EmoVid: A Multimodal Emotion Video Dataset for Emotion-Centric Video Understanding and Generation

Zongyang Qiu[1,2*], **Bingyuan Wang**[1*], **Xingbei Chen**[1], **Yingqing He**[3], **Zeyu Wang**[1,3†]

[1]The Hong Kong University of Science and Technology (Guangzhou), China
[2]Fudan University, China
[3]The Hong Kong University of Science and Technology, Hong Kong SAR, China
zyqiu22@m.fudan.edu.cn, bwang667@connect.hkust-gz.edu.cn, xchen053@connect.hkust-gz.edu.cn,
yhebm@connect.ust.hk, zeyuwang@ust.hk

Abstract

Emotion plays a pivotal role in video-based expression, but existing video generation systems predominantly focus on low-level visual metrics while neglecting affective dimensions. Although emotion analysis has made progress in the visual domain, the video community lacks dedicated resources to bridge emotion understanding with generative tasks, particularly for stylized and non-realistic contexts. To address this gap, we introduce **EmoVid**, the first multimodal, emotion-annotated video dataset specifically designed for artistic media, which includes cartoon animations, movie clips, and animated stickers. Each video is annotated with emotion labels, visual attributes (brightness, colorfulness, hue), and text captions. Through systematic analysis, we uncover spatial and temporal patterns linking visual features to emotional perceptions across diverse video forms. Building on these insights, we develop an emotion-conditioned video generation technique by fine-tuning the Wan2.1 model. The results show a significant improvement in both quantitative metrics and the visual quality of generated videos for text-to-video and image-to-video tasks. EmoVid establishes a new benchmark and protocol for affective video computing. Our work not only offers valuable insights into visual emotion analysis in artistic videos but also provides practical methods for enhancing emotional expression in video generation. The extended version and the dataset are available on the following links.

Project Page — https://zane-zyqiu.github.io/EmoVid
Extended version —
https://www.arxiv.org/abs/2511.11002

Introduction

Video is a powerful medium for storytelling and expression, with emotion playing a key role in viewer engagement (Cao et al. 2022). While recent video generation models have improved in visual coherence and motion, they have paid limited attention to emotional expressiveness (Kalateh et al. 2024). This is especially true in creative applications like

---

*These authors contributed equally.
†Corresponding author.

comic portrait animation, sticker (meme) creation, and cinematic editing, where emotional expressiveness is essential but underexplored (Wang, Chen, and Wang 2025).

In recent years, emotional content analysis and generation have achieved great progress in language, speech, and images, and have gained considerable attention within multimodal contexts (Kalateh et al. 2024). In video, multimodal approaches have shown promising results in improving tasks such as sentiment analysis, emotion-driven content creation, and interactive video generation (Pandeya and Bhattarai 2021). However, most of these studies focused on human dialogue and realistic styles. The integration of emotion in stylized video understanding and generation—taking into account both creative context, stylistic features, and the emotional undercurrents—remains underexplored.

In this paper, we introduce EmoVid, the first large-scale and emotion-labeled video dataset focusing on stylized and non-realistic content. As shown in Figure 1, EmoVid consists of videos in three categories: cartoon animation, movie clips, and animated stickers (GIFs). We adopt the Mikels' eight-emotion scheme (Mikels et al. 2005) (amusement, awe, contentment, excitement, anger, disgust, fear, sadness), a widely used discrete set originally curated for affective image studies. Each clip is annotated with emotion labels, color attributes, and captioned via a vision-language model (VLM). Through this dataset, we explored the emotional patterns in videos, such as emotion distributions, color-emotion correlations, temporal transitions of emotion, and semantic links. We believe that these insights can enhance tasks like comic animation and stylized video generation.

To demonstrate the effectiveness of EmoVid, we propose a benchmark for both T2V and I2V tasks, evaluating the visual quality and emotion accuracy of AI-generated videos. We also fine-tune the Wan2.1 (Wan et al. 2025) model on our data. Experiments show a significant gain in emotional expressiveness when emotion is explicitly incorporated as a prior into video generation tasks. We also designed a video generation pipeline, which can generate animated stickers of any character with any emotion. This is of great value for today's network communication and can also be further used in the production of animations and movies. Together, EmoVid contributes to both affective computing and stylized video generation by linking emotion understanding
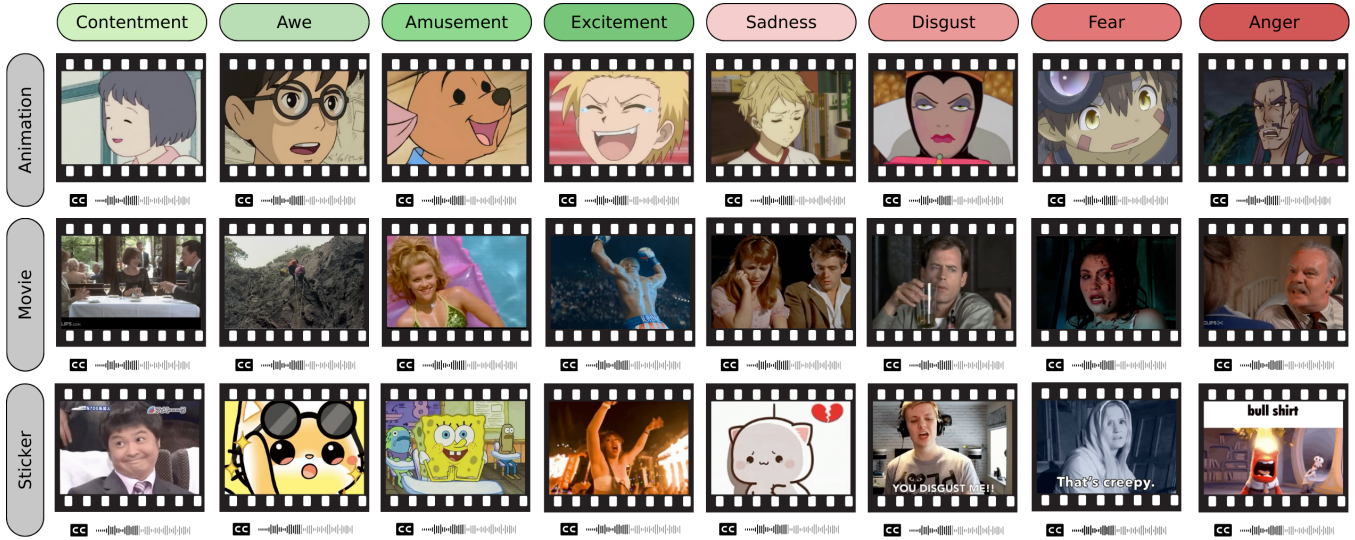
Figure 1: **Overview of the EmoVid dataset.** The dataset spans eight emotion categories—*Contentment, Awe, Amusement, Excitement, Sadness, Disgust, Fear, and Anger*—and three content domains: *Animation, Movie, and Sticker*. The dataset captures diverse emotional expressions in various visual styles and contexts, demonstrating both multimodal richness (with associated text and audio) and cross-domain generality.

with practical generative tasks.

In summary, we make the following contributions:

- We introduce EmoVid, a large-scale, emotion-labeled video dataset focusing on stylized and non-realistic content, and present a scalable benchmark, evaluation metrics, and protocol for emotional enhancement in video generation.

- We explore both spatial and temporal emotional patterns in the EmoVid dataset, as well as their relationship with text captions or other visual attributes.

- We demonstrate EmoVid's utility in generation and editing tasks by fine-tuning the Wan2.1 model, which shows significant improvement in emotional expression.

Related Work

Emotion Analysis and Affective Computing

Affective computing has gained increasing attention in recent years, particularly in textual, auditory, and visual modalities. Early research mostly focused on text sentiment analysis using lexical features and semantic understanding (Wang et al. 2025b). Subsequently, auditory emotion recognition emerged as a reliable signal through speech prosody, pitch, and tone analysis (Zadeh et al. 2016). Visual emotion recognition, especially via facial expression and gesture, has gained momentum more recently but remains less mature in comparison to text and audio (Zhu et al. 2024). More recently, multimodal affective computing has become a key focus (Das and Singh 2023), but comprehensive multimodal tasks remain scarce due to challenges in data alignment, modality imbalance, and dataset availability.

Within the visual modality, affective computing first progressed on static images, exploring the affective content of images via color histograms, facial attributes, and compositional cues (Pang, Zhu, and Ngo 2015). Nevertheless, static images lack temporal dynamics, which is essential for modeling transitions and nuances in affect. Consequently, video-based affective computing has gained traction, with datasets such as AffectNet and LIRIS-ACCEDE supporting sequence-based modeling (Mollahosseini, Hasani, and Mahoor 2017; Baveye et al. 2015). Despite progress, few efforts have been made to recognize video emotions within creative domains, such as film, performance, or storytelling, where affect is most deeply embedded and semantically rich.

Video Generation and Editing

Recent advances in video generation have shown remarkable capability across diverse domains, such as human motion synthesis (Tulyakov et al. 2018), natural scene rendering (Wang, Liu et al. 2018), and short video creation (Ho, Saharia et al. 2022). In creative applications—such as animation, film production, and meme/sticker creation—generative models like VideoCrafter (Yang, Xu et al. 2023) have demonstrated strong visual coherence and temporal smoothness, but current research primarily focuses on visual quality, realism, or aesthetic control, with little emphasis on affective expressiveness (Ma et al. 2025). Notably, meme or sticker generation inherently carries affective signals, yet emotional intent is typically implicit and lacks formal integration into generation frameworks.

While affect has been discussed in video synthesis through domains such as facial expression transfer (Zakharov et al. 2019), and gesture-guided animation (Cao, Yang et al. 2022), these discussions are often limited to human-centric or conversational tasks. Affective conditioning has been explored via latent space alignment (Ji, Wang et al. 2023) or emotion labels in prompt-based genera-

Dataset	Modalities	Size	Content	Emotion Labels
CAER (Lee et al. 2019)	v, a	12h (13k clips)	TV shows	7 cls
MELD (Poria et al. 2019)	v, a, t	1.4h (1.4k clips)	Human dialogue	7 cls
DEAP (Koelstra et al. 2011)	v, a	2h (120 clips)	Music videos	V-A-D
VEATIC (Ren et al. 2024)	v	3h (124 clips)	In-the-wild	V-A
MEAD (Wang et al. 2020)	v, a	40h (10k clips)	Human face	7 cls + 3 intensity levels
DH-FaceEmoVid-150 (Liu et al. 2025)	v, t	150h (18k clips)	Human face	6 cls + 4 compound
EmoVid (Ours)	**v, a, t**	**39h (22k clips)**	**Animation, Movie, Sticker**	**8 cls (Mikels)**

Table 1: **Comparison of EmoVid with other emotional video datasets.** We focus on modalities, size, content types, and emotion label schemes. Modalities are abbreviated as follows: *v = Video, a = Audio, t = Text*. Emotion labels include either discrete categories (e.g., 7 cls = 7 emotion classes) or dimensional annotations such as *Valence–Arousal (V–A)* and *Valence–Arousal–Dominance (V–A–D)*.

tion (Guo, Huang et al. 2023), yet these approaches rarely address stylized and non-realistic domains where emotion is central to narrative structure, such as animated films or cinematic scene generation (Wang et al. 2025a). The gap between affective modeling and creative video generation suggests a pressing need to bridge semantic emotion representation with generative visual storytelling.

Emotion-related Datasets

Affective computing has been supported by a growing number of emotion datasets across textual, auditory, and visual modalities. Text-based corpora such as SemEval (Rosenthal, Farra, and Nakov 2017) and GoEmotions (Demszky, Movshovitz-Attias et al. 2020) provide fine-grained emotion labels for sentiment and intent understanding. In the auditory domain, datasets like RAVDESS (Livingstone and Russo 2018) include speech with emotion expressions. Visual emotional datasets evolved from static-image datasets such as Emotion6 (Peng, Wang et al. 2015) and EmoSet (Yang et al. 2023). However, static images lack temporal continuity, limiting their utility in studying emotional dynamics and transitions. This has motivated the development of video-based datasets for affective modeling.

Several datasets have attempted to bridge the gap between affective labeling and video modality. MELD (Poria et al. 2019), DEAP (Koelstra et al. 2011), and VEATIC (Ren et al. 2024) offer emotion-annotated videos, but are limited in either modality (e.g., lacking audio and text), domain focus (e.g., only dialogues or music videos), or size. Facial datasets such as MEAD (Wang et al. 2020) and DH-FaceEmoVid-150 (Liu et al. 2025) focus on constrained emotional expressions, offering high precision for facial affect but limited diversity in content and setting. These datasets are suitable for emotion recognition but not ideal for video generation, where varied visuals, rich context, and narrative emotional arcs are critical. In contrast, EmoVid introduces a large-scale, multimodal (video, audio, text) dataset with 22,758 clips covering animation, film scenes, and stickers—domains where emotional content is not only embedded but essential to semantics.

The EmoVid Dataset

As discussed in the former sections, the primary challenge of emotion-enhanced video generation lies in the lack of stylized emotional video datasets. We analyzed previous datasets specifically in emotion and video-related fields, and summarized their features in terms of size, category, content, and emotion labels, as shown in Table 1. Through the analyses, we find that existing datasets either lack in scale or fail to include all necessary modalities, which hinders the progress of multimodal emotion analysis. Moreover, they all focus exclusively on real-world scenarios (primarily human facial expressions), limiting the effective transfer of emotional priors to general video generation tasks.

Data Collection

To fill this gap, we created the EmoVid dataset, the first large-scale multimodal video dataset with fine-grained emotion labels, which comprises 2,807 animation face clips, 13,255 movie clips, and 6,696 animated stickers. The dataset includes high-quality emotional annotations, as well as other relevant visual attributes such as brightness, colorfulness, and hue, along with a textual caption for each video clip. Basic information is provided in Table 2, and more detailed information can be found in the extended version.

For the animation clips, we source data from the MagicAnime dataset, which contains 3,000 clips of cartoon faces from American, Chinese, and Japanese cartoons (Xu et al. 2025). The movie clips are retrieved using the metadata and code provided by Condensed Movies (Bain et al. 2020). As movie videos are mostly several minutes long, we segment them using the PySceneDetect tool (Castellano 2025). Only clips within 4–30 seconds were retained for further analysis. For the animated stickers, we used the Tenor API (Tenor, Inc. 2025) to search for GIFs based on the eight primary

Type	Clips	Avg	SD	Vid	Aud	Cap
Total	**22758**	**6.18**	**4.53**	–	–	–
Animation	2807	5.12	2.65	✓	✓	✓
Movie	13255	8.75	4.71	✓	✓	✓
Sticker	6696	2.91	2.18	✓	✗	✓

Table 2: **Basic statistics of the EmoVid dataset.** *Avg* is the average duration (in seconds), and *SD* is the standard deviation. *Vid*, *Aud*, and *Cap* indicate whether each clip includes video, audio, and textual caption, respectively.

emotion labels and their synonyms summarized by Yang et al. (2023). Each clip is manually verified to ensure it accurately expresses the intended emotion. More details are included in the extended version.

Videos in EmoVid are annotated at the clip level, and the annotation includes the following aspects:

Emotion. We employ the widely-used Mikels emotion model, which categorizes emotions into eight types: *amusement, awe, contentment, excitement, anger, disgust, fear, and sadness*. As the Valence-Arousal Model (Russell 1980) also plays an important role in intuitive understanding of emotions, we sorted the valence and arousal of the eight emotions as in Figure 2 according to the work of Warriner, Kuperman, and Brysbaert (2013).

Given the trade-off between labeling accuracy and resource consumption, we adopt a human-machine collaborative method to obtain the labels. We first conducted a comparative experiment on the EmoSet dataset, the results of which are detailed in the extended version. We find that fine-tuning VLMs on the same data domain significantly improves emotion labeling accuracy, and NVILA-Lite-2B (Liu et al. 2024b) exhibits classification performance comparable to that of humans. We pick 20% of the animation and movie data to be annotated by human annotators, with each clip tagged as one of the eight emotions or as *no specific emotion*. As emotions are ambiguous and open to interpretation, each video is annotated by three people, and the video is retained only when at least two annotators provide the same result. For the remaining 80% of the data, we use the NVILA-Lite-2B model (fine-tuned on the manually labeled data) to annotate the clips. To assess annotation quality, we randomly select 1% of the videos as a validation set. Three human annotators independently annotate the same set. We then calculated pairwise Cohen's kappa scores across the four annotation sources (three humans and the VLM). The results indicate a small difference ($< 4\%$) between the inter-human kappas and the human-VLM kappas, indicating that the VLM provides labels of similar quality to humans.

Attributes and Captions. We computed three low-level visual attributes for each video clip: *colorfulness, brightness, and hue*, based on the HSV color space. Specifically, we sampled every 20 frames from each clip, and every pixel is represented by a triplet (h_i, s_i, v_i). For hue H, each pixel

value h_i is defined as an angle on the color wheel in the range $[0, 360)$, and the overall H is defined as the angle of the vector resulting from the sum of these vectors:

$$\phi = \text{atan2}\left(\sum_i \sin(h_i), \sum_i \cos(h_i)\right) \times \frac{180°}{\pi} \quad (1)$$

$$H = \text{round}\left((\phi + 360°) \mod 360°\right) \quad (2)$$

The modulo operation ensures that H lies within the range $[0, 360)$. Colorfulness (C) was defined as the normalized average of the saturation channel (S), and brightness (B) was calculated using the value channel (V):

$$C = \text{round}\left(\frac{1}{N}\sum_{i=1}^{N} S_i, 1\right), B = \text{round}\left(\frac{1}{N}\sum_{i=1}^{N} V_i, 1\right). \quad (3)$$

where N is the total number of pixels sampled from selected frames. Both C and B were normalized between 0 and 1 and rounded to one decimal place.

Additionally, we generated high-quality captions for each clip using the NVILA-8B-Video model to facilitate further training and evaluation based on the dataset.

To sum up, beyond categorical emotion labels, EmoVid provides rich multimodal annotations including:

- **Audio tracks** aligned with each video to enable audio-visual emotion fusion.
- **Low-level visual features** including *brightness*, *colorfulness*, and *hue*, quantitatively extracted to support emotion attribution analysis.
- **Free-form captions** generated by a VLM, describing the perceived content and sentiment of each video.

Analysis of EmoVid

Properties of EmoVid

EmoVid is a large-scale, multimodal video dataset designed for emotion-aware video understanding and generation tasks. The dataset consists of 22,758 videos, with a total duration of 140,580 seconds. Among them, 10,049 clips
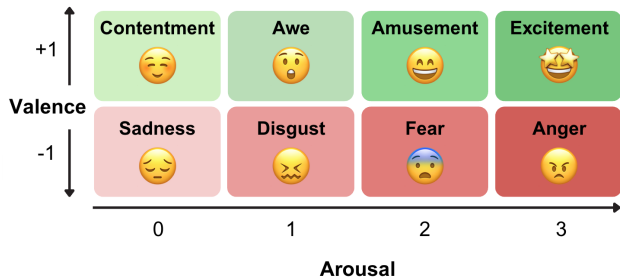

Figure 2: **Relationship between different emotions.** We refer to Warriner, Kuperman, and Brysbaert (2013) to arrange emotion categories on the valence-arousal model.
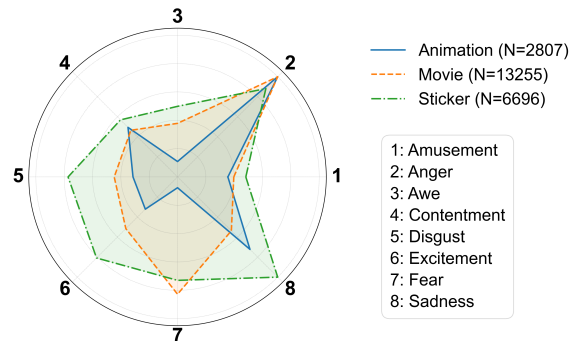

Figure 3: **Emotion distribution across three video categories.** The number of videos in the three video types was normalized to better illustrate the distribution of data across different emotions.
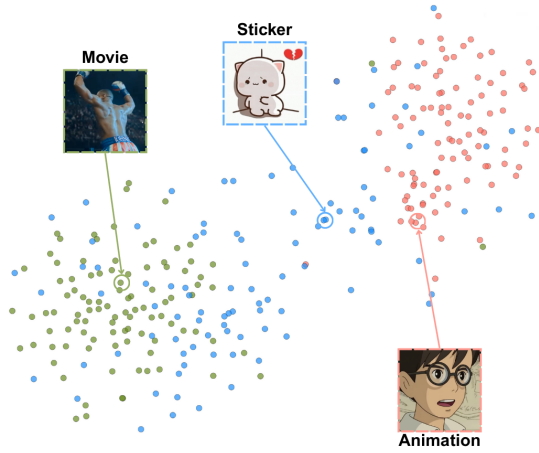
Figure 4: **t-SNE visualization of video features.** Animation and Movie clusters are separated, with Sticker samples overlapping both, reflecting their hybrid content characteristics.

have human-annotated emotion labels (282 animation clips, 2,771 movie clips, and 6,996 sticker clips). The emotion labels of the rest are generated by a fine-tuned VLM, the quality of which is verified in the above experiments.

As depicted in Figure 4, the t-SNE visualization illustrates the clear distribution among the three data types. We observe significant differentiations between Animation and Movie types, while Sticker data points are intermediate, exhibiting overlaps with both Animation and Movie clusters. This aligns intuitively with expectations regarding content similarity across these categories.

EmoVid covers eight discrete emotion categories aligned with the Mikels model, including *amusement, anger, awe, contentment, disgust, excitement, fear*, and *sadness*. The videos are sourced from three representative domains—*animation*, *movie*, and *sticker* content—each contributing different emotional intensities, stylistic traits, and temporal structures. Figure 3 summarizes the distribution of emotion labels across domains. From the figure, we observe that the emotion distribution in animation and movie video types is relatively imbalanced, with emotions like anger and sadness appearing more frequently, while amusement and awe are underrepresented. This reflects the natural distribution of emotions in real-life settings because these videos are all collected from real-world contexts and annotated.

Together, these properties make EmoVid an interpretable and extensible benchmark for video-based emotion understanding and generation. The dataset serves as a foundation for multimodal tasks such as emotional storytelling, text-to-video (T2V) or image-to-video (I2V) generation, and emotionally grounded video editing.

Emotional Structure and Dynamics

Visual attributes. The inclusion of visual attributes in the dataset is intended to facilitate further research on the relationship between emotion and color expression. In Figure 5, the positive-to-total emotion ratio demonstrates a gen-

eral upward trend concerning colorfulness and brightness attributes, which is consistent with our conventional knowledge. We also calculate the average value of colorfulness, brightness, and hue for each of the eight emotions. Positive-valence categories are brighter and slightly more colorful than negative-valence ones, while high-arousal emotions tend to be darker but more colorful than low-arousal emotions. ANOVA (St, Wold et al. 1989) confirms statistically significant ($p < 0.01$) but small effects ($\eta^2 < 1\%$), and the details are in the extended version.

Temporal analysis. Different from static data, a key advantage of video datasets lies in the additional temporal dimension, which enables the exploration of how emotions evolve over time. Since the movie clips are extracted from continuous segments of films, we are able to perform temporal analysis based on them. We analyse the first-order Markov transition matrix derived from consecutive movie clips, which can be found in the extended version, unveiling a three-stage emotional dynamic. First, all eight emotions exhibit strong self-persistence—particularly fear (0.53), anger (0.46), and amusement (0.46)—indicating that once an affective state is established, the visual stream tends to maintain it over short temporal windows. Second, transitions are markedly more frequent within the same valence polarity than across it (typically 0.08–0.18 versus < 0.08). Third, negative emotions reveal a chain-like escalation pattern: sadness →fear/anger and fear →anger—suggesting a possible "defense-attack" progression that makes negative sequences harder to dissolve (Blanchard et al. 1977). Together, these findings corroborate a "hold, intra-valence drift, arousal leap" trajectory, providing explicit understandings for emotion-aware video generation, editing, and pacing strategies.

Text Caption. To further explore the relationship between content and emotion, we extracted the five most frequently occurring 2–4 word phrases from video captions associated with each emotion, after filtering out redundant or noisy expressions. These phrases provide concrete semantic cues reflective of emotional content. For instance, under
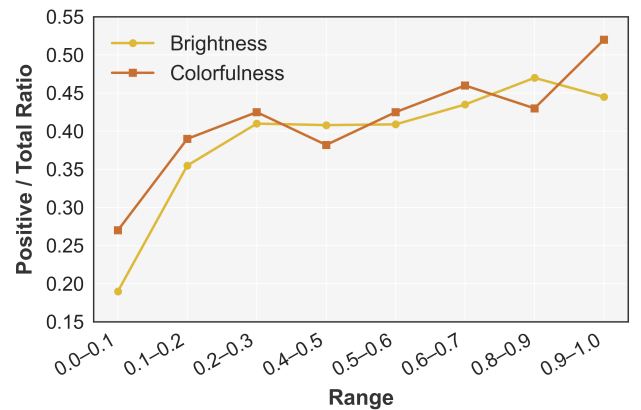

Figure 5: Positive-to-total emotion ratio across bins of colorfulness and brightness, exhibiting a distinct upward trend.

Task	Method	FVD↓	CLIP↑	SD↑	Flicker↓	EA-2cls↑	EA-8cls↑
T2V	VideoCrafter-V2	610.1	0.3012	–	0.0184	80.42	42.50
	HunyuanVideo	**552.6**	0.2776	–	**0.0116**	76.87	40.41
	CogVideoX	584.0	0.3013	–	0.0213	82.91	44.58
	WanVideo (before)	594.3	0.2982	–	0.0091	84.17	44.16
	WanVideo (after)	573.7	**0.3021**	–	0.0143	**88.33**	**48.33**
I2V	DynamiCrafter512	**512.3**	–	**0.7288**	0.0280	90.41	71.25
	HunyuanVideo	544.6	–	0.7244	**0.0233**	89.17	70.00
	CogVideoX	528.4	–	0.7214	0.0331	90.83	70.83
	WanVideo (before)	517.9	–	0.7146	0.0325	91.25	71.30
	WanVideo (after)	517.8	–	0.7193	0.0324	**94.58**	**76.25**

Table 3: **Quantitative results on the EmoVid benchmark.** Evaluation covers T2V and I2V tasks. Emotion Accuracy (EA) is assessed using binary (EA-2cls) and 8-category (EA-8cls) emotion classification metrics, highlighting the improved capability of finetuned models in capturing and reproducing emotional content.

amusement we found frequent phrases such as "funny reaction," "merry moment," and "laughing together." In contrast, *fear* often included phrases like "dark tunnel," "screaming sound," and "approaching danger." The detailed information can be found in the extended version.

Evaluation of EmoVid

To rigorously evaluate the effectiveness and utility of EmoVid, we construct a comprehensive benchmark and perform both quantitative and qualitative analyses. Specifically, we sample 240 videos across 3 video types with 8 distinct emotion labels, selecting 10 representative videos for each category. The test videos are all sampled from human-annotated ones and have been double-checked to guarantee the best quality. For each video, we use its corresponding caption and modify it by appending an explicit emotional label (e.g., *The video is in the "amusement" emotion*).

We test four SOTA T2V models—VideoCrafter-V2 (Chen et al. 2024), HunyuanVideo (Kong et al. 2024), CogVideoX-5B (Yang et al. 2024), and Wan2.1-T2V-14B (Wan et al. 2025). We also conduct experiments on four I2V models—DynamiCrafter512 (Xing et al. 2024), HunyuanVideo-I2V, CogVideoX-I2V, and Wan2.1-I2V-480P.

Furthermore, we fine-tune both T2V and I2V Wan2.1 models on EmoVid, excluding the data in the benchmark with the LoRA technique (Hu et al. 2022). We conducted fine-tuning using the DiffSynth Studio framework (ModelScope 2025) on an H20 GPU with 96 GB memory. To balance the distribution of training data, we did not use the entire set of movie clips. Instead, the final training dataset consisted of 2,727 animation clips, 8,000 movie clips, and 6,616 sticker clips. The LoRA configuration was set with rank=32, learning_rate=1e-4, training_epoch=3, and batch_size=1, the same as default settings in DiffSynth Studio.

Quantitative Results

To measure the emotional accuracy of generated videos, we use the fine-tuned VLM employed during dataset annotation to classify the generated videos and adopt the same metrics as Yang et al. (2023). **EA-2cls** measures binary accuracy by checking whether the predicted emotion matches the ground-truth valence, while **EA-8cls** measures top-1 accuracy across the eight discrete emotions. We evaluated model performance using a diverse set of metrics:

- **FVD** (Ge et al. 2024) measures the overall visual fidelity.
- **CLIP Score** (Hessel et al. 2021) quantifies semantic alignment between text and video.
- **SD Score** (Liu et al. 2024a) measures consistency between video and its first frame.
- **Temporal Flicker** (Huang et al. 2024) captures temporal instability across frames.
- **EA-2cls** and **EA-8cls** measure binary and full-class emotion accuracy.

Table 3 shows that our fine-tuned model *WanVideo (after)* has better performance in both tasks over the baseline *WanVideo (before)*. In the T2V setting, while general metrics like FVD and CLIP show comparative performance, the emotion alignment metrics (EA-2cls and EA-8cls) exhibit clear gains, indicating stronger emotional expression fidelity in the generated videos. Similar trends are observed in the I2V scenario, where the fine-tuned model surpasses all competitors, achieving the highest emotion classification accuracy of **92.08%** (2-class) and **72.92%** (8-class).

Qualitative Results

To further assess the effectiveness of fine-tuning pretrained models on our dataset, we qualitatively compare video outputs of the original Wan2.1-I2V-480P model and those of the fine-tuned version. As shown in Figure 6(a), the baseline model often fails to capture the correct emotional tone—e.g., generating neutral or mismatched expressions—whereas the fine-tuned model exhibits more precise emotional articulation, such as heightened facial expressions, contextual cues, and mood-consistent motion patterns. These improvements highlight the value of EmoVid not only as a benchmarking dataset, but also as a valuable resource for emotion-specific downstream tasks.

We also utilize the LoRA-trained I2V model to generate animated stickers conditioned on different characters and emotions, as illustrated in Figure 6(b). The results demonstrate that our model is capable of producing vivid emotional
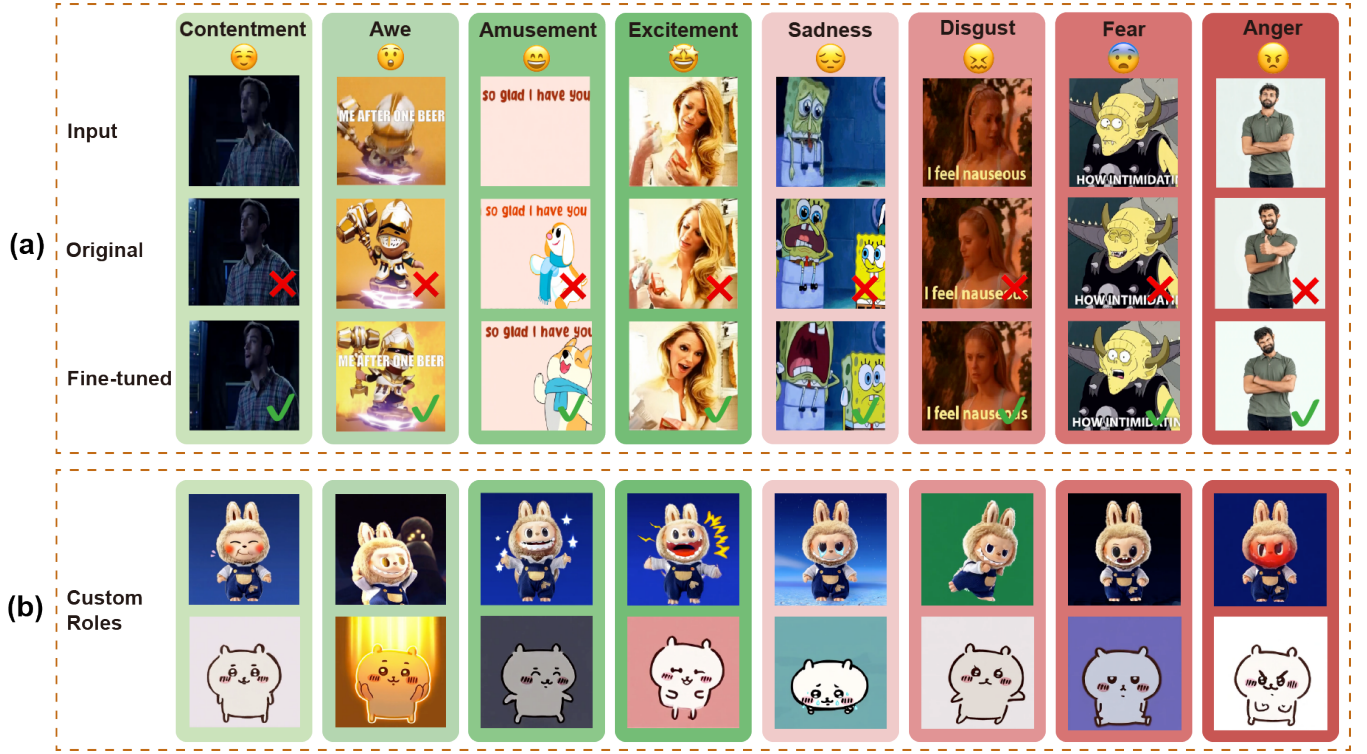
Figure 6: **Qualitative results.** (a) Comparison between the original Wan2.1 I2V model and our fine-tuned one. The ✓indicates better emotional alignment. (b) Emotion-conditioned animated sticker generation using the fine-tuned Wan2.1 I2V model.

expressions, which can be applied to social media platforms. In addition, we employ a multi-LoRA approach on the T2V model, combining our emotion-aware LoRA with other Lo-RAs that encode character identity or visual style priors, to generate videos with specific emotional attributes. More results can be found in the extended version.

Discussion

Through comparative experiments, we validate the effectiveness of the EmoVid benchmark. The quantitative results demonstrate that models fine-tuned with EmoVid data exhibit superior emotional accuracy, and the qualitative comparisons further support this conclusion. Such improvement mainly comes from cases where the baseline model failed to express the intended emotion, while the fine-tuned version captured it more precisely. Importantly, EmoVid is designed for artistic and creative scenarios where emotional expression is central. Such contexts include films and operas, miniseries, and emotionally evocative social media content such as expressive memes. In these applications, emotional clarity is often more important than realism or temporal fidelity. Our benchmark captures this focus by emphasizing emotional accuracy as the primary evaluation criterion.

Finally, we anticipate that EmoVid will be useful far beyond the scope of model benchmarking. Potential downstream applications include emotion-aware avatar generation, expressive media content synthesis, and controllable video editing based on emotional cues. As generative mod-

els become increasingly capable, datasets like EmoVid can help ground their outputs in meaningful human affect.

Conclusion

This paper has introduced **EmoVid**, the first large-scale emotion-annotated video dataset tailored specifically for creative contexts, including animations, movie clips, and animated stickers. EmoVid fills a critical gap by providing high-quality multimodal annotations—emotion labels, visual attributes, and textual captions—enabling deeper analysis of the interplay between visual features, temporal dynamics, and emotional perception. Through extensive experiments, we demonstrated EmoVid's capability in fine-tuning state-of-the-art generative models, resulting in significant improvements in emotional expressiveness for both T2V and I2V tasks.

By establishing a new benchmark for affective video computing, EmoVid not only advances fundamental research in emotion-driven video understanding and generation but also supports practical applications in fields such as animation, filmmaking, and social media communication. Our work is based on the assumption that each clip conveys a specific emotion. However, due to the complex nature of human emotion, the real-world expressions can be highly detailed and composite. In addition, the audio component of the dataset can be better leveraged to build a truly unified video-audio-text multimodal model. We will continue to explore these directions in our future work.

Ethics Statement

In recognition of the importance of copyright and privacy protection, we will only provide access to our dataset strictly for non-commercial research use by academic institutions. Any use of the dataset must comply with relevant intellectual property laws and ethical research standards. Redistribution or commercial use is prohibited.

Acknowledgments

References

Bain, M.; Nagrani, A.; Brown, A.; and Zisserman, A. 2020. Condensed Movies: Story-Based Retrieval with Contextual Embeddings. In *Proceedings of the Asian Conference on Computer Vision*.

Baveye, Y.; Dellandrea, E.; Chamaret, C.; and Chen, L. 2015. LIRIS-ACCEDE: A Video Database for Affective Content Analysis. *IEEE Transactions on Affective Computing*, 6(1): 43–55.

Blanchard, R. J.; Blanchard, D. C.; Takahashi, T.; and Kelley, M. J. 1977. Attack and Defensive Behaviour in the Albino Rat. *Animal Behaviour*, 25: 622–634.

Cao, W.; Yang, W.; et al. 2022. GestureDiffuCLIP: Gesture-Conditioned Diffusion Model for Zero-Shot Emotive Body Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cao, W.; Zhang, K.; Wu, H.; Xu, T.; Chen, E.; Lv, G.; and He, M. 2022. Video Emotion Analysis Enhanced by Recognizing Emotion in Video Comments. *International Journal of Data Science and Analytics*, 14(2): 175–189.

Castellano, B. 2025. PySceneDetect.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. arXiv:2401.09047.

Das, R.; and Singh, T. D. 2023. Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges. *ACM Computing Surveys*, 55(13s): 1–38.

Demszky, D.; Movshovitz-Attias, D.; et al. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. *arXiv preprint arXiv:2005.00547*.

Ge, S.; Mahapatra, A.; Parmar, G.; Zhu, J.-Y.; and Huang, J.-B. 2024. On the Content Bias in Fréchet Video Distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guo, R.; Huang, X.; et al. 2023. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *arXiv preprint arXiv:2307.04725*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-Free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*.

Ho, J.; Saharia, C.; et al. 2022. Imagen Video: High-Definition Video Generation with Diffusion Models. *arXiv preprint arXiv:2210.02303*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*, 1(2): 3.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Ji, Z.; Wang, W.; et al. 2023. EmoVideo: Emotion-Controllable Text-to-Video Generation via Affective Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kalateh, S.; Estrada-Jimenez, L. A.; Hojjati, S. N.; and Barata, J. 2024. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access*.

Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2011. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1): 18–31.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuan-Video: A Systematic Framework for Large Video Generative Models. *arXiv preprint arXiv:2412.03603*.

Lee, J.; Kim, S.; Kim, S.; Park, J.; and Sohn, K. 2019. Context-Aware Emotion Recognition Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10143–10152.

Liu, H.; Sun, W.; Di, D.; Sun, S.; Yang, J.; Zou, C.; and Bao, H. 2025. MoEE: Mixture of Emotion Experts for Audio-Driven Portrait Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26222–26231.

Liu, Y.; Cun, X.; Liu, X.; Wang, X.; Zhang, Y.; Chen, H.; Liu, Y.; Zeng, T.; Chan, R.; and Shan, Y. 2024a. Eval-Crafter: Benchmarking and Evaluating Large Video Generation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22139–22149.

Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; Li, X.; Fang, Y.; Chen, Y.; Hsieh, C.-Y.; Huang, D.-A.; Cheng, A.-C.; Nath, V.; Hu, J.; Liu, S.; Krishna, R.; Xu, D.; Wang, X.; Molchanov, P.; Kautz, J.; Yin, H.; Han, S.; and Lu, Y. 2024b. NVILA: Efficient Frontier Visual Language Models. arXiv:2412.04468.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLOS ONE*.

Ma, Y.; Feng, K.; Hu, Z.; Wang, X.; Wang, Y.; Zheng, M.; He, X.; Zhu, C.; Liu, H.; He, Y.; et al. 2025. Controllable Video Generation: A Survey. *arXiv preprint arXiv:2507.16869*.

Mikels, J. A.; Fredrickson, B. L.; Larkin, G. R.; Lindberg, C. M.; Maglio, S. J.; and Reuter-Lorenz, P. A. 2005. Emotional Category Data on Images from the International Affective Picture System. *Behavior Research Methods*, 37(4): 626–630.

ModelScope. 2025. DiffSynth-Studio. https://github.com/modelscope/DiffSynth-Studio/. Apache-2.0 license.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1): 18–31.

Pandeya, Y. R.; and Bhattarai, B. 2021. Deep-Learning-Based Multimodal Emotion Classification for Music Videos. *Sensors*.

Pang, L.; Zhu, S.; and Ngo, C.-W. 2015. Deep Multimodal Learning for Affective Analysis and Retrieval. *IEEE Transactions on Multimedia*, 17(11): 2008–2020.

Peng, K.; Wang, J.; et al. 2015. Mixed Emotional Image Dataset with Subjective and Objective Labels. In *Proceedings of the ACM International Conference on Multimodal Interaction*.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536.

Ren, Z.; Ortega, J.; Wang, Y.; Chen, Z.; Guo, Y.; Yu, S. X.; and Whitney, D. 2024. VEATIC: Video-Based Emotion and Affect Tracking in Context Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4467–4477.

Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of SemEval*.

Russell, J. A. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6): 1161.

St, L.; Wold, S.; et al. 1989. Analysis of Variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, 6(4): 259–272.

Tenor, Inc. 2025. Tenor – Animated GIF Search Engine.

Tulyakov, S.; Liu, M.-Y.; Yang, X.; and Kautz, J. 2018. MoCoGAN: Decomposing motion and content for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Wang, B.; Chen, Q.; and Wang, Z. 2025. Diffusion-Based Visual Art Creation: A Survey and New Perspectives. *ACM Computing Surveys*, 57(10): 1–37.

Wang, B.; Meng, H.; Cao, R.; Cai, Z.; Li, L.; Ma, Y.; Chen, Q.; and Wang, Z. 2025a. MagicScroll: Enhancing Immersive Storytelling with Controllable Scroll Image Generation. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 431–441. IEEE.

Wang, B.; Shi, Q.; Wang, X.; Zhou, Y.; Zeng, W.; and Wang, Z. 2025b. EmotionLens: Interactive Visual Exploration of the Circumplex Emotion Space in Literary Works via Affective Word Clouds. *Visual Informatics*, 9(1): 84–98.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation. In *Proceedings of the European Conference on Computer Vision*, 700–717. Springer.

Wang, T.-C.; Liu, M.-Y.; et al. 2018. Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems*.

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods*, 45(4): 1191–1207.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; and Wong, T.-T. 2024. Dynami-Crafter: Animating Open-Domain Images with Video Diffusion Priors. In *Proceedings of the European Conference on Computer Vision*, 399–417. Springer.

Xu, S.; Wang, B.; Cai, Z.; Fu, F.; Ma, Y.; Lee, T.; Yu, H.; and Wang, Z. 2025. MagicAnime: A Hierarchically Annotated, Multimodal and Multitasking Dataset with Benchmarks for Cartoon Animation Generation. arXiv:2507.20368.

Yang, J.; Huang, Q.; Ding, T.; Lischinski, D.; Cohen-Or, D.; and Huang, H. 2023. EmoSet: A Large-Scale Visual Emotion Dataset with Rich Attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20383–20394.

Yang, Y.; Xu, Y.; et al. 2023. VideoCrafter: Open Diffusion Models for High-Quality Video Generation. *arXiv preprint arXiv:2310.19512*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv preprint arXiv:1606.06259*.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhu, X.; Guo, C.; Feng, H.; Huang, Y.; Feng, Y.; Wang, X.; and Wang, R. 2024. A Review of Key Technologies for Emotion Analysis Using Multimodal Information. *Cognitive Computation*, 16(4): 1504–1530.