IMML Lab
Interactive Multimodal
Machine Learning Lab

CVPR JUNE 3-7, 2026 DENVER COLORADO

Cross-Modal Emotion Transfer for Emotion Editing in Talking Face Video

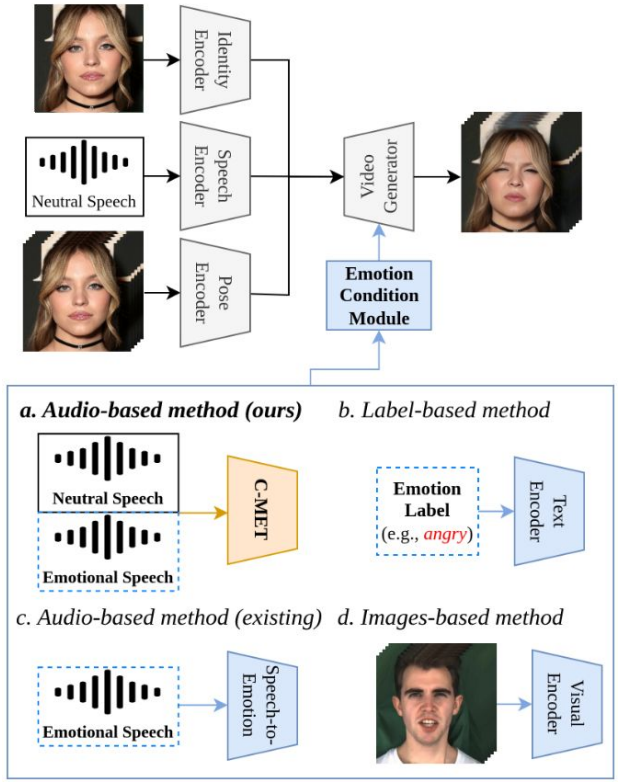
Chanhyuk Choi, Taesoo Kim, Donggyu Lee, Siyeol Jung, Taehwan Kim

Ulsan National Institute of Science and Technology (UNIST)

Emotion Editing in Talking Face Video

Emotional Talking Face Generation = (1) + (2)

(1) Generating a talking face
(2) **Editing facial expressions with an emotion prompt such as label, audio, and images**

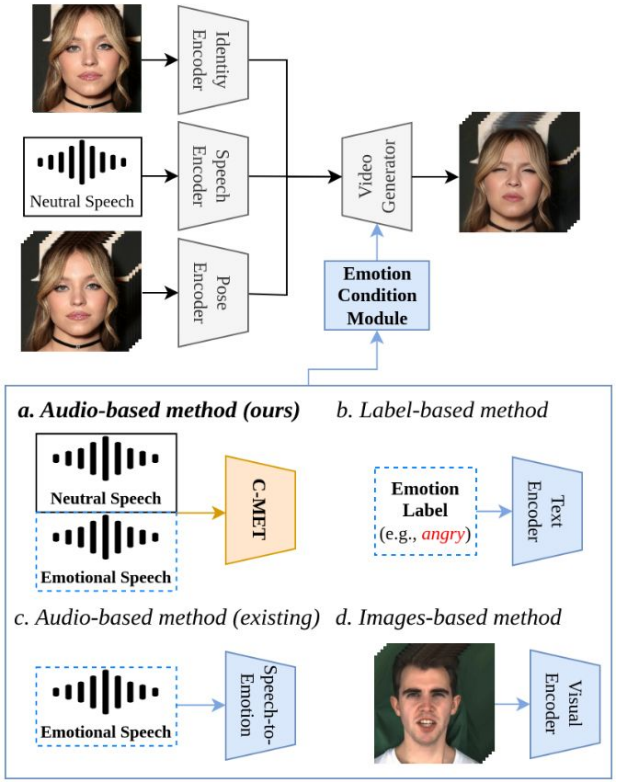


Emotion Editing in Talking Face Video

Emotional Talking Face Generation = (1) + (2)

(1) Generating a talking face
(2) Editing facial expressions with an emotion prompt such as label, audio, and images

Existing Methods		
Emotion Prompt	*Representative Method*	***Key Limitation***
🏷️ Label	*EAT (ICCV 2023)*	Only discrete basic emotions
🔊 Audio	*FLOAT (ICCV 2025)*	Emotion entangled with linguistic content
🖼️ Image	*EDTalk (ECCV 2024)*	Requires clean frontal-view reference



Emotion Editing in Talking Face Video

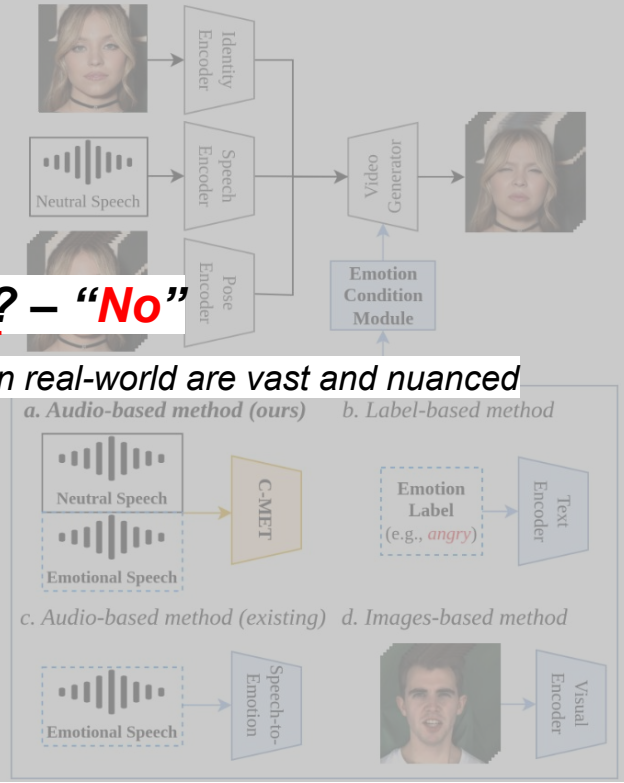
Emotional Talking Face Generation = (1) + (2)

(1) Generating a talking face
(2) Editing facial expressions with an emotion prompt such as label, audio, and images

Can they model extended emotions (e.g, sarcasm)? – "No"

why is expressing these emotions important? – emotions in real-world are vast and nuanced

Emotion Prompt	*Representative Method*	***Key Limitation***
🏷️ Label	*EAT (ICCV 2023)*	Only discrete basic emotions
🔊 Audio	*FLOAT (ICCV 2025)*	Emotion entangled with linguistic content
🖼️ Image	*EDTalk (ECCV 2024)*	Requires clean frontal-view reference



Can We Model Extended Emotions? – "Yes"

Emotion prompt:

"So why don't you give me your number?"

Neutral
(Annotated Emotion)

Sarcastic
(Actual Emotion)

Source: MELD (dialogue 5, utterance 8) [2]

Input Video *Edited Results*

(neutral) **Ours*** EAT FLOAT EDTalk

* , exhibiting a more pronounced widening of the lip corners compared to the baselines.

[1] Cross-Modal Emotion Transfer for Emotion Editing in Talking Face Video (CVPR 2026)
[2] MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations (ACL 2019)

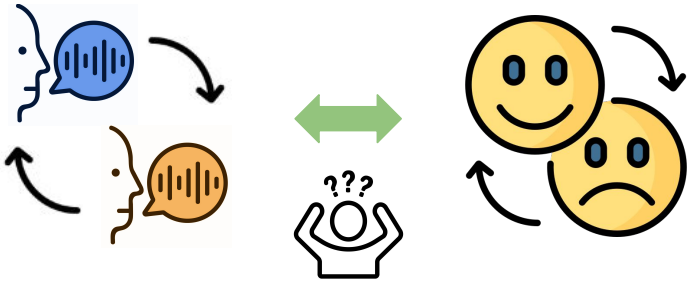
Motivation: Leveraging Expressive TTS Methods

- Leverage synthetic emotional speeches:

  Recent expressive TTS (e.g., EmoKnob, Gemini TTS) can synthesize any target emotion — including sarcasm, empathy, or charisma

- **Key open problem:**

  Bridging the domain gap between audio and visual emotion representations (cross-modal mapping) remains difficult



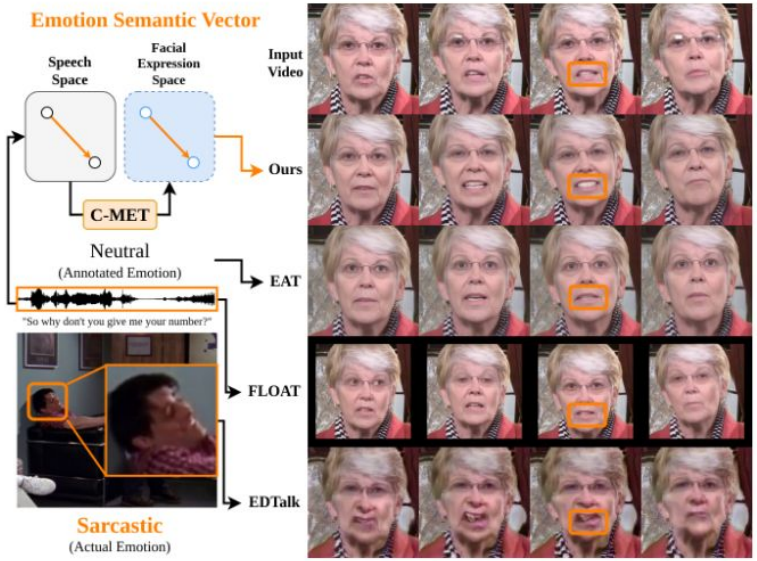
Cross-Modal Emotion Transfer (C-MET)

➢ Key Idea:

- Define **emotion semantic vectors** in speech and facial expression space, respectively*.
- Map these vectors by **cross-modal emotion transfer (C-MET)** learning**.

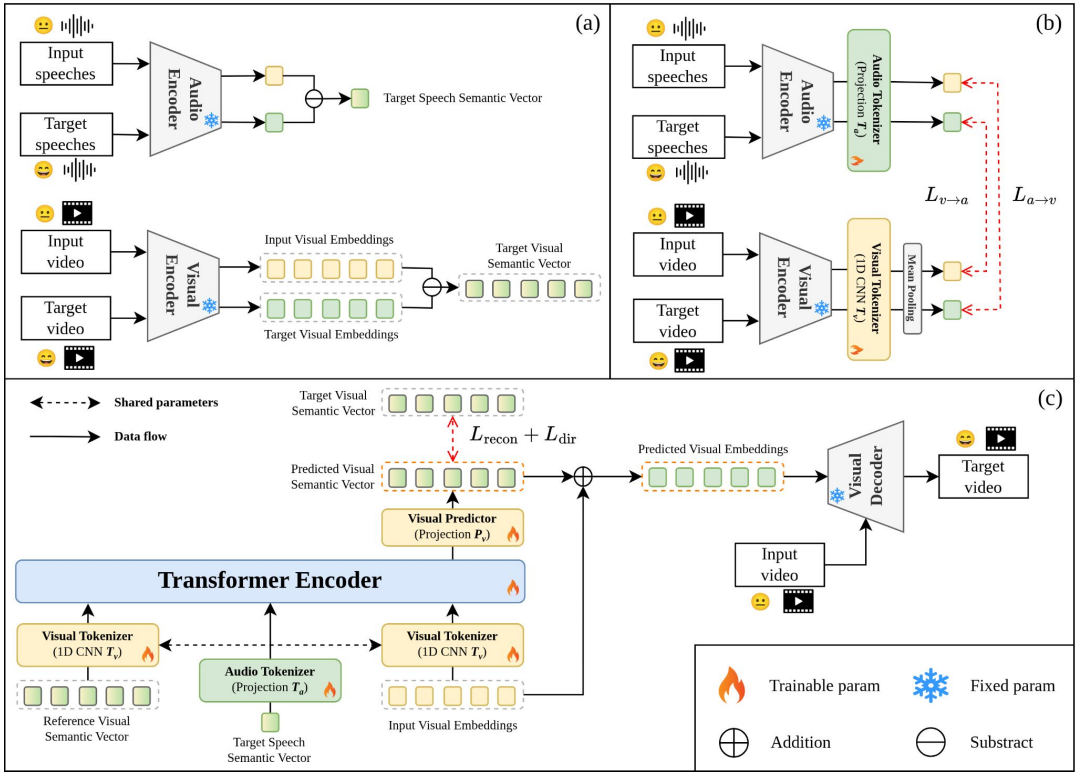
➢ Contributions:

- 🥇 First method for extended emotion editing via cross-modal audio-visual semantic vectors
- 🔌 Lightweight plug-and-play module for existing disentanglement-based models
- 📈 +14% emotion accuracy on MEAD & CREMA-D over state-of-the-art

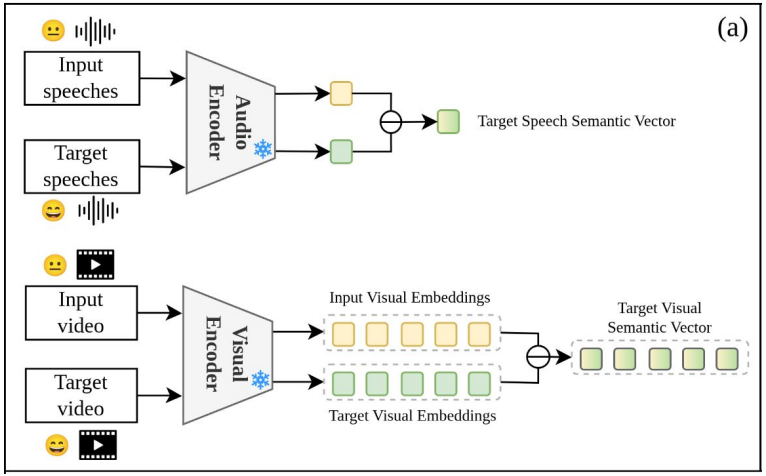


* In disentangled space, this vector is obtained by subtracting the embedding of two different emotional expression with sampling-and-averaing strategy.
** We leverage the Transformer encoder architecture to enable cross-modal transfer.

Overview of Cross-Modal Emotion Transfer (C-MET)



Cross-Modal Emotion Transfer (C-MET)



(a) Formulate Semantic Vectors

1) Extract input and target embeddings using pretrained audio* and visual encoders**,
2) Compute the semantic vectors by subtracting the target embeddings from the inputs***.

input emotion (i) and target emotion (j)

$$f_a^{i \to j} = f_a^j - f_a^i$$

$$f_{v,1:T}^{i \to j} = f_{v,1:T}^j - f_{v,1:T}^i$$

* emotion2vec+large [2] is adopted as the pretrained audio encoder.
** The facial expression encoder of EDTalk [3] is used as the pretrained visual encoder.
*** We randomly sample ten samples for each modality and average them to reduce noise and stabilize learning.

[1] Cross-Modal Emotion Transfer for Emotion Editing in Talking Face Video (CVPR 2026)
[2] emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation (ACL Findings 2024)
[3] EDTalk: Efficient Disentanglement for Emotional Talking Head Synthesis (ECCV 2024 oral)

Cross-Modal Emotion Transfer (C-MET)

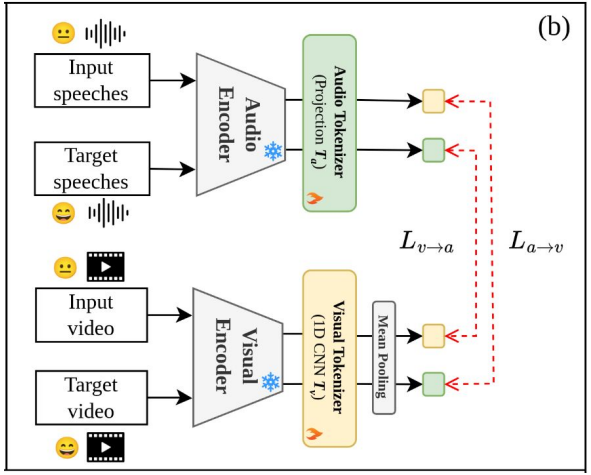
(b) Contrastive Learning on Multimodal Tokens

1) Construct the audio tokenizer* and visual tokenizer** to extract the audio and visual tokens, respectively,
2) Define the multimodal token contrastive loss [3] as:

$$L_{v\to a} = -\sum_{i\in B} \log \frac{e^{(\mathrm{sim}(v^i,a^i)/\tau)}}{e^{(\mathrm{sim}(v^i,a^i)/\tau)} + \sum_{j,i\neq j} e^{(\mathrm{sim}(v^i,a^j)/\tau)}} \tag{1}$$

$$L_{a\to v} = -\sum_{i\in B} \log \frac{e^{(\mathrm{sim}(a^i,v^i)/\tau)}}{e^{(\mathrm{sim}(a^i,v^i)/\tau)} + \sum_{j,i\neq j} e^{(\mathrm{sim}(a^i,v^j)/\tau)}} \tag{2}$$

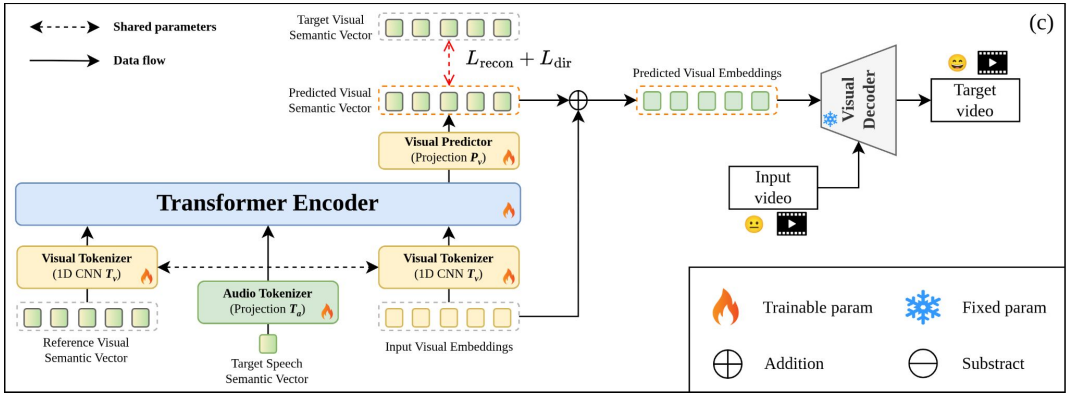
$$L_{\mathrm{cnt}} = \frac{L_{v\to a} + L_{a\to v}}{2} \tag{3}$$



* We construct the visual tokenizer using 1D convolution layers [2], and
** the audio tokenzier using projection layers

[1] Cross-Modal Emotion Transfer for Emotion Editing in Talking Face Video (CVPR 2026)
[2] Identity-preserving talking face generation with landmark and appearance priors (CVPR 2023)
[3] Ma-avt: Modality alignment for parameter-efficient audio-visual transformers (CVPR 2024)

Cross-Modal Emotion Transfer (C-MET)



$$L_{i\to j} = \sum_{t=1}^{T} \left\| f_{v,t}^{i\to j} - \hat{f}_{v,t}^{i\to j} \right\|_2$$

$$L_{j\to i} = \sum_{t=1}^{T} \left\| f_{v,t}^{j\to i} - \hat{f}_{v,t}^{j\to i} \right\|_2$$

$$L_{\text{recon}} = L_{i\to j} + L_{j\to i}$$

$$L_{\text{dir}} = 1 + \frac{\langle \hat{f}_v^{i\to j}, \hat{f}_v^{j\to i} \rangle}{||\hat{f}_v^{i\to j}|| \, ||\hat{f}_v^{j\to i}||}$$

(c) Cross-Modal Emotion Transfer Learning

1) A multimodal transformer encoder [2] is used to regress the target visual semantic vectors, guided by the speech semantic vectors.
2) Since vectors consider forward and reverse, we define tow loss terms as the above.
3) Train the C-MET model with the final loss: $L = L_{\text{recon}} + \lambda_{\text{cnt}} \cdot L_{\text{cnt}} + \lambda_{\text{dir}} \cdot L_{\text{dir}}$

[1] Cross-Modal Emotion Transfer for Emotion Editing in Talking Face Video (CVPR 2026)
[2] Attention is all you need (NIPS 2017)

Quantitative results on MEAD and CREMA-D

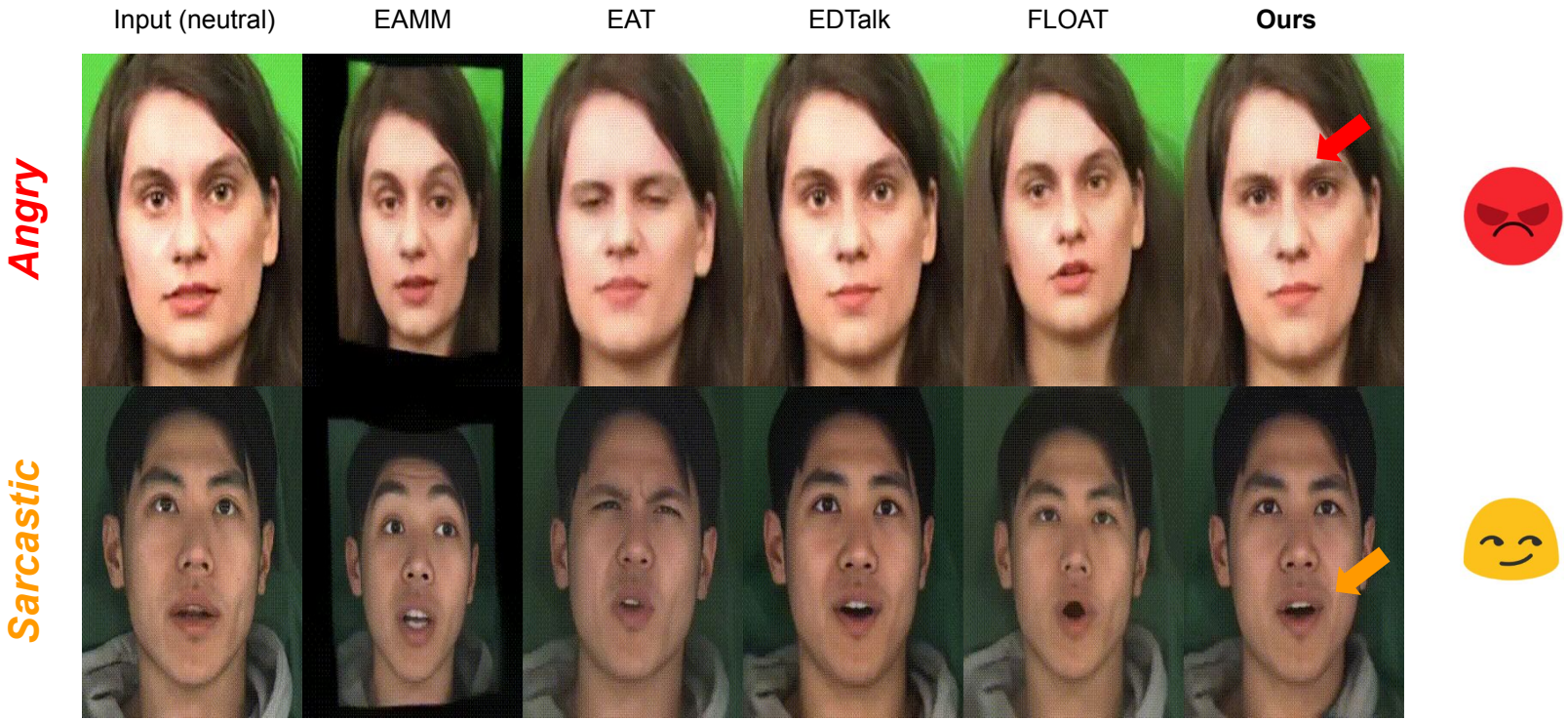
* The model is trained on the MEAD training set [2].
** Then, it is evaluated on the MEAD test set and CREMA-D [3].

Method	Emotion Source Type	MEAD					CREMA-D				
		AITV ↓	FID ↓	FVD ↓	$\text{Sync}_{\text{conf}}$ ↑	Acc_{emo} ↑	AITV ↓	FID ↓	FVD ↓	$\text{Sync}_{\text{conf}}$ ↑	Acc_{emo} ↑
EAMM [24]	Images	3.745	161.602	474.446	6.0609	18.81	6.481	206.168	628.344	4.1134	19.15
EAT [58]	Label	12.575	90.974	330.722	8.0528	41.56	8.055	50.855	320.795	5.9862	39.97
EDTalk [50]	Images	2.827	**76.423**	**293.904**	**8.0529**	41.99	1.590	**42.376**	**288.162**	**6.3569**	29.69
FLOAT [25]	Audio	**1.434**	92.799	368.081	7.1632	13.21	**0.846**	52.933	365.770	4.9860	29.11
C-MET (Ours)	Audio	2.643	90.804	329.862	7.9996	**55.91**	1.561	50.028	309.828	6.2887	**43.47**

Table 1. **Quantitative comparison with state-of-the-art methods.** Each method is evaluated on the MEAD and CREMA-D datasets. To assess emotion editing, we input a neutral talking-face video while varying the emotion source: images (EAMM, EDTalk), label (EAT), and audio (FLOAT, ours). Best and second-best results are shown in **bold** and underline, respectively. For emotion editing in talking-face videos, achieving higher Acc_{emo} is the primary objective, while other perceptual attributes are expected to be preserved with minimal degradation.

[1] Cross-Modal Emotion Transfer for Emotion Editing in Talking Face Video (CVPR 2026)
[2] MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation (ECCV 2020)
[3] CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset (TAFFC 2014)

Qualitative results on CREMA-D and MEAD



Ablation study on training loss

Loss			Metric
L_{recon}	L_{cnt}	L_{dir}	Acc_{emo} ↑
✓			49.43
✓	✓		53.46
✓	✓	✓	**55.91**

Table 2. We evaluate the impact of ablation on training loss.

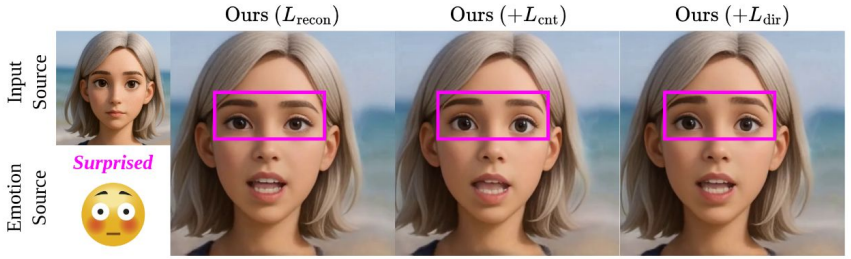
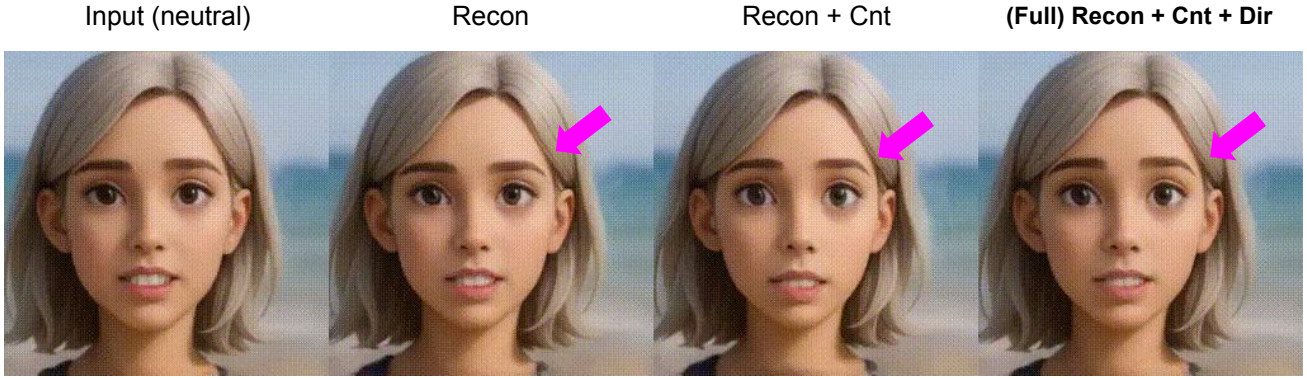


Figure 5. Qualitative analysis of ablation in the training loss.



Ablation study on disentanglement networks

C-MET can be used as **a plug-and-play module** into existing disentanglement-based talking face generation models with

1) **Enhancement of facial expressions,**
2) **Faster inference speed.**

Disentanglement Network	Metric	
	AITV ↓	Acc_{emo} ↑
PD-FGC [57]	1.247	33.36
w/ Ours	**1.180**	**36.82**
EDTalk [50]	2.827	41.99
w/ Ours	**2.643**	**55.91**

Table 3. Effect of integrating C-MET into disentanglement networks on MEAD.

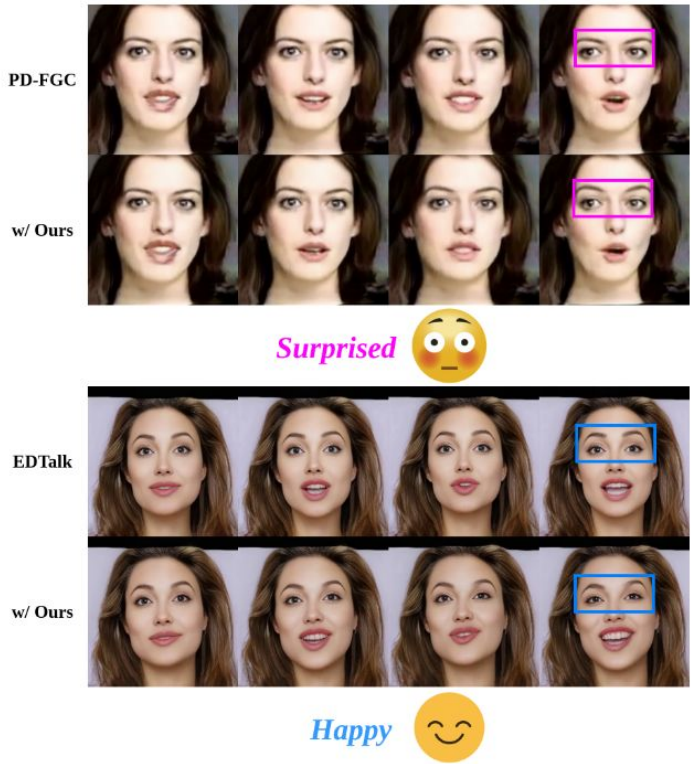


Figure 6. Qualitative analysis of integrating C-MET into disentanglement networks.

User Study

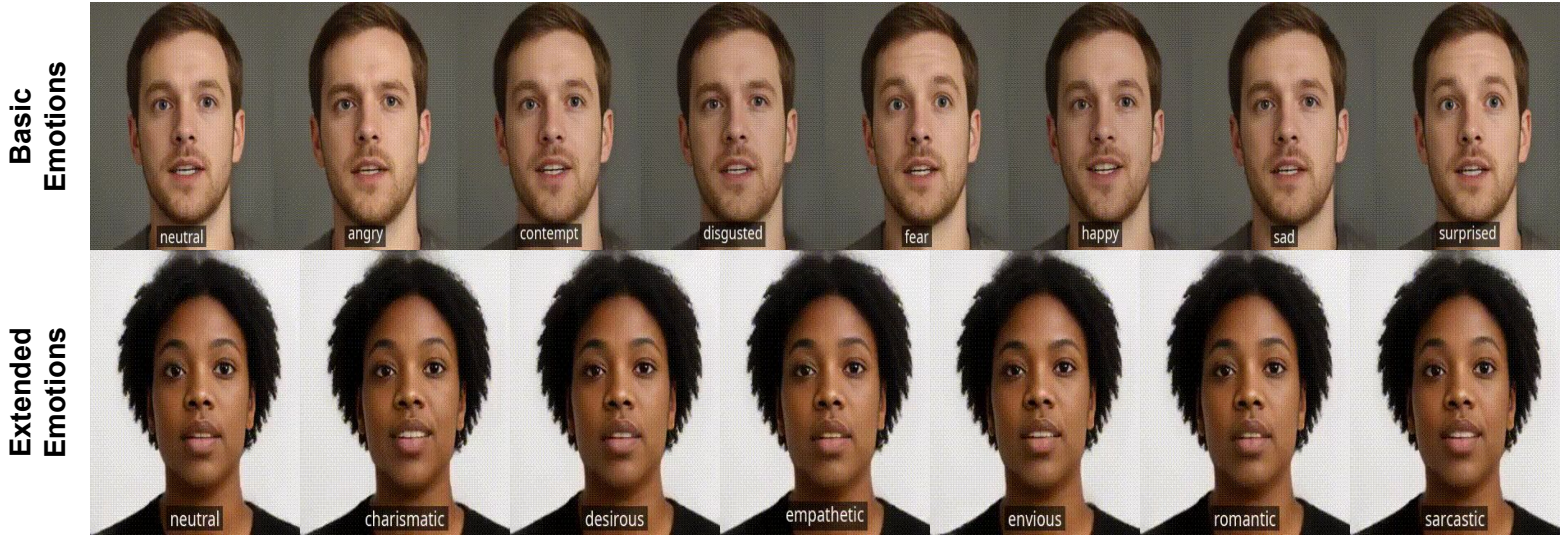
	Metric	Ours	EAMM	Tie	Ours	EAT	Tie	Ours	EDTalk	Tie	Ours	FLOAT	Tie
Basic Emotion	Emotional Expression (%)	**77.8**	10.4	11.8	**61.6**	21.4	17.1	**42.4**	22.0	35.7	**84.5**	14.9	0.6
	Visual Quality (%)	**77.8**	10.6	11.6	**61.4**	22.4	16.3	**40.6**	23.5	35.9	**81.4**	18.2	0.4
	Lip Synchronization (%)	**71.4**	14.5	14.1	**58.2**	24.7	17.1	**40.4**	23.3	36.3	**79.0**	20.4	0.6
Extended Emotion	Emotional Expression (%)	**91.0**	6.7	2.2	**80.4**	17.1	2.4	**51.2**	36.5	12.2	**86.9**	11.8	1.2
	Visual Quality (%)	**90.8**	8.8	0.4	**77.8**	19.4	2.9	**48.0**	39.8	12.2	**87.1**	11.4	1.4
	Lip Synchronization (%)	**87.6**	9.8	2.7	**78.4**	19.4	2.9	**45.3**	39.6	15.1	**86.7**	11.8	1.4

Table 4. **User study results across basic and extended emotions.** We report the percentage of participants who preferred our method, a baseline, or rated them equally (tie), in terms of emotional expression, visual quality, and lip synchronization. Our method consistently outperforms all baselines across both emotion categories.

Conclusion

Take-home messages:

- First enabling **extended facial expressions** beyond discrete labels (e.g., *sarcasm*)
- Audio-driven facial expression editing via **cross-modal semantic vectors**
- **+14% improvement in emotion accuracy** for emotion editing in talking face video
- **Lightweight plug-and-play module** for disentanglement based models



IMML Lab
Interactive **M**ultimodal
Machine **L**earning **Lab**

CVPR JUNE 3-7, 2026 DENVER COLORADO

THANK YOU

Project page: https://chanhyeok-choi.github.io/C-MET/

HuggingFace Demo: https://huggingface.co/spaces/coldhyuk/C-MET

Scan QR code for Paper and Code