Unsupervised Disentanglement of Content and Style via Variance-Invariance Constraints

Yuxuan Wu[♭], Ziyu Wang[♯♭], Bhiksha Raj[♮♭], Gus Xia[♭♯]

[♭] MBZUAI [♯] New York University Shanghai [♮] Carnegie Mellon University

Presenter: Yuxuan Wu

Motivation

- Abstraction is essential and natural in human intelligence

- Abstraction can be modeled as representation disentanglement:
  - ***Content***---the information to be communicated
  - ***Style***---the particular way content is “loaded”

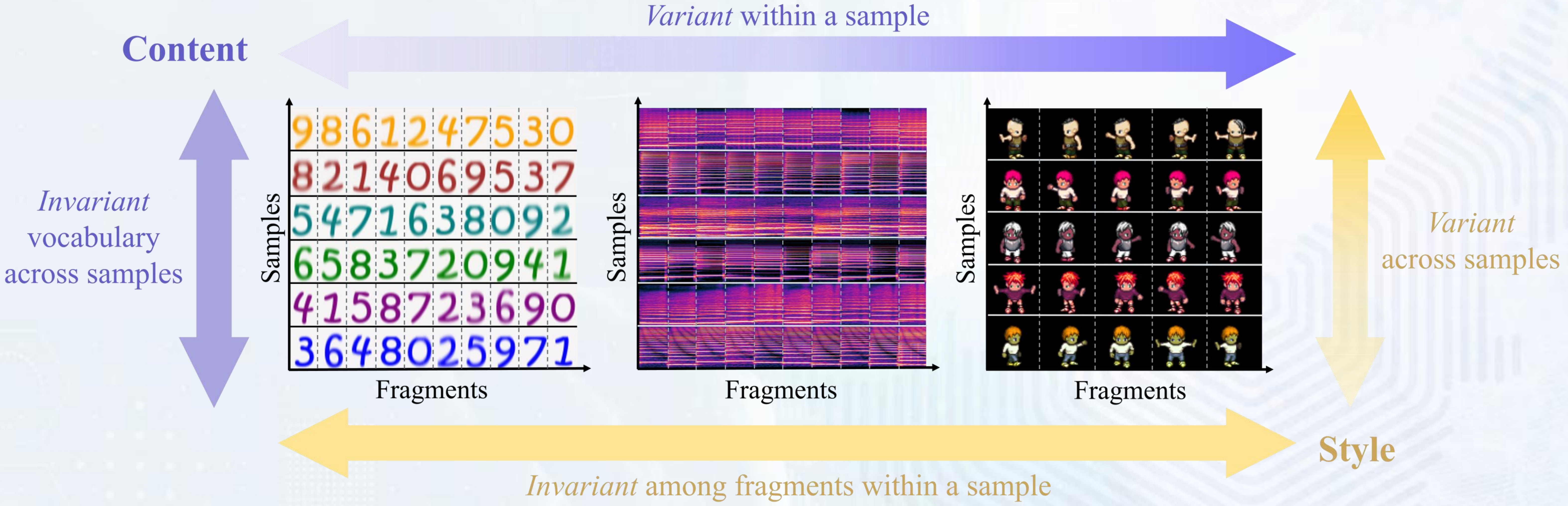
Motivation

- Common practice of content-style disentanglement:
  - Strongly or weakly supervised
    - Pretrained representations, explicit labels, or even paired data
  - Rely on domain-specific knowledge

- A Human-like learning process can be more natural!
  - Domain-general approach
  - Generalizable to new styles
  - More interpretable

The Meta-Level Content-Style Difference

- Because of different “roles” in communication, they have **distinct patterns of variation**

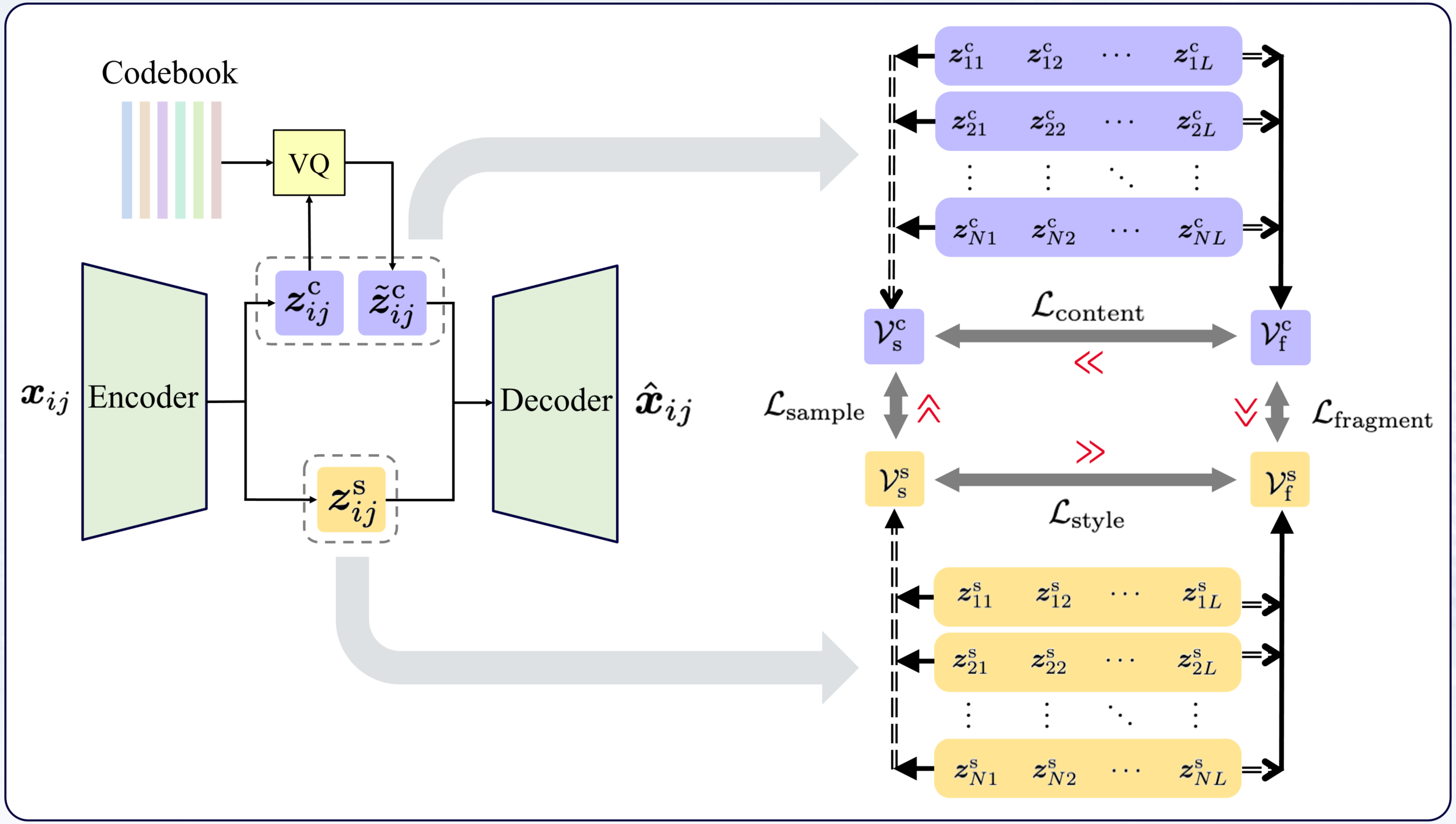


I’m informative and universal!

I’m consistent and unique!

V3: Variance-Versus-Invariance

- Learning content and style through V3, on a branched autoencoder



Left: a VQ content branch, and a regular style branch

Right: the V3 constraints

Double-dashed arrows (==>): measuring the variability

Solid arrows (→): Taking the average

Experiments Results

- V3 achieves **better disentanglement** of content and style over unsupervised baselines

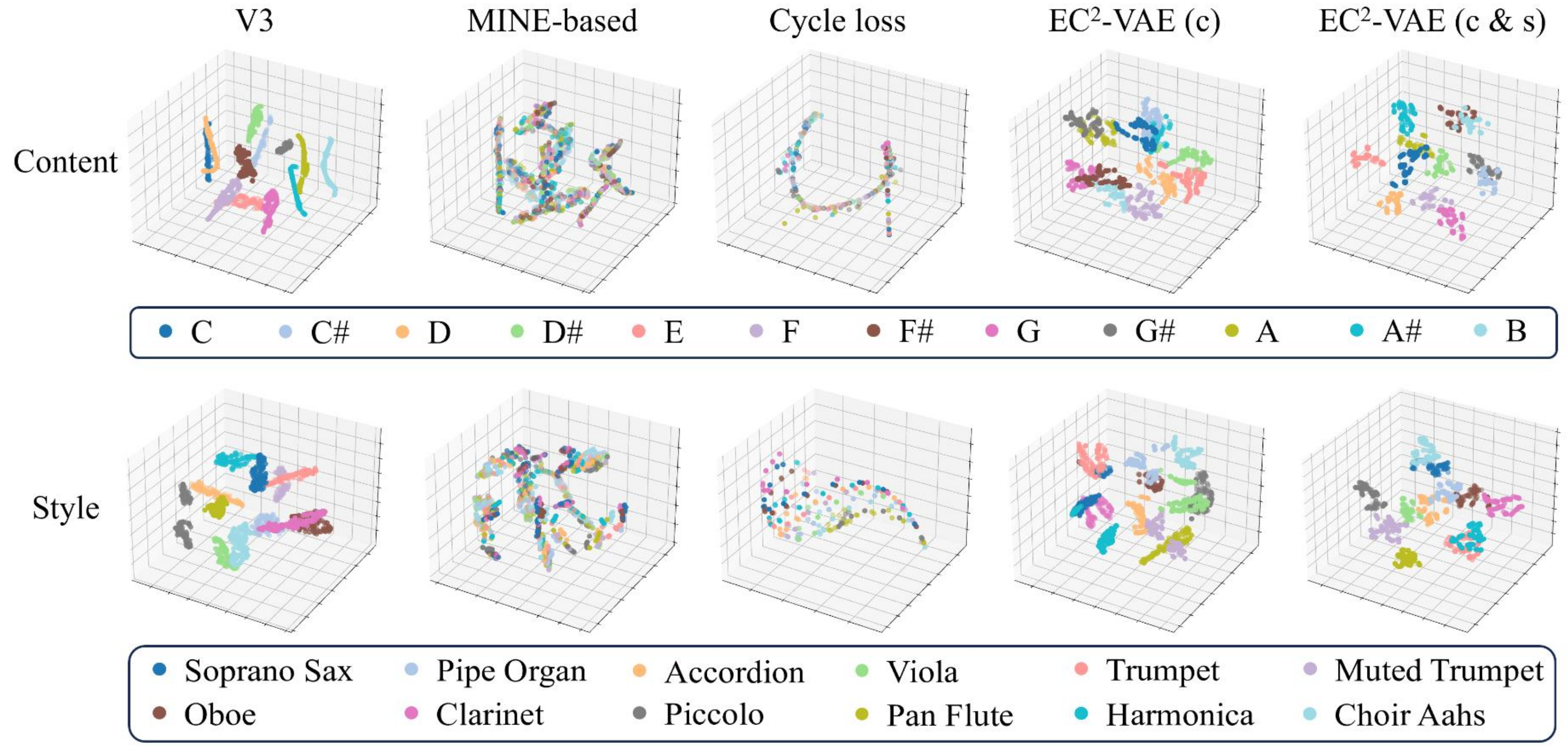


Figure 8: t-SNE visualization of the learned pitch (content) and timbre (style) representations on InsNotes when there is no codebook redundancy ($K = 12$).

Table 3: Linear probing accuracies (in %) for content (digit) classification on SVHN.

Method	K	$z^c \uparrow$	$z^s \downarrow$
V3	20	**40.6**	**18.5**
MINE-based	20	36.0	20.8
Cycle loss	20	16.8	21.2
β-VAE	-	21.8	
Raw input	-	21.4	
EC^2-VAE (c)	-	97.0	21.2

Table 5: Linear probing accuracies (in %) for content (phoneme) classification on Libri100.

Method	K	$z^c \uparrow$	$z^s \downarrow$
V3	80	**52.1**	**40.4**
MINE-based	80	28.6	51.6
Cycle loss	80	16.1	50.5
β-VAE	-	11.0	
Raw input	-	31.8	
EC^2-VAE (c)	-	78.1	18.2

Experiments Results

- V3 surpasses weakly supervised models in few-shot **OOD styles generalization**

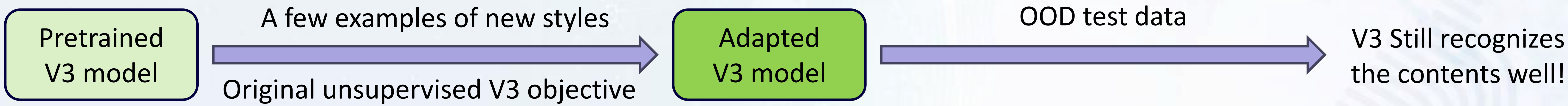


Table 7: Content classification accuracies (in %) on data with OOD styles.

Pretraining		Continuous Training		PhoneNums				InsNotes			
Method	Supervision	Supervision	Self-boost	0-shot	1-shot	5-shot	10-shot	0-shot	1-shot	5-shot	10-shot
V3	No	No	No	57.8	91.3	**97.1**	**99.0**	90.5	**97.6**	**97.8**	**99.2**
EC^2-VAE (c)	Yes	No	No	**84.2**	**92.1**	92.2	92.7	87.1	87.2	89.4	91.2
EC^2-VAE (c)	Yes	No	Yes	**84.2**	91.8	92.1	92.4	87.1	94.6	95.0	95.1
CNN Classifier	Yes	No	No	59.5	59.5	59.5	59.5	**92.6**	92.6	92.6	92.6
CNN Classifier	Yes	No	Yes	59.5	80.2	82.2	82.7	**92.6**	87.6	85.9	85.3
EC^2-VAE (c)	Yes	Yes	No	84.2	94.6	98.8	99.2	87.1	97.7	98.9	99.8
CNN Classifier	Yes	Yes	No	59.5	81.2	82.4	83.5	92.6	91.9	91.3	89.1

Experiments Results

- The learned content symbols have better **interpretability**

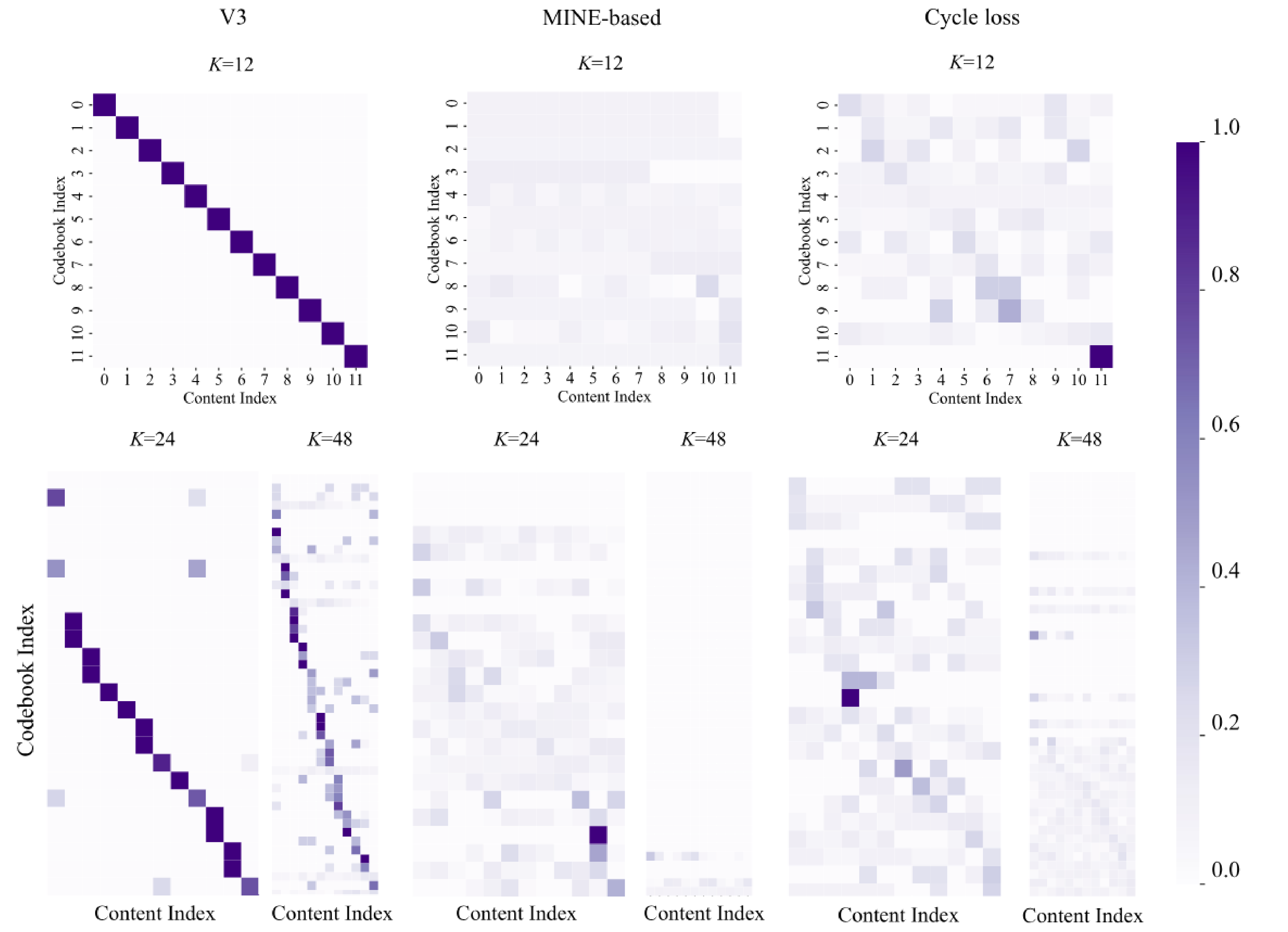


Figure 13: Confusion matrices of learned codebooks on InsNotes. The horizontal axes show pitch labels from “C” to “B”, and the vertical axes show codebook atoms sorted by ground truth pitch labels.

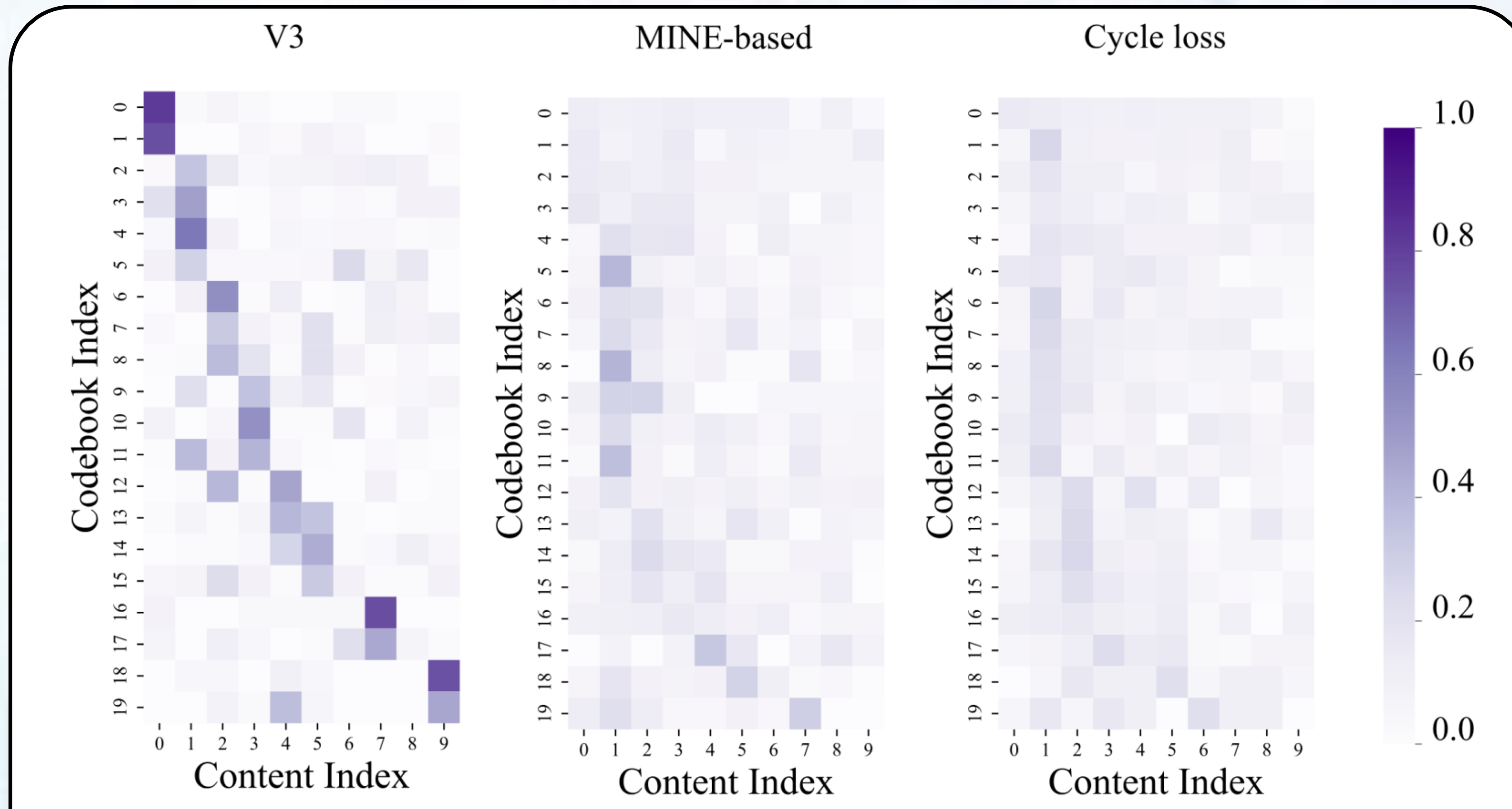


Figure 14: Confusion matrices of learned codebooks on SVHN. The horizontal axes show digit labels from “0” to “9”, and the vertical axes show codebook atoms sorted by ground truth digit labels.

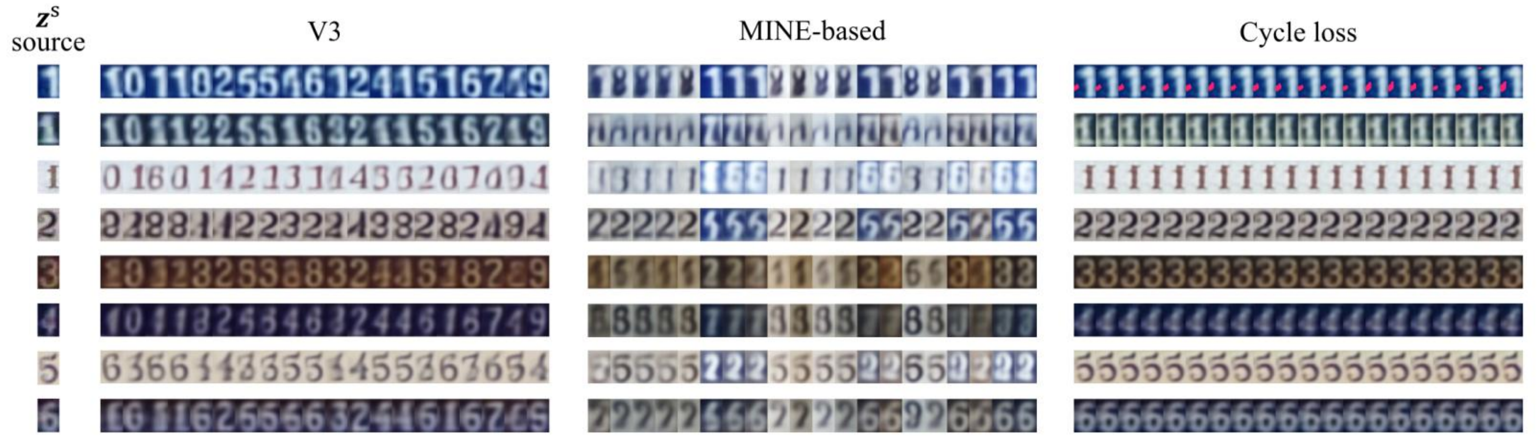


Figure 3: Comparison of generated images by recombining z^s from given sources in SVHN and all z^c in the learned codebook.

Conclusions

- We present ***V3***, a **domain-general** and **intuitive** method for unsupervised content-style disentanglement

  - V3 exhibits good **disentanglement** performance on tasks of different domains
  - V3 shows better **generalizability** on OOD styles compared to supervised methods
  - V3 achieves high **interpretability** of learned content symbols

Thanks for watching!

✌️ ✌️ ✌️