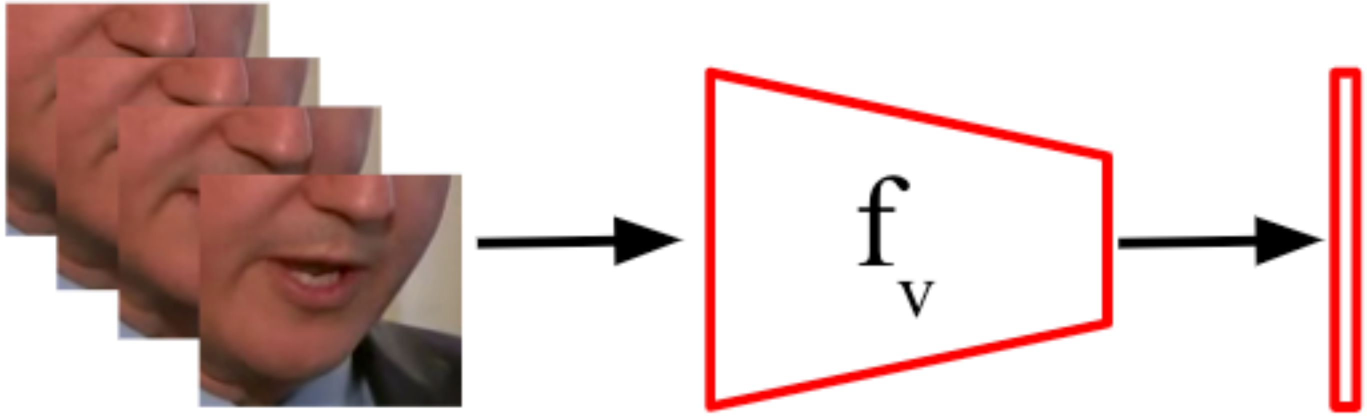
Audio-Visual Predictive Coding for Self-Supervised Visual Representation Learning

Mani Kumar[1], Michel Valstar[1], Michael Pound[1], Timo Giesbrecht[2]

[1]University of Nottingham, [2]Unilever R&D Port Sunlight, UK

- **Problem Statement:** To learn a **visual representation function (f_v)** from **unlabeled video data**



Directly Supervised Representation Learning

- Labeled Data: $\{X, Y\}$

$$X \xrightarrow{f} Y$$

Self-Supervised Representation Learning

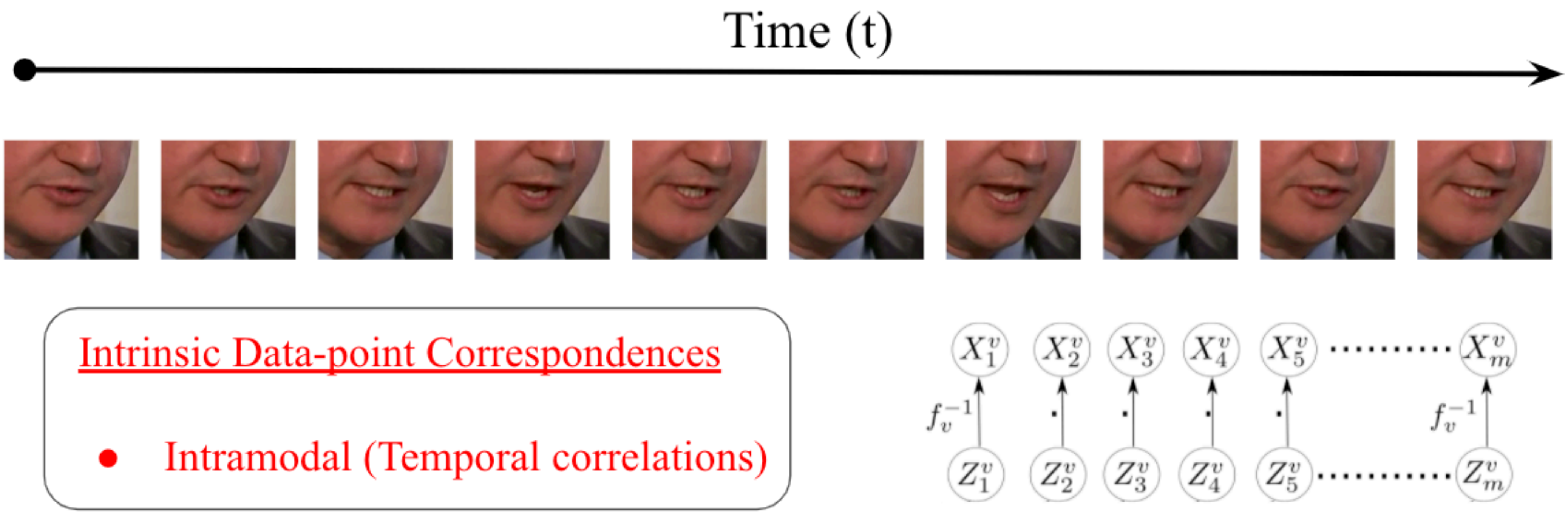
- Unlabeled Data: $\{X\}$

$\Rightarrow$ Proxy learning task: $\{X, \widehat{Y}\}$

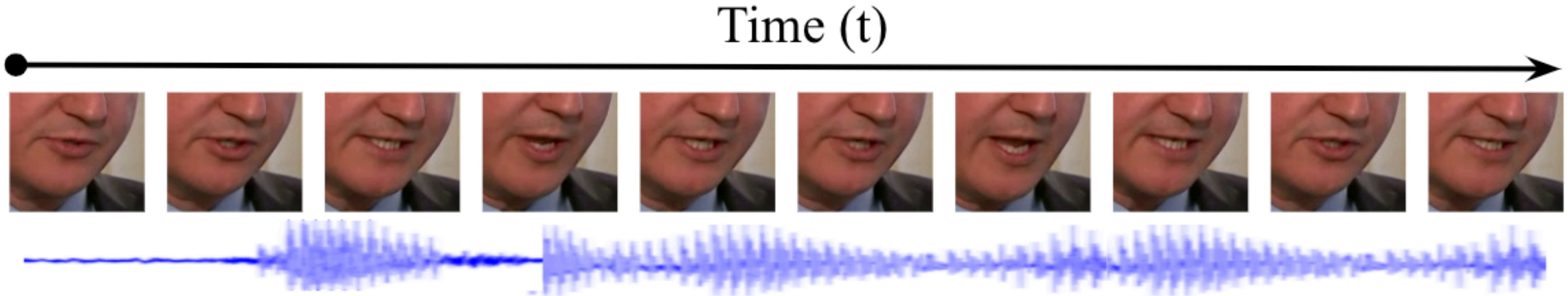
$$X \xrightarrow{\widehat{f}} \widehat{Y}$$

Unlabeled Data Points: Intrinsic Correspondences

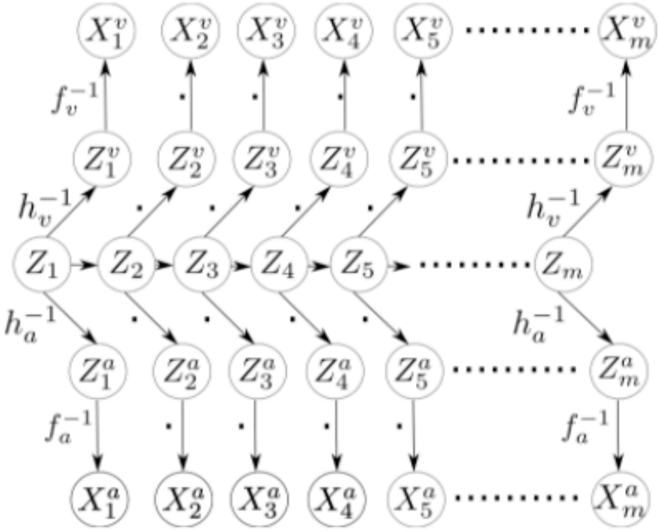
Data-points as i.i.d (independent and identically distributed) samples

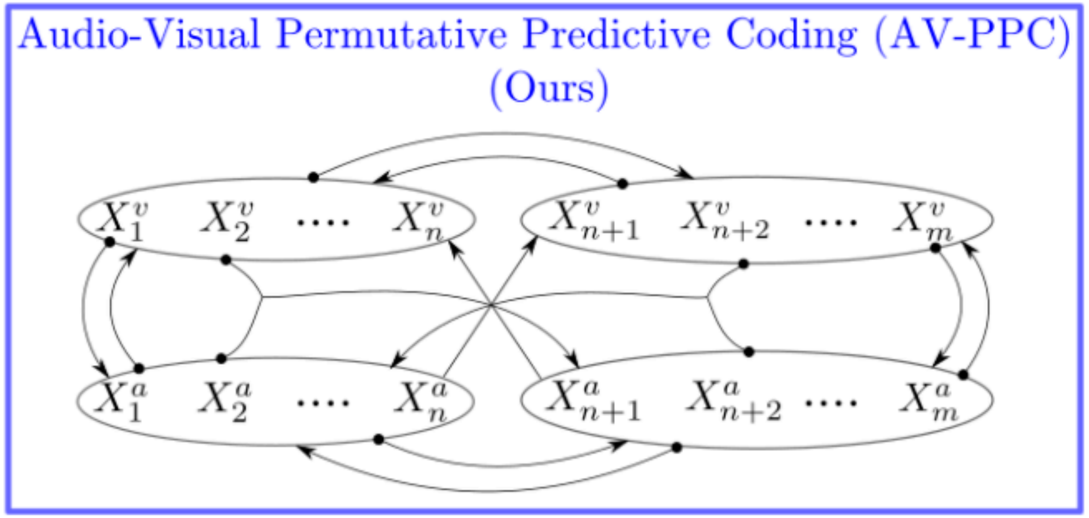

Time (t)
Intrinsic Data-point Correspondences
• Intramodal (Temporal correlations)
X_1^v X_2^v X_3^v X_4^v X_5^v X_m^v
f_v^{-1} f_v^{-1}
Z_1^v Z_2^v Z_3^v Z_4^v Z_5^v Z_m^v

Audio-Visual Predictive Coding

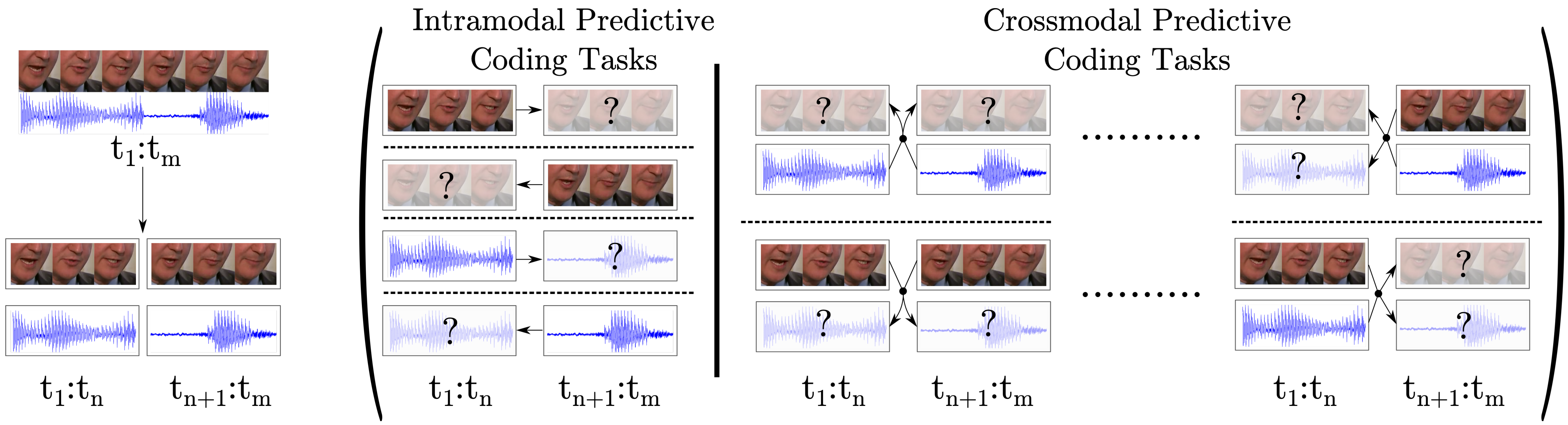


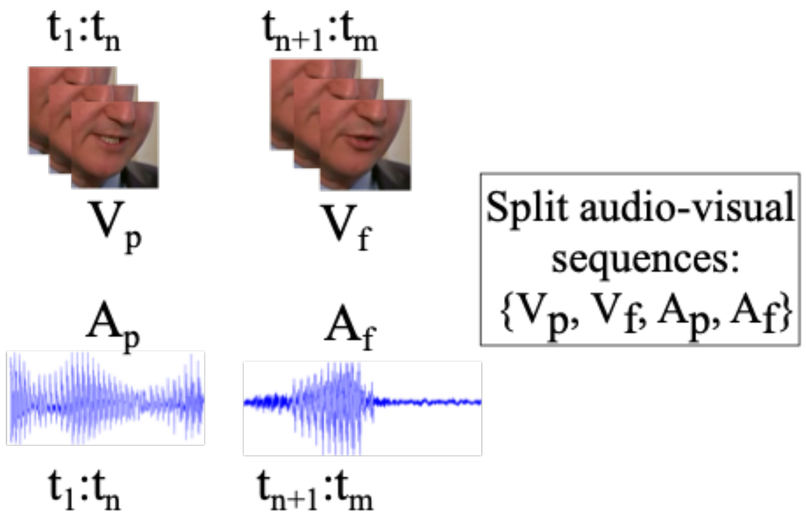
Exploiting temporal and crossmodal correspondences jointly





Audio-Visual Permutative Predictive Coding

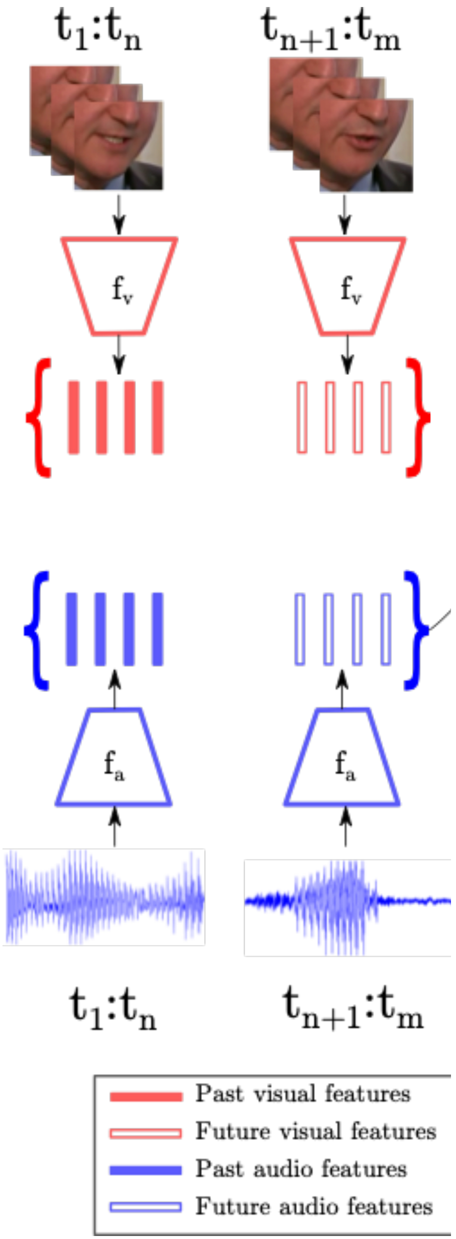




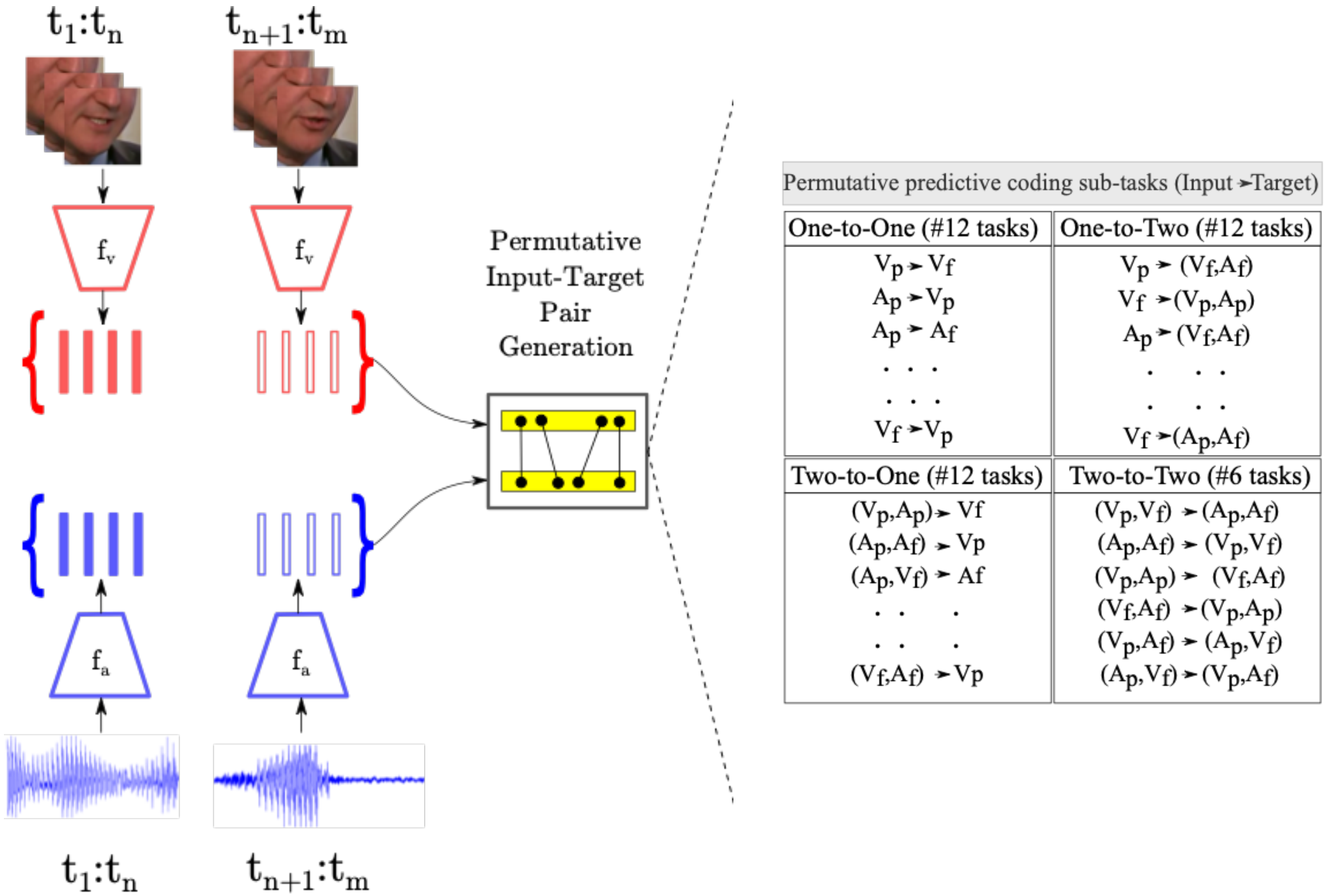
Permutative predictive coding sub-tasks (Input ➢ Target)

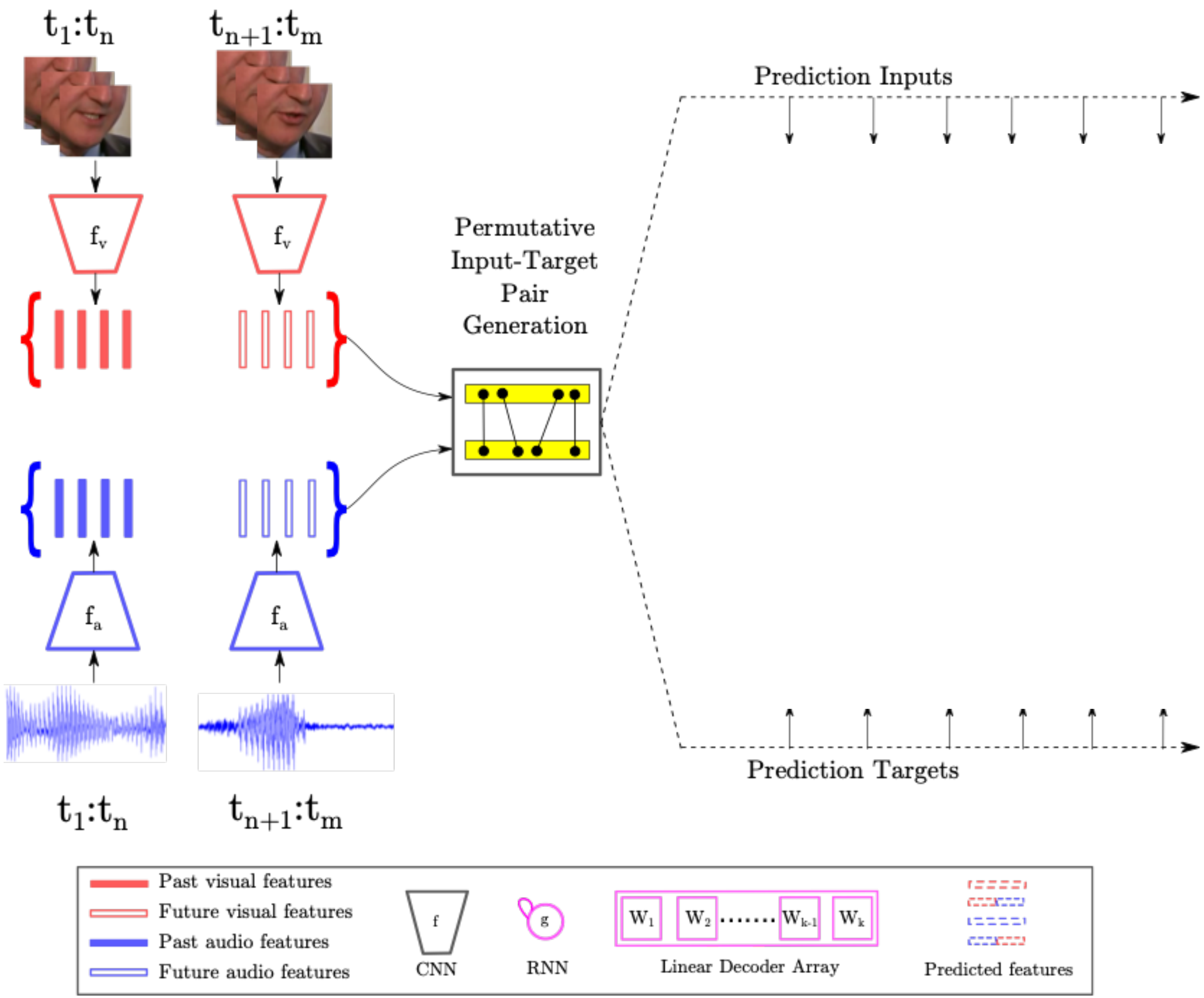
One-to-One (#12 tasks)	One-to-Two (#12 tasks)
V_p ➢ V_f	V_p ➢ (V_f,A_f)
A_p ➢ V_p	V_f ➢ (V_p,A_p)
A_p ➢ A_f	A_p ➢ (V_f,A_f)
. . .	. . .
. . .	. . .
V_f ➢ V_p	V_f ➢ (A_p,A_f)

Two-to-One (#12 tasks)	Two-to-Two (#6 tasks)
(V_p,A_p) ➢ Vf	(V_p,V_f) ➢ (A_p,A_f)
(A_p,A_f) ➢ Vp	(A_p,A_f) ➢ (V_p,V_f)
(A_p,V_f) ➢ Af	(V_p,A_p) ➢ (V_f,A_f)
. . .	(V_f,A_f) ➢ (V_p,A_p)
. . .	(V_p,A_f) ➢ (A_p,V_f)
(V_f,A_f) ➢ Vp	(A_p,V_f) ➢ (V_p,A_f)

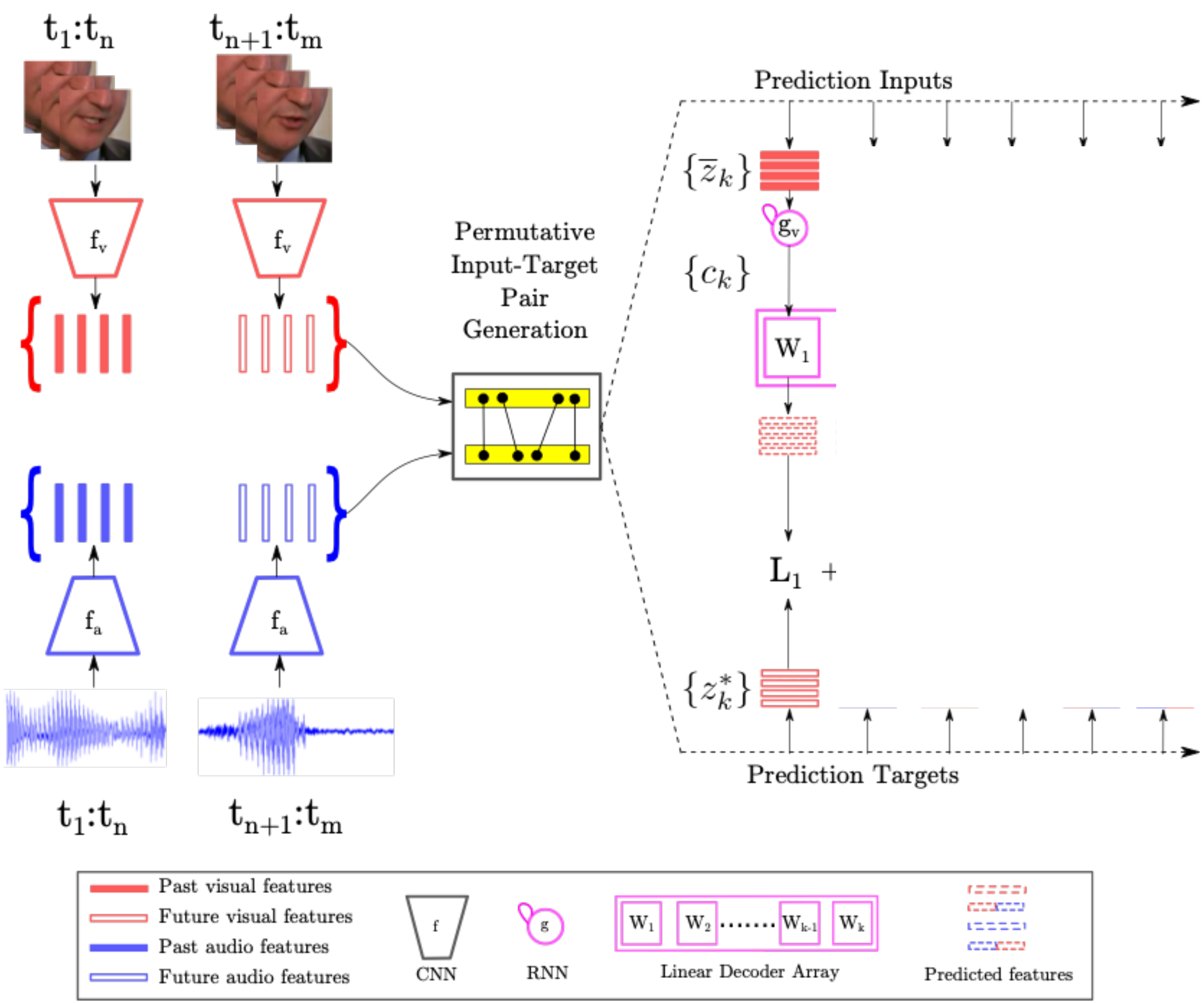


t1:tn
tn+1:tm
fv
fv
fa
fa
t1:tn
tn+1:tm
Past visual features
Future visual features
Past audio features
Future audio features

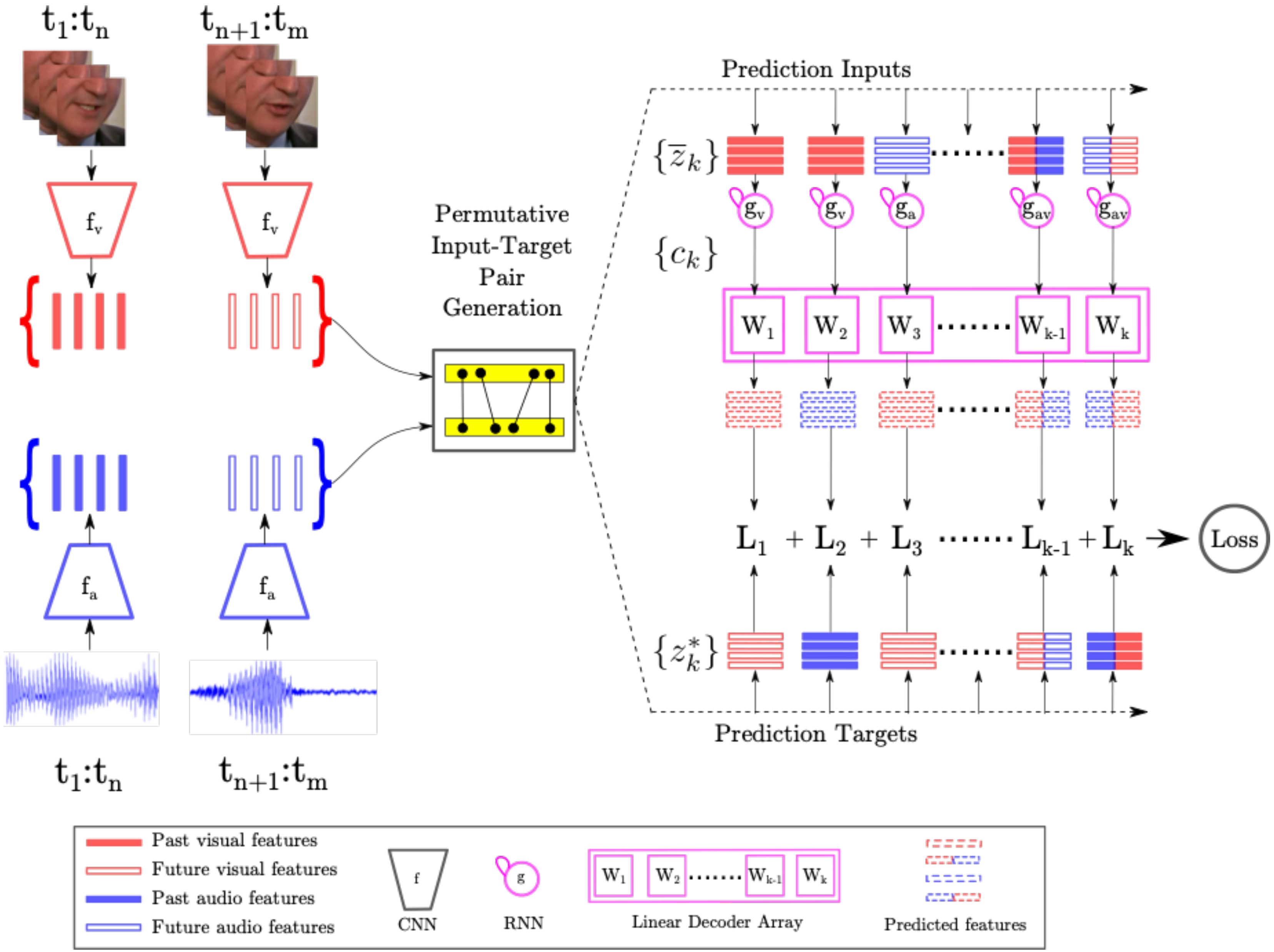

t_1:t_n
t_{n+1}:t_m
f_v
f_v
Permutative Input-Target Pair Generation
f_a
f_a
t_1:t_n
t_{n+1}:t_m
Permutative predictive coding sub-tasks (Input ➢Target)
One-to-One (#12 tasks)
V_p ➢ V_f
A_p ➢ V_p
A_p ➢ A_f
V_f ➢ V_p
One-to-Two (#12 tasks)
V_p ➢ (V_f,A_f)
V_f ➢ (V_p,A_p)
A_p ➢ (V_f,A_f)
V_f ➢ (A_p,A_f)
Two-to-One (#12 tasks)
(V_p,A_p) ➢ Vf
(A_p,A_f) ➢ Vp
(A_p,V_f) ➢ Af
(V_f,A_f) ➢ Vp
Two-to-Two (#6 tasks)
(V_p,V_f) ➢ (A_p,A_f)
(A_p,A_f) ➢ (V_p,V_f)
(V_p,A_p) ➢ (V_f,A_f)
(V_f,A_f) ➢ (V_p,A_p)
(V_p,A_f) ➢ (A_p,V_f)
(A_p,V_f) ➢ (V_p,A_f)



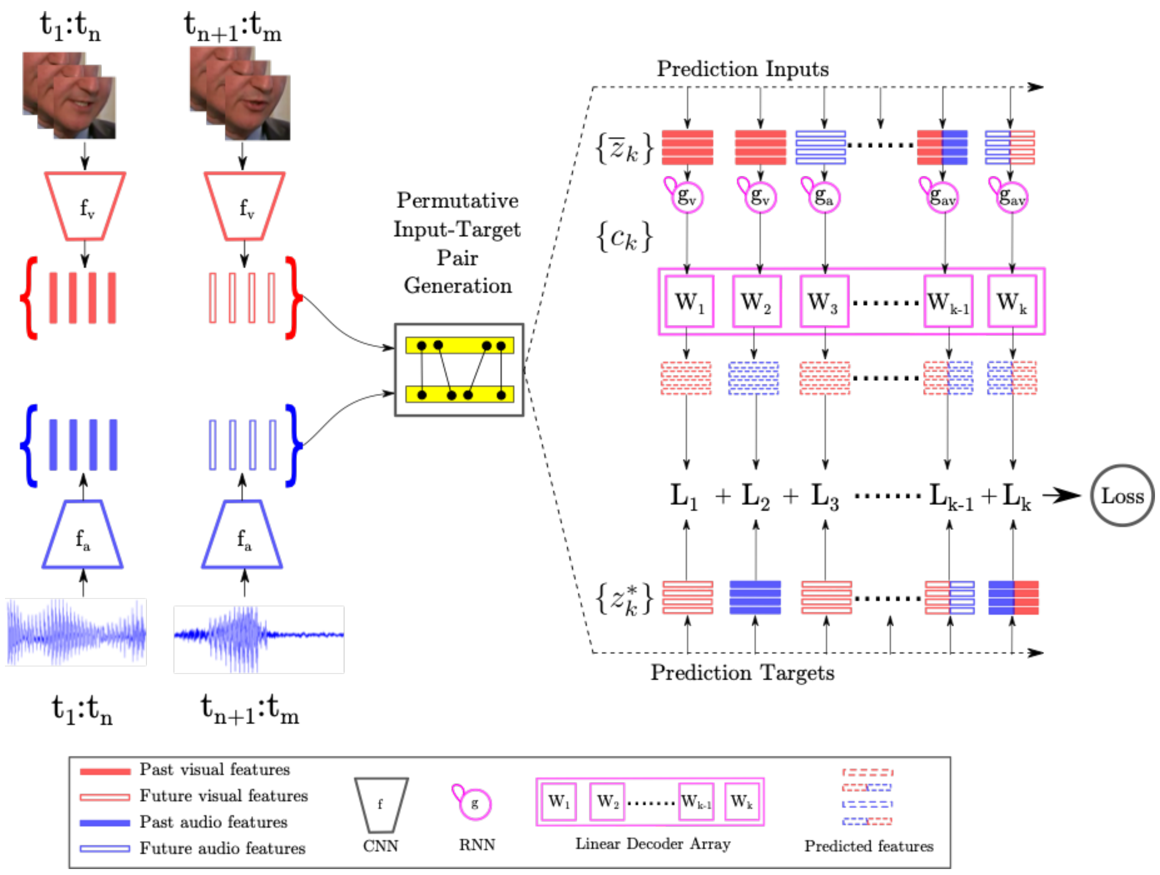
t_1:t_n
t_{n+1}:t_m
f_v
f_v
f_a
f_a
t_1:t_n
t_{n+1}:t_m
Permutative Input-Target Pair Generation
Prediction Inputs
Prediction Targets
Past visual features
Future visual features
Past audio features
Future audio features
f
CNN
g
RNN
W_1 W_2 .... W_{k-1} W_k
Linear Decoder Array
Predicted features



t1:tn
tn+1:tm
fv
fa
Permutative
Input-Target
Pair
Generation
Prediction Inputs
{z̄k}
gv
{ck}
W1
L1 +
{z*k}
Prediction Targets
Past visual features
Future visual features
Past audio features
Future audio features
CNN
RNN
Linear Decoder Array
Predicted features



t_1:t_n
t_{n+1}:t_m
f_v
f_a
Permutative
Input-Target
Pair
Generation
Prediction Inputs
$\{\bar{z}_k\}$
g_v
g_a
g_{av}
$\{c_k\}$
W_1 W_2 W_3 W_{k-1} W_k
$L_1 + L_2 + L_3 \cdots\cdots L_{k-1} + L_k$
Loss
$\{z_k^*\}$
Prediction Targets
Past visual features
Future visual features
Past audio features
Future audio features
CNN
RNN
Linear Decoder Array
Predicted features



Contrastive Learning: InfoNCE Loss

(Noise Contrastive Estimation)

$$I(z_k^*; c_k) = \sum_{z_k^*, c_k} p(z_k^*, c_k) log(\frac{p(z_k^*|c_k)}{p(z_k^*)})$$

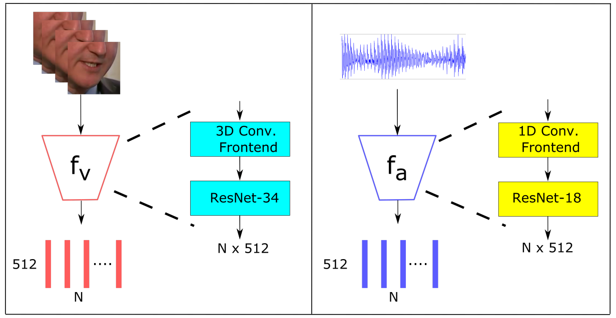
$$f_k(z_k^*, c_k) \propto \frac{p(z_k^*|c_k)}{p(z_k^*)}$$

$$f_k(z_k^*, c_k) = exp(z_k^{*T}.W.c_k)$$

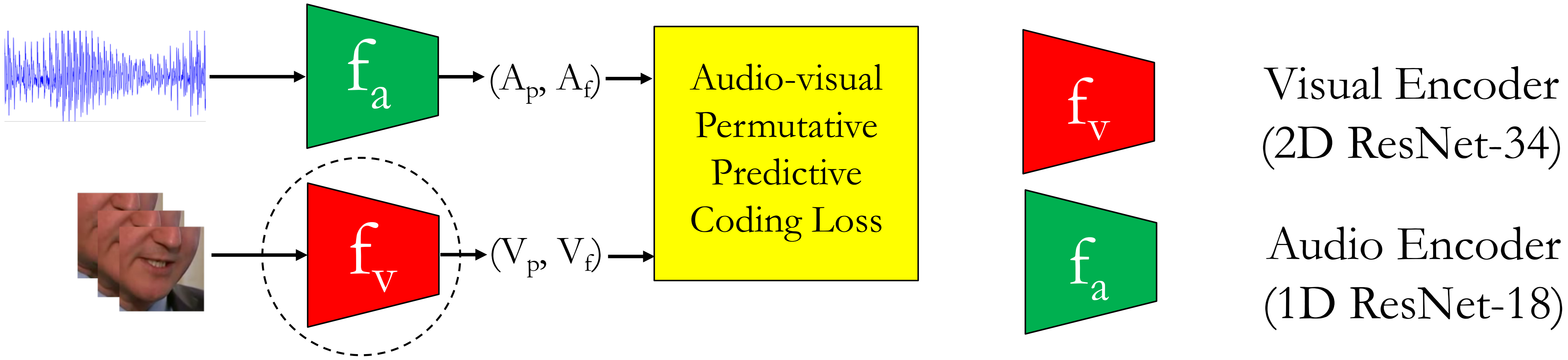
$$L_i = -E_B[log \frac{f_k(z_k^*, c_k)}{\sum_{z_j \in B} f_k(z_j, c_k)}]$$

where B is a mini batch of N samples,

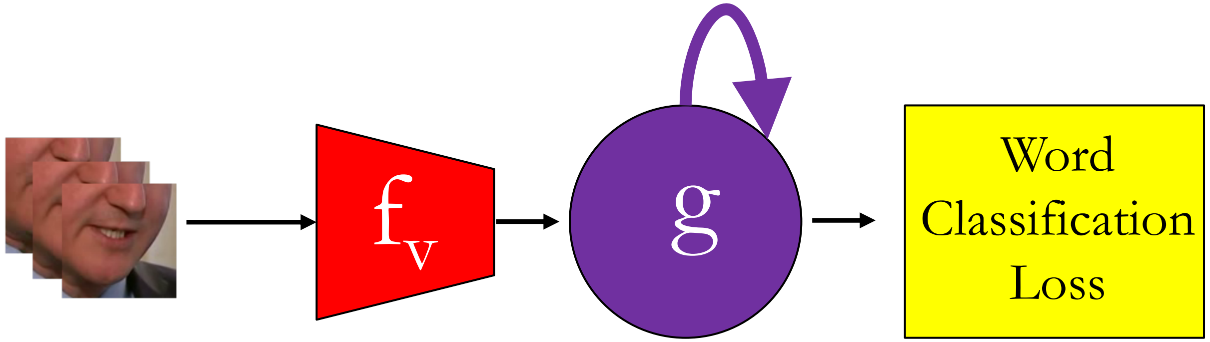
- with 1 positive sample and N-1 negative samples



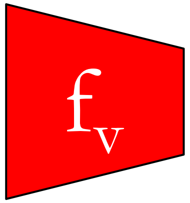
Summary of Self-Supervised Learning Stage



Downstream Evaluation Task: Lip-Reading



Task: Predict the word uttered in a video
Dataset: LRW with 500-word class labels
Metric: Word Classification Rate (WCR)



Visual Encoder
(2D ResNet-34)

Temporal Model
(GRU/Temporal Conv)

Evaluation Protocol: Measure WCR

A. before finetuning the visual encoder
B. after finetuning the visual encoder
    A. using the entire train data and
    B. using small amounts of train data.

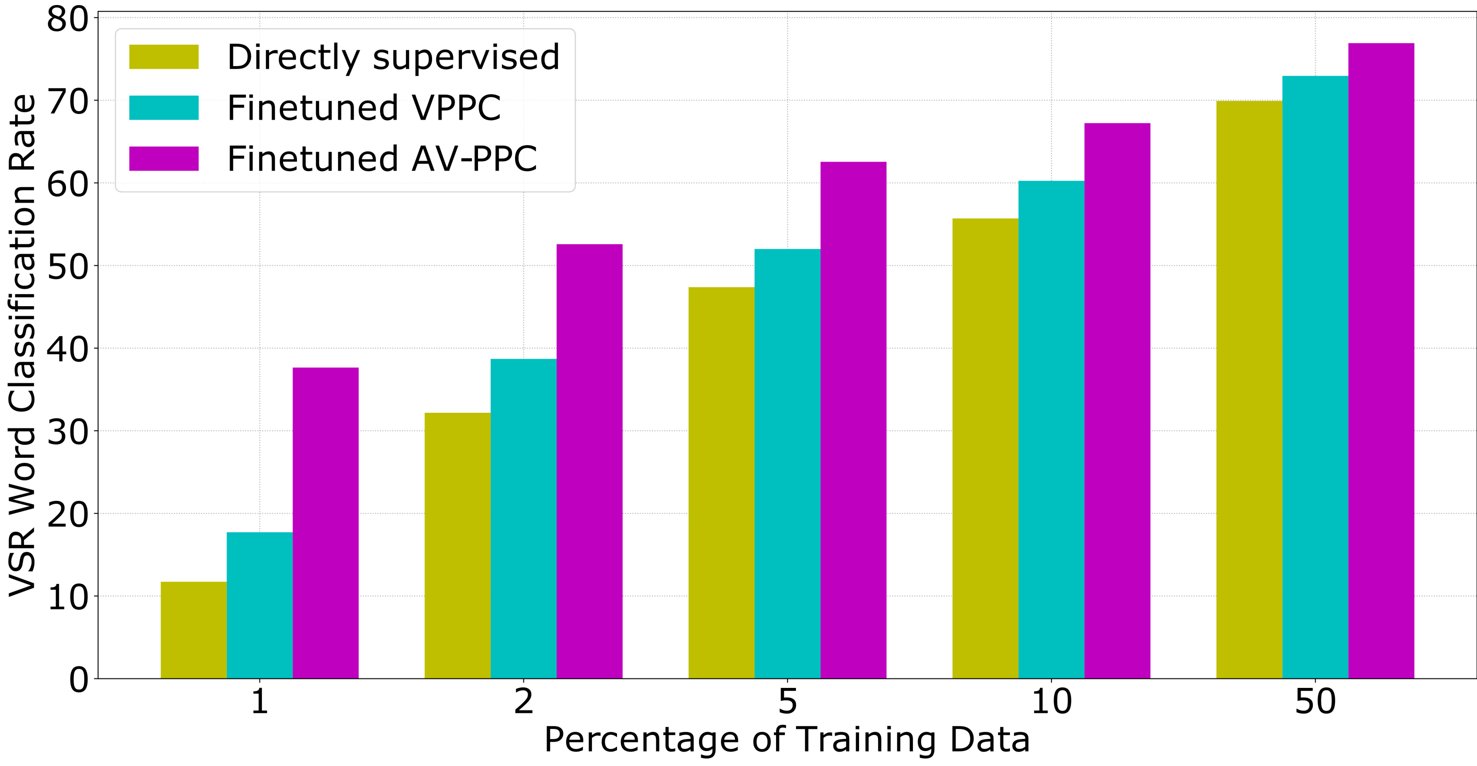
Performance of Different Proxy Tasks on the Lip-Reading Task (Word Classification Rates)

Proxy Task	**Using Temporal Conv**	**Using GRU**
AV Synchronization	50.70 (74.17)	55.26 (76.92)
Time-Arrow	52.42 (75.80)	59.88 (78.26)
AV Correspondence	56.22 (74.23)	61.90 (77.90)
Vis. Permutative Pred. Coding	60.77 (77.95)	67.62 (81.76)
AudVis. Permutative Pred. Coding (ours)	**76.47 (80.44)**	**80.30 (83.16)**

Data-Efficiency Evaluation

- Number of labeled instances required to learn lip-reading task

- With 1% of train data (10 instances per word class),
    - Our method: 38% WCR
    - Fully-supervised: 11% WCR



Take-home Idea

A potential approach to unsupervised representation learning:

Leveraging rich intrinsic data-point correspondences
- temporal and cross-modal semantic correlations -
as natural supervision signals in the self-supervised setting.

Thank You