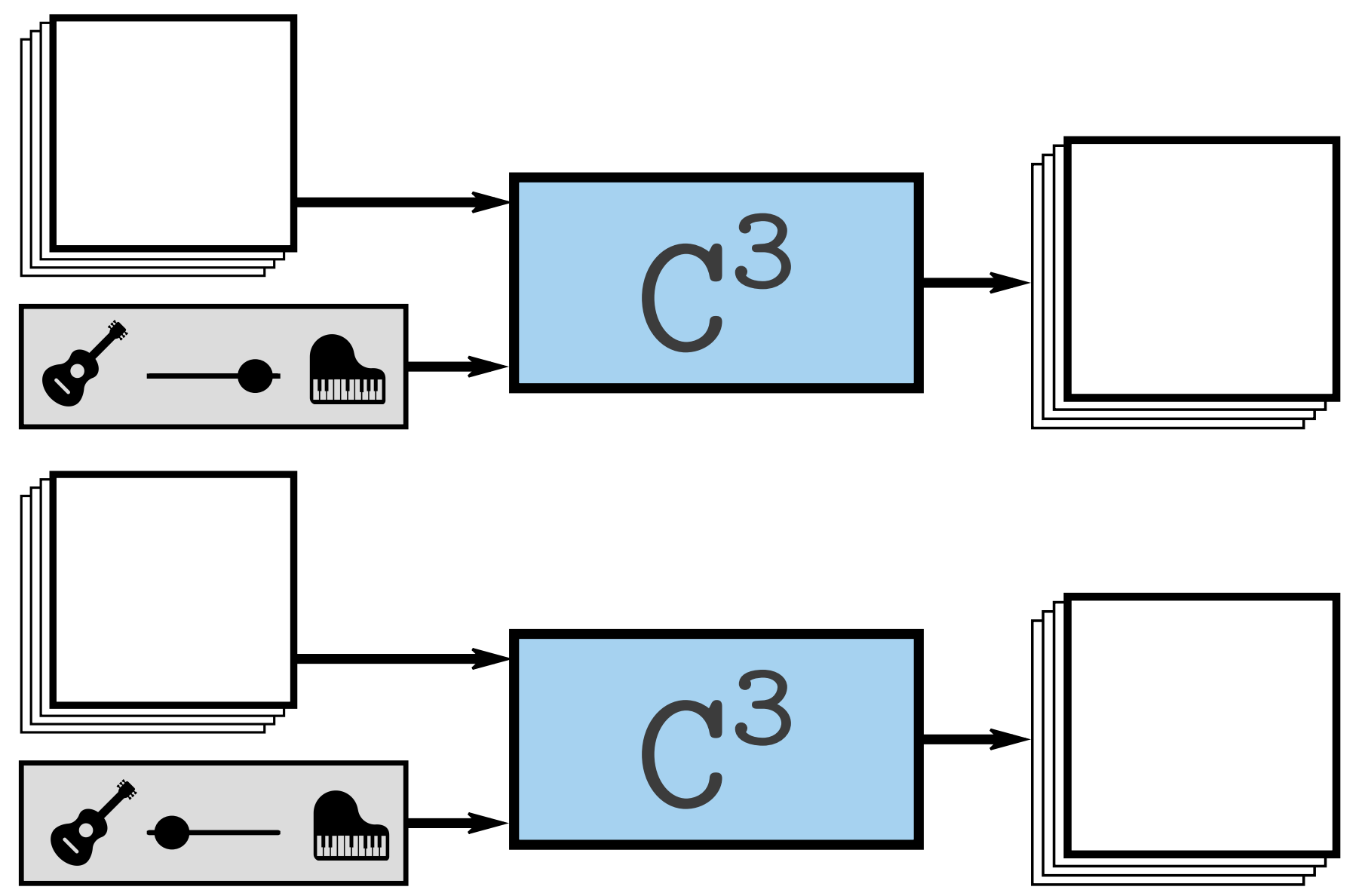
Overview

Previous work: Visual Centrifuge



The visual centrifuge: Model-free layered video representations, Alayrac, Carreira and Zisserman, CVPR, 2019.

Controllable Compositional Centrifuge



Goal: be able to separate a video into its natural layers, and to control which of the separated layers to attend to.

Contributions:

- *Compositional architecture* (C²) for layer decomposition.
- Augment the architecture to leverage external cues such as audio for *controllability* (C³).

Conclusion

- New proposed compositional architecture can better handle automatically generated transparency and especially occlusions.
- Layers are correctly selected based on sound cues with accuracy close to 80%.

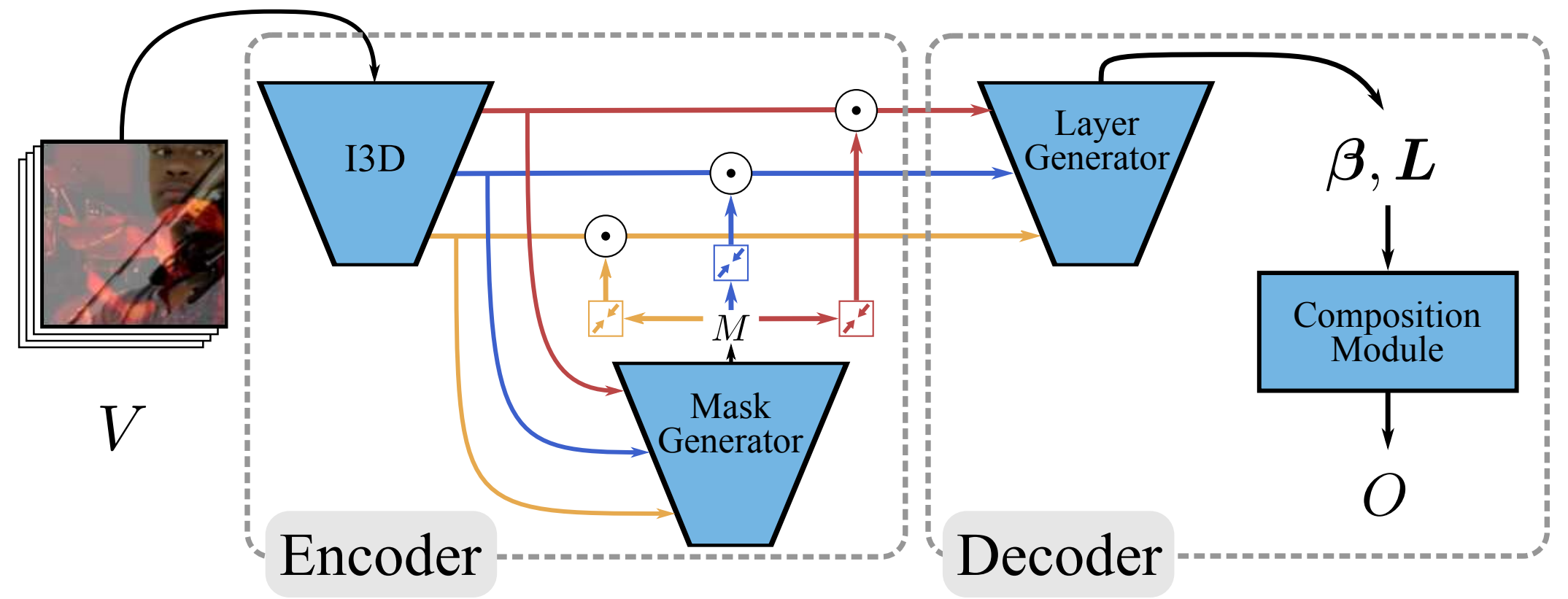
The Approach

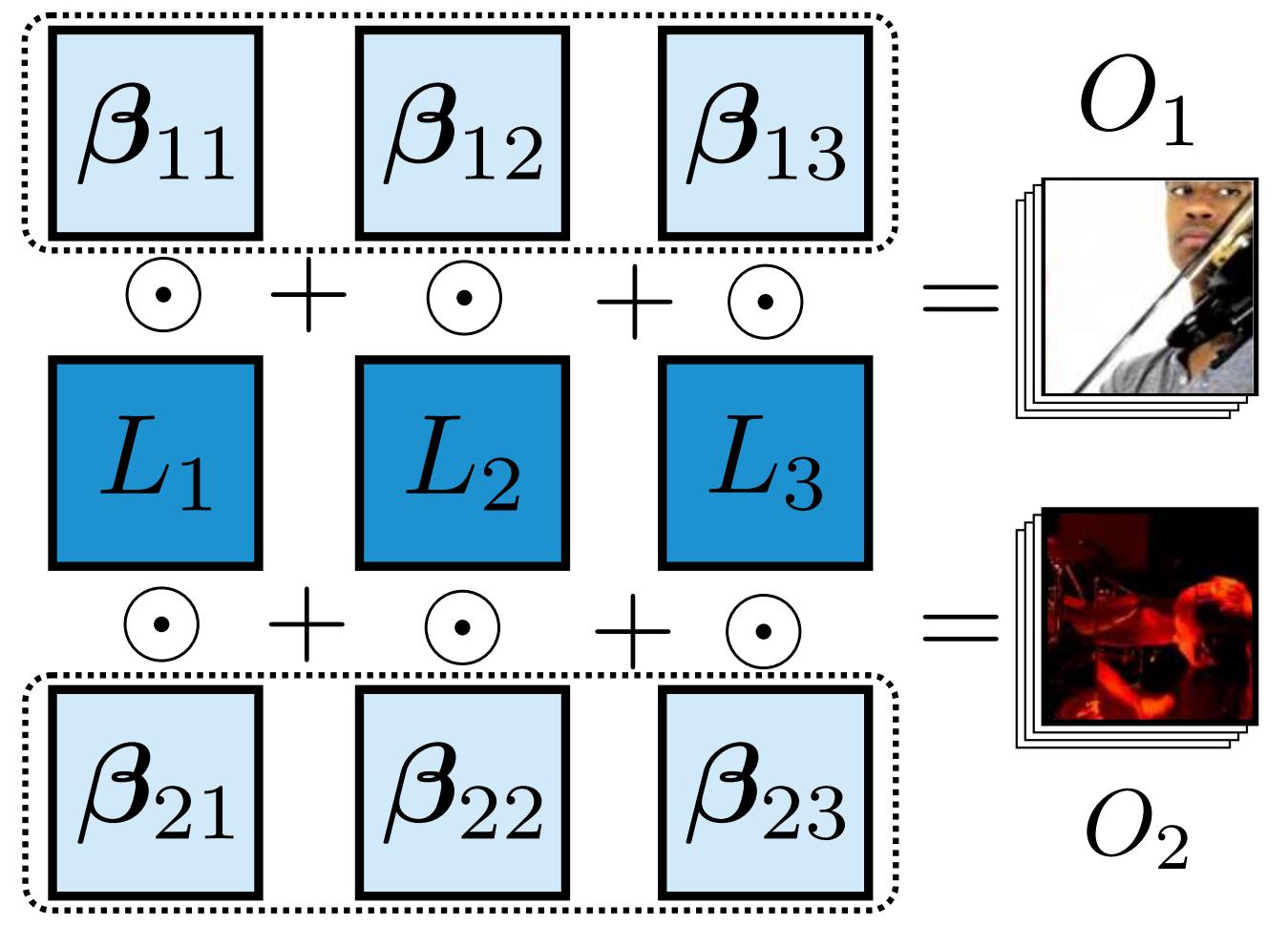
1 Architecture for layer decomposition: C²

High level goal: incorporate priors tailored to layer decomposition.

Modified Encoder architecture with grouped channel masking for handling occlusions and transparencies:

$$\tilde{F}_l^c = M_l^c \odot F_l^c$$




$\Sigma = 1$
$\boldsymbol{\beta}_{11}$
$\boldsymbol{\beta}_{12}$
$\boldsymbol{\beta}_{13}$
L_1
L_2
L_3
$\Sigma = 1$
$\boldsymbol{\beta}_{21}$
$\boldsymbol{\beta}_{22}$
$\boldsymbol{\beta}_{23}$
O_1
O_2

Composition module

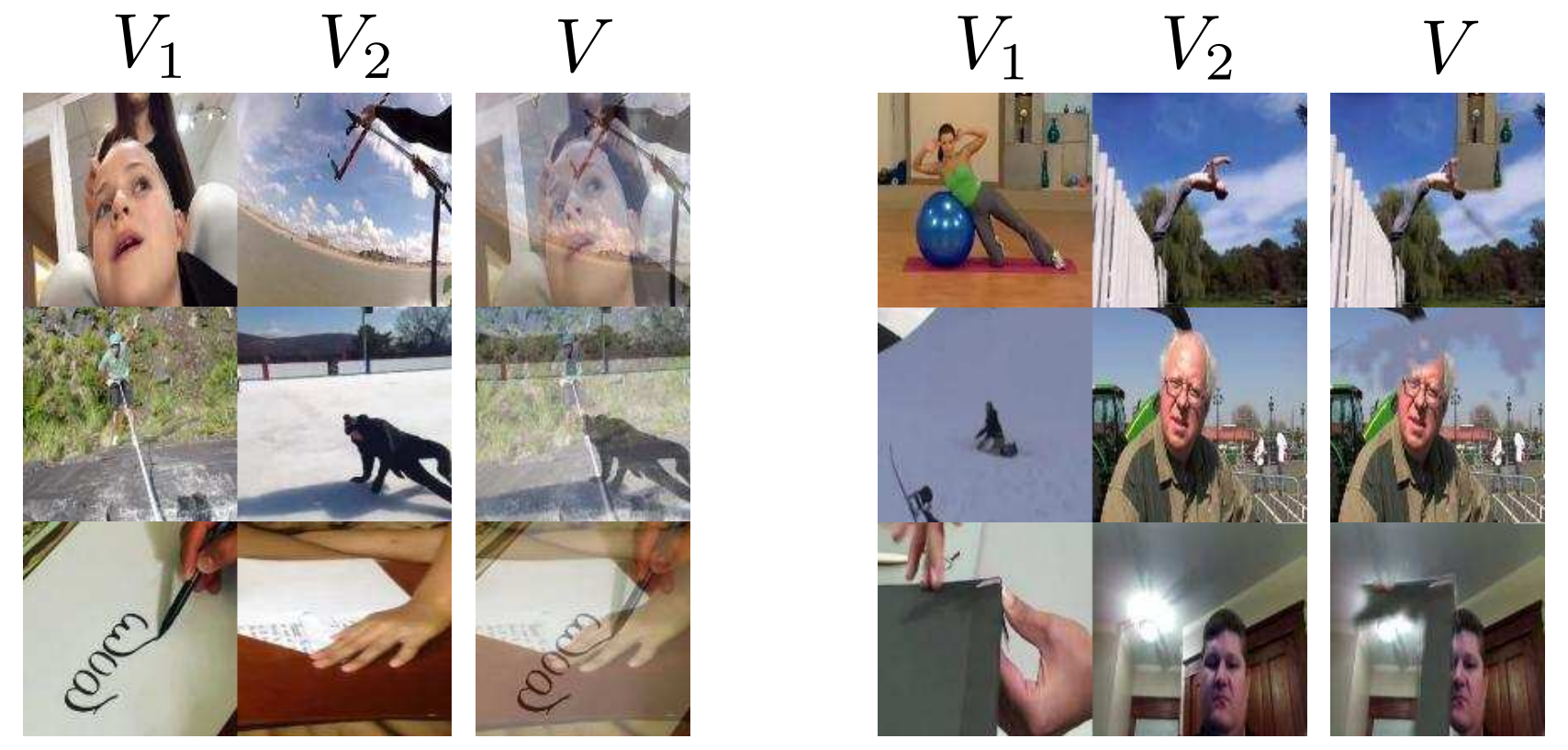
Imposing compositionality:

The decoder produces m layers L and composing coefficients $\boldsymbol{\beta}$ that are then composed as follows:

$$O_i = \sum_j \boldsymbol{\beta}_{ij} \odot L_j$$

3 Training procedure

Generating **training data** from Kinetics600:



Transparencies

Occlusion

Training losses:

- Without control:

$$\mathcal{L}_{\text{pil}}\left(\{V_1, V_2\}, \boldsymbol{O}\right) = \min_{(i,j)|i \neq j} \ell(V_1, O_i) + \ell(V_2, O_j)$$

- With control:

Deterministic Control loss:

$$\mathcal{L}_{\text{det}}\left(\{V_1, V_2\}, \boldsymbol{O}\right) = \ell(V_1, O_1) + \ell(V_2, O_2)$$

Internal Prediction loss:

$$\mathcal{L}_{\text{reg}}\left(V_1, \boldsymbol{s}\right) = \sum_{i=1}^{n} |s_i - \ell(V_1, \mathbf{sg}(O_i))|$$

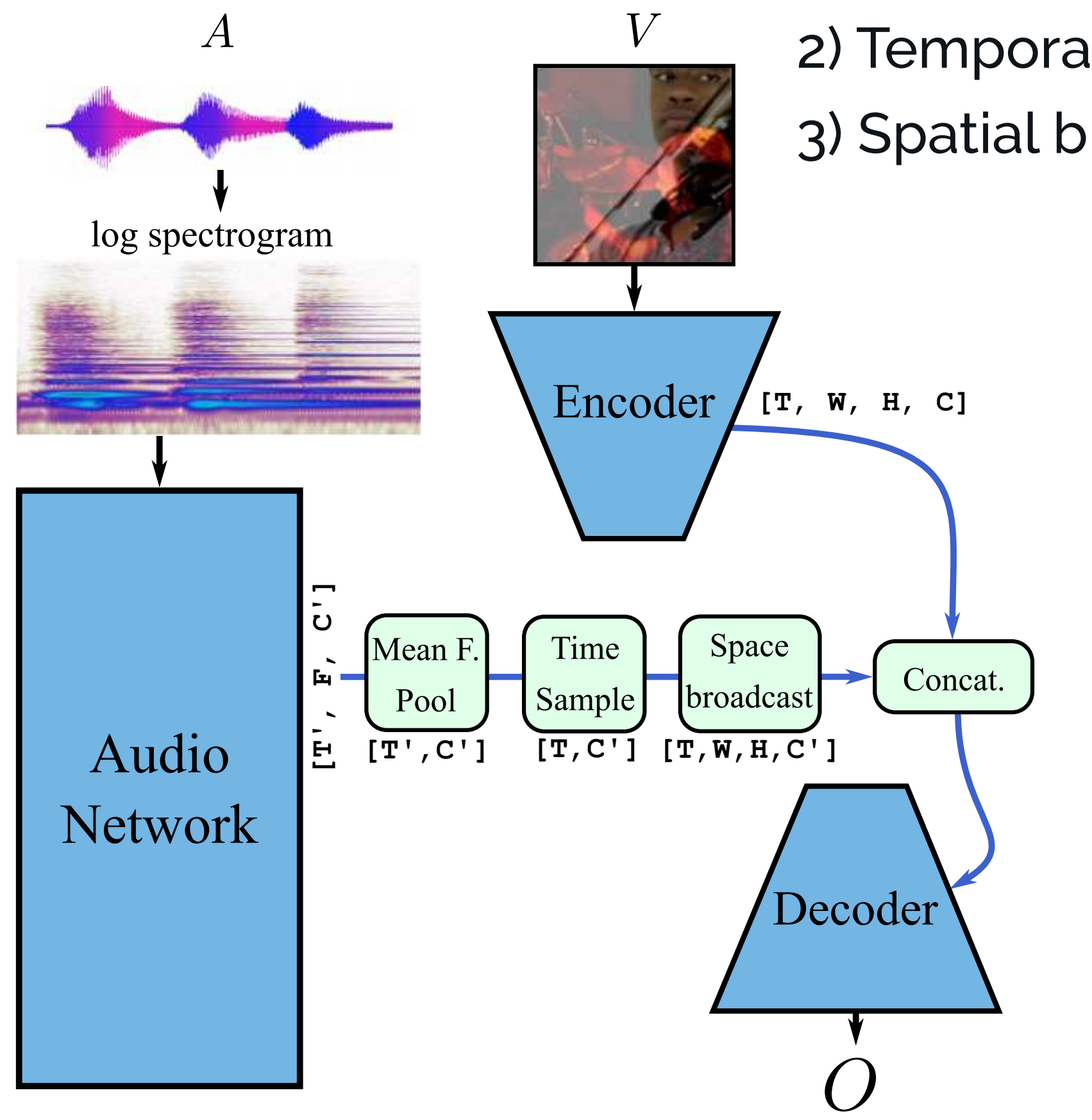
2 Controllable Compositional Centrifuge: C³

High level goal: have control over the output of the decoder by attending to an external cue, here an audio signal.

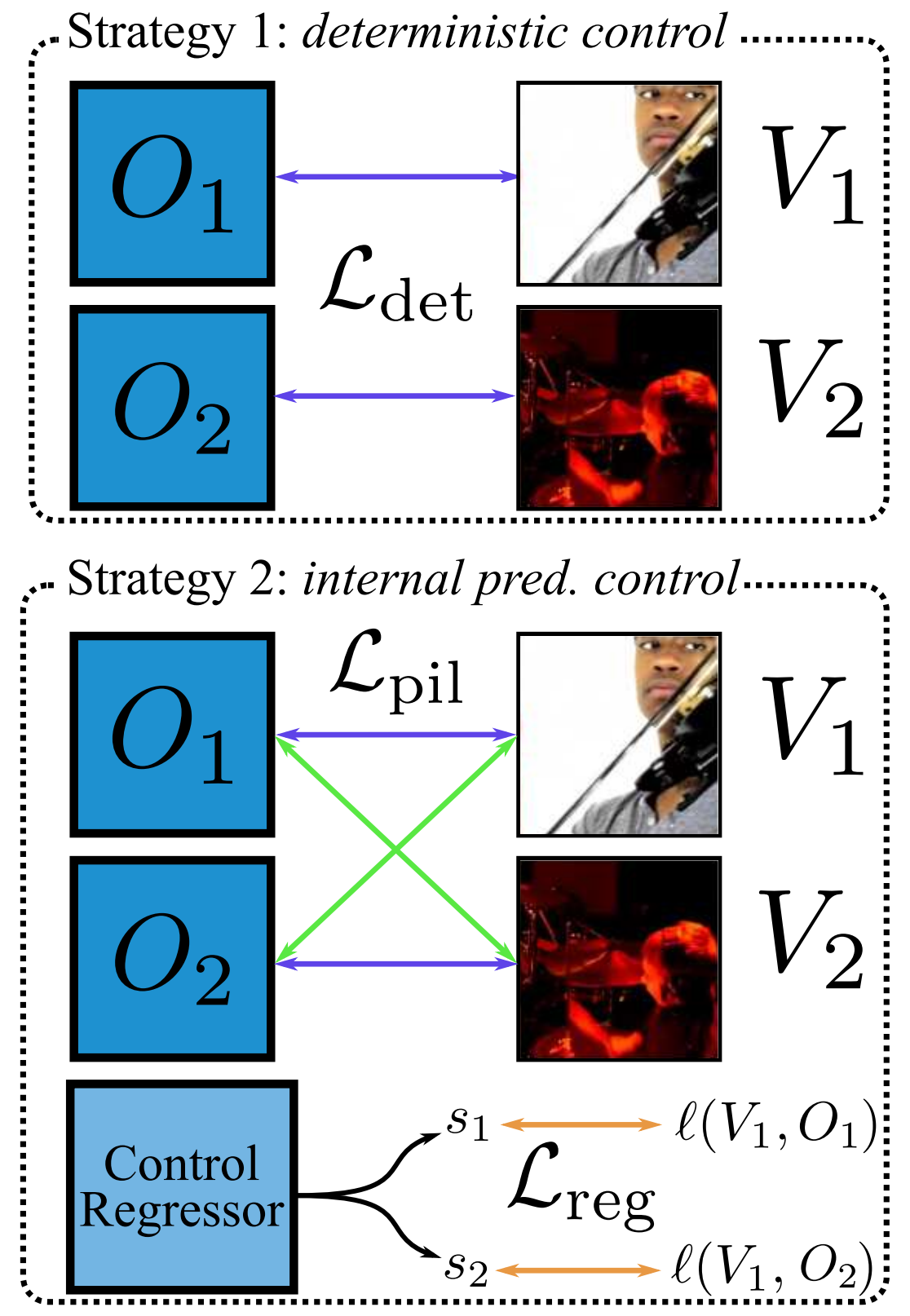
Audio network: VGG-like net on log spectrogram.

Audio visual fusion:

1) Average pooling over frequency
2) Temporal sampling
3) Spatial broadcast



Attention control. Two strategies are proposed:

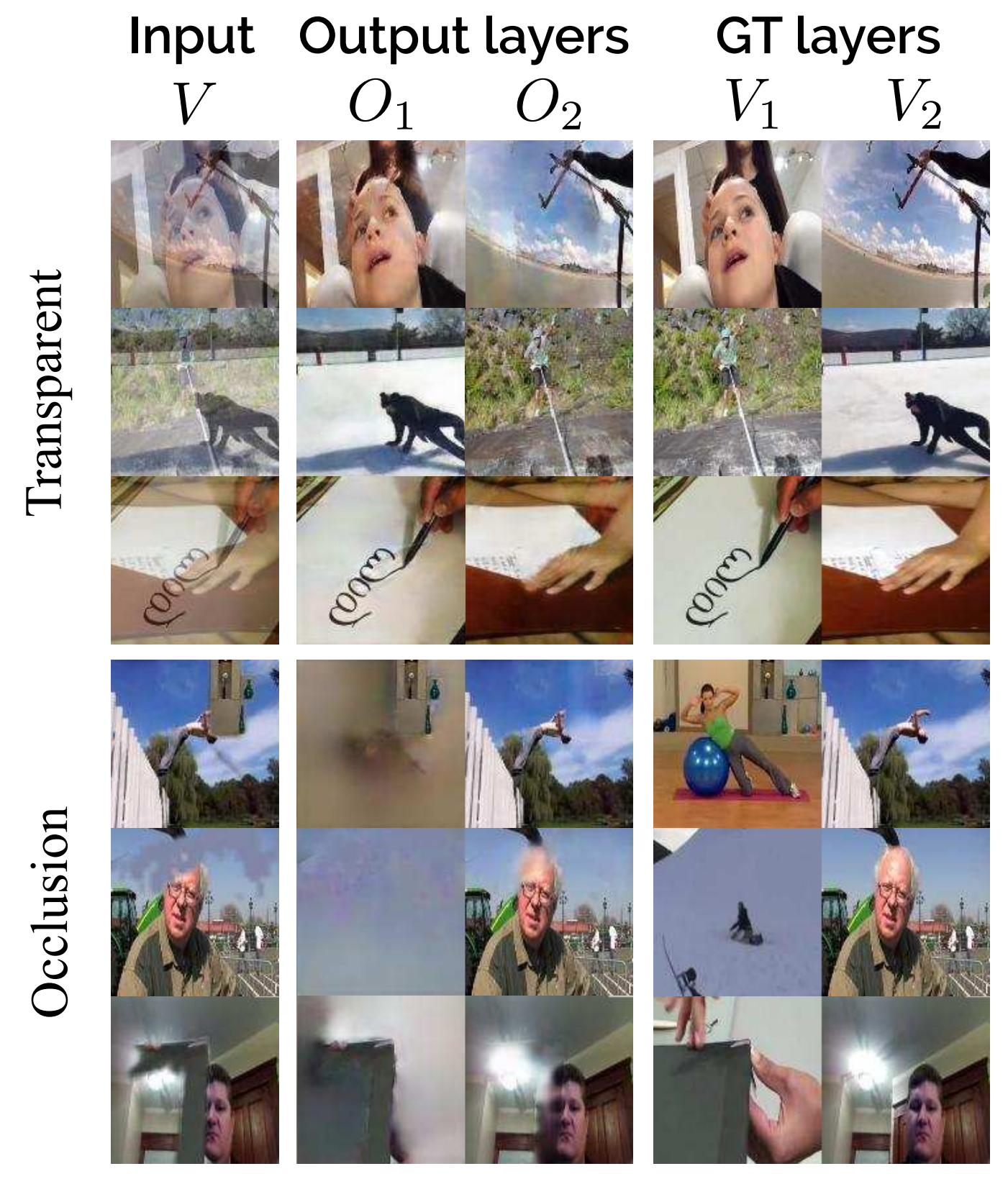

Strategy 1: deterministic control
O_1
O_2
$\mathcal{L}_{\text{det}}$
V_1
V_2
Strategy 2: internal pred. control
O_1
O_2
$\mathcal{L}_{\text{pil}}$
V_1
V_2
Control Regressor
s_1
s_2
$\mathcal{L}_{\text{reg}}$
$\ell(V_1, O_1)$
$\ell(V_1, O_2)$

***Deterministic:** desired video is forced to be output in a specific slot.*

***Internal prediction:** the network regresses where the desired output is going to be.*

Experiments

Compositional Centrifuge: C².



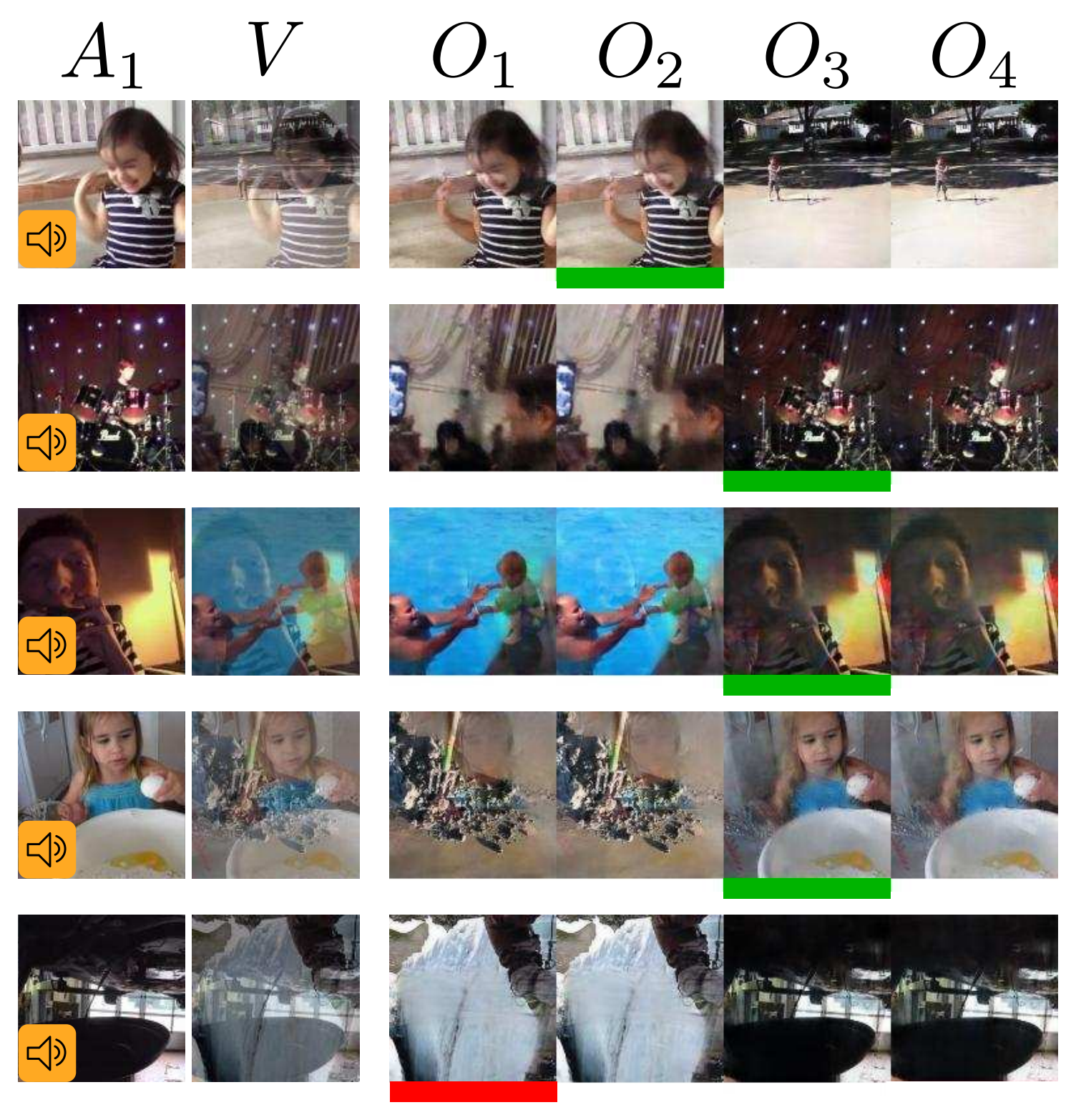
Qualitative results

Visualisation of the compositional outputs

Model	Loss (Transp.)	Loss (Occl.)	Size
Identity	0.364	0.362	–
Centrifuge	0.149	0.253	22.6M
CentrifugePC	0.135	0.264	45.4M
C² w/o masking	0.131	0.200	23.4M
C²	**0.120**	**0.190**	27.1M

Ablation study of the proposed improvements

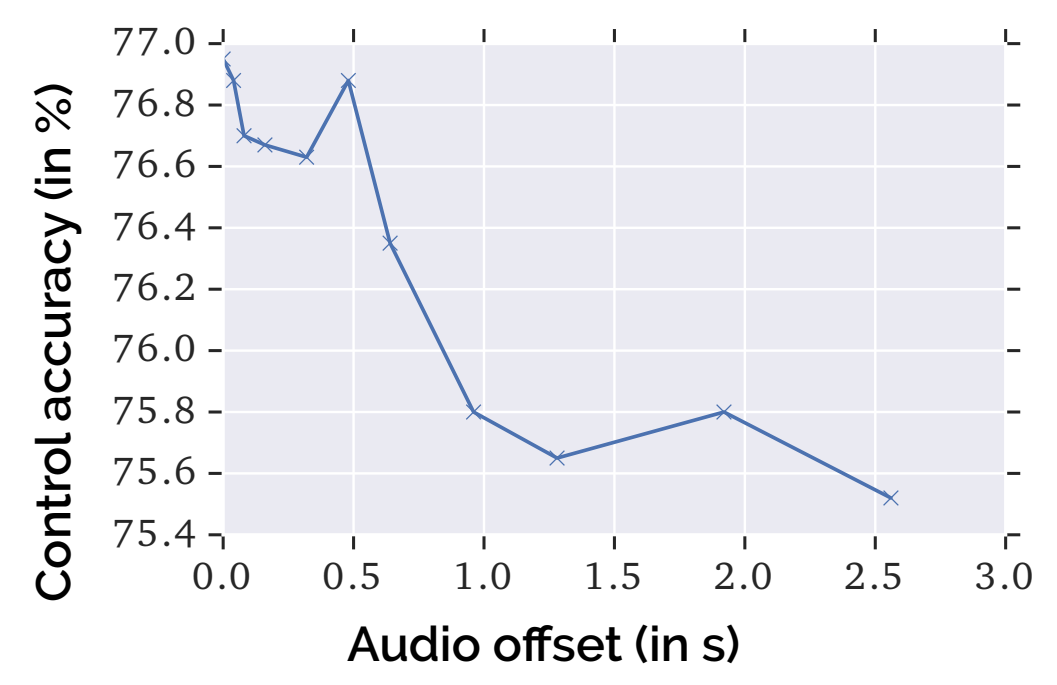
Controllable Compositional Centrifuge: C³.



Qualitative results

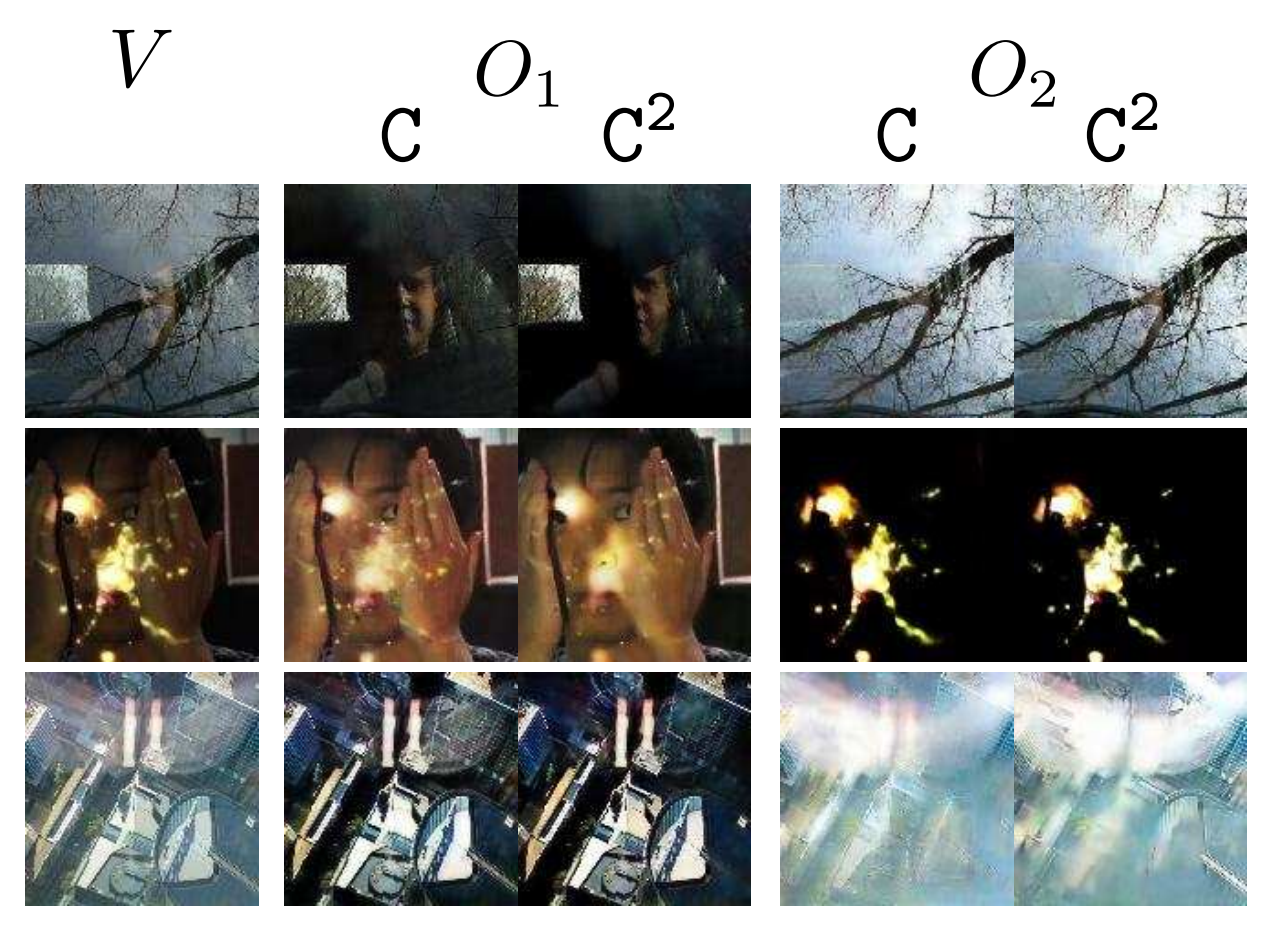
Model	Loss (Transp.)	Control Acc.
C²	0.120	50% (chance)
C³ w/ deterministic control	0.191	79.1%
C³ w/ internal prediction	0.119	77.7%

Internal prediction strategy has the best trade off between reconstruction error and control accuracy.



Effect of shifting the control audio signal on control accuracy.

Downstream tasks.



Real world videos decomposition.

Mode	Acc. (Transp.)	Acc. (Occl.)
I3D – pure video	59.5	59.5
I3D	22.1	21.3
CentrifugePC + I:	34.4	21.5
C² + I3D	40.1	24.7

Action recognition results.