Noname manuscript No.
(will be inserted by the editor)

L-VOCAL: Language-based Video Colorization with Audio Alignment

Zheng Chang†, Shuchen Weng†, Huan Ouyang†, Yuchen Hong, Lihan Lin, Si Li ✉, Boxin Shi

Received: date / Accepted: date

Abstract While language-based video colorization addresses the inherent ambiguity of color assignment, language descriptions typically focus on central objects, neglecting the crucial context of emotional tone and surrounding environment necessary for accurate film colorization. In this paper, we introduce L-VOCAL, a novel framework for language-based video colorization that leverages audio alignment to supplement context not explicitly provided by language. L-VOCAL pretrains an alignment model to establish correspondences between color and audio, enabling the learning of emotional tone and environmental atmosphere. Subsequently, these aligned audio features guide the colorization process through specially designed condition injection modules. We additionally contribute L-VACOLOR, a new dataset tailored for this task, consisting of cinematic clips with diverse color and audio tones for training and evaluation. Extensive experimental results demonstrate that L-VOCAL produces colorization results that more accurately reflect filmmakers' artistic expression.

† Equal contribution.

✉ Si Li (Corresponding author)
E-mail: lisi@bupt.edu.cn

Huan Ouyang · Zheng Chang · Lihan Lin · Si Li
School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China.

Shuchen Weng
Beijing Academy of Artificial Intelligence, China.

Yuchen Hong · Boxin Shi
State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, China.

Keywords Video Colorization · Audio-Driven Editing · Diffusion Model · Cross-Modal Learning

1 Introduction

Old films serve as a vital medium for cultural heritage. However, early black-and-white films are typically monochromatic, which inherently limits their ability to convey emotional nuances and capture the full visual spectrum of real scenes, potentially diminishing the viewing experience for contemporary audiences. To address this issue, researchers explore ways to breathe new life into old films through colorization techniques (Welsh et al, 2002; Liu et al, 2008; Cheng et al, 2015).

Automatic video colorization methods (Liu et al, 2024b; Zhao et al, 2023; Lei and Chen, 2019) rely on semantic cues in the luminance channel to predict corresponding colors. Consequently, these methods suffer from inherent ambiguity in mapping grayscale to color, resulting in multiple semantically plausible colors for a single object. Language-based colorization methods (Li et al, 2024; Bozic et al, 2024; Chang et al, 2024, 2023) use language descriptions to guide the color assignment, further improving controllability and flexibility. However, these methods focus on central objects (*e.g.*, underlined words in Fig. 1), neglecting other relevant elements for accurate film colorization (*e.g.*, the emotional tone and surrounding environment).

As a multi-modal medium, film integrates rich emotional and environmental information through its audio component. Filmmakers often leverage the interplay of audio and visuals to connect with audiences, resulting in an intrinsic correspondence between audio and color. This audio-visual interplay is evident in two aspects: *(i)* **Emotional tone.** Audio amplifies the emotional impact and is frequently used to establish color-emotion relationships to evoke specific emotional responses

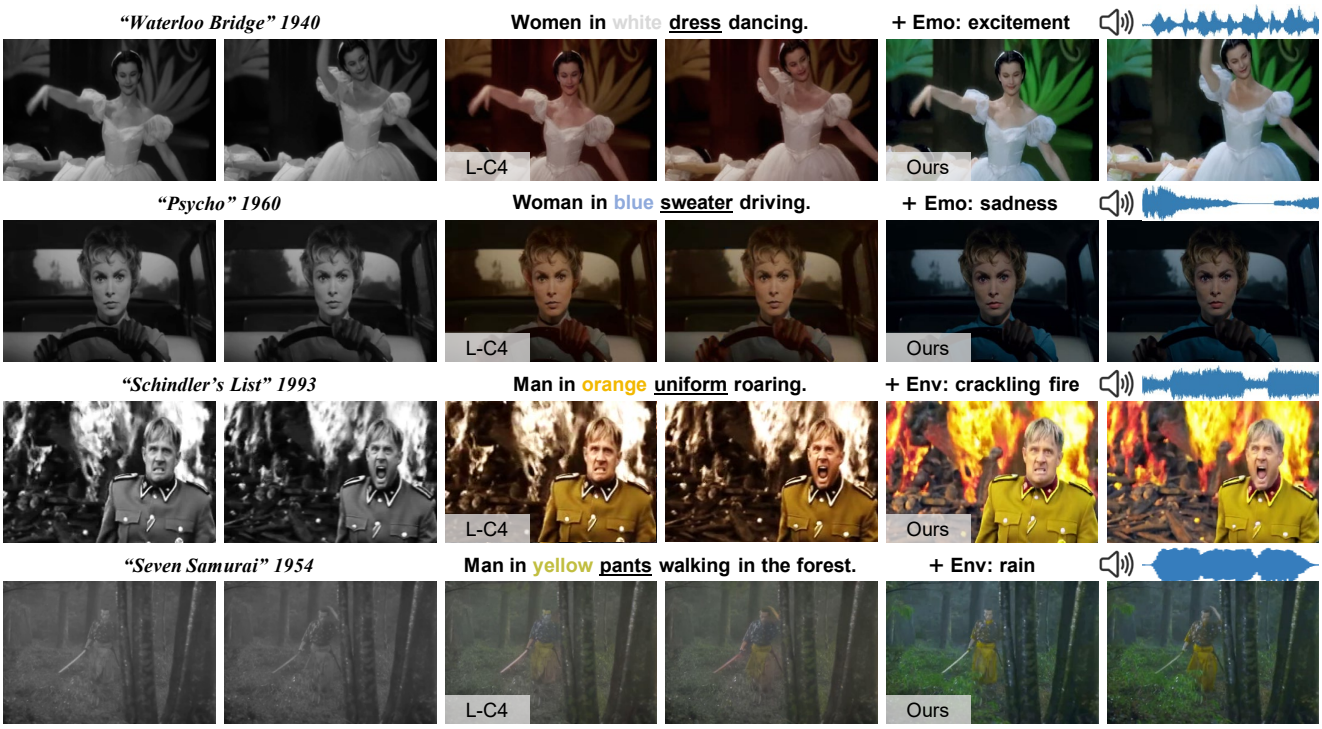

Fig. 1 Colorization results for black-and-white sound films, compared to the language-based method (Chang et al, 2024). By introducing film audio to supplement context not explicitly provided by language descriptions, L-VOCAL effectively produces colorization results with an appropriate emotional tone (first and second rows) and an enriched environmental atmosphere (third and fourth rows).

(Palmer et al, 2013; Barbiere et al, 2007) (*e.g.*, exciting music paired with vibrant colors and sad music with dull colors). *(ii)* **Environmental atmosphere.** Audio creates an immersive atmosphere, providing crucial environmental cues to enrich the viewer's scene understanding (*e.g.*, the sounds of crackling fire and rain). These observations motivate us to incorporate the audio semantics into the language-based video colorization process.

In this paper, we propose **L-VOCAL**, a novel **L**anguage-based **V**ide**O** **C**olorization with **A**udio a**L**ignment for the first time. To guide the colorization of central objects based on language descriptions, we build the model upon the pre-trained text-to-video generation model (WanTeam, 2025), leveraging its generative priors to colorize monochrome videos. The grayscale features are extracted to preserve the structure of original videos. To equip the pre-trained model with the audio semantics, we additionally develop an audio-color alignment model, which effectively enriches audio features with corresponding emotional and environmental context. During colorization, emotionally-aware audio features are used to modulate video features, guiding the colorization towards the appropriate emotional tone (*e.g.*, the bright tone in Fig. 1 first row and the dull tone in Fig. 1 second row). The environmentally-aware audio features are injected into the denoising network via cross-attention, improving the environmental atmosphere (*e.g.*, the crackling fire in Fig. 1 third row and the heavy rain in Fig. 1 fourth row).

For robust training and comprehensive evaluation of our L-VOCAL, we construct the **L-VACOLOR** dataset, consisting of film clips with **L**anguage descriptions and aligned **V**isual and **A**uditory samples for **COLOR**ization. A tailored filtering process selects film clips to ensure the accompanying audios provide clear emotional tone or environmental atmosphere. We further divide the dataset into 50K training and 50 testing samples, used for both audio-color alignment and colorization models. Each clip is additionally annotated with captions to describe the colors of central objects, enabling flexible and user-friendly interaction via language descriptions.

Our contributions can be summarized as follows:

- We establish the correspondence between color and audio, leveraging the emotional tone and environmental atmosphere to suggest film colors implicitly.
- We incorporate emotionally-aware and environmentally-aware audio features, producing colorization results that better reflect the filmmakers' artistic expression.
- We introduce the L-VACOLOR dataset featuring diverse film colors and audio tones, enabling the training and evaluation of audio-color alignment and colorization models.

2 Related work

2.1 Video colorization

The target of video colorization is to accurately assign semantically appropriate and visually appealing colors to monochrome videos while ensuring temporal consistency between frames. Although various strategies are proposed to improve colorization results (*e.g.*, self-regularization (Lei and Chen, 2019), adversarial learning (Zhao et al, 2023), and optical flow estimation (Liu et al, 2024b)), automatic video colorization methods predict colors based solely on luminance, which makes the colorization task inherently ill-posed. Exemplar-based video colorization methods (Wan et al, 2022; Zhang et al, 2019; Iizuka and Simo-Serra, 2019; Yang et al, 2024b) utilize reference images or user-selected frames to create implicit correspondences with monochrome frames. While they can produce plausible colorization results, their effectiveness is heavily dependent on the relevance and quality of the chosen exemplars, which restricts their application scenarios. Language-based video colorization methods (Chang et al, 2024; Liu et al, 2023) allow for flexible control over object colors. However, these methods tend to focus on the central objects as users tend to only describe objects they are interested in. This results in neglecting the color tone and environment atmosphere, and motivates us to introduce audio cues to assist in producing accurate colorization results.

2.2 Multimodal feature alignment

Multimodal alignment models are designed to facilitate more effective feature fusion by projecting features from different modalities into a common space. CLIP (Radford et al, 2021) achieves text-image alignment through contrastive learning on large-scale text-image pairs. Inspired by their remarkable success, CLAP (Elizalde et al, 2023) aligns text and audio features, demonstrating strong performance across various audio classification tasks. ImageBind (Girdhar et al, 2023) learns a unified embedding across six modalities (*i.e.*, image, text, audio, depth, thermal, and IMU data), extending the zero-shot capabilities to these multiple modalities and enabling novel cross-modal applications (*e.g.*, audio-to-text and audio-to-image generation). SCAV (Tsiamas et al, 2024) leverages sequential contrastive audio-visual learning and achieves significant performance improvements in retrieval tasks. To align emotions between modalities, Emo-CLIM (Stewart et al, 2024) introduces an emotional embedding space for images and music through emotion-supervised contrastive learning. The success of these methods in aligning diverse modalities motivates us to investigate a nuanced alignment between audio cues and visual color, to better model the rich audio-visual interplay presented in films.

2.3 Multimodal-guided visual editing

Although recent language-guided visual editing methods (Brooks et al, 2023; Feng et al, 2024; Liu et al, 2024a) have demonstrated significant advancements, they struggle to capture the rich semantic nuances required for complex editing goals (*e.g.*, distinguishing between heavy rain and thunderstorms). As an alternative, researchers explore approaches for incorporating audio semantics into image editing. Lee *et al.*(Lee et al, 2022b) project audio into a multimodal embedding space and optimize image generation by aligning audio representations. SonicDiffusion (Biner et al, 2024) introduces the audio-image cross-attention layer, enabling audio-guided image generation and editing. Recently, audio has been further applied to capture continuous and dynamic features in video editing. TPOS (Jeong et al, 2023) combines temporal semantics and amplitude features of audio to guide the audio-reactive video content generation. Soundini (Lee et al, 2023b) incorporates sound-guided visual effects (*e.g.*, explosion and lightning) into specific regions with a zero-shot approach. AudioScenic (Shen et al, 2024) preserves foreground content while modifying backgrounds through a temporally-aware audio semantic injection process. Inspired by these achievements, we leverage audio to supplement the context not explicitly provided by language descriptions for the video colorization.

3 L-VACOLOR dataset

The datasets used by previous video colorization methods typically lack accompanying audio, as these methods generally do not incorporate audio cues. Although audio-driven video editing models collect a large-scale synchronized video and audio recordings (Chen et al, 2020), they primarily consist of unedited real-world clips, still lacking the intentional artistic expression rendering in films. To further explore the correspondence between audio and color in films, we construct the L-VACOLOR dataset, consisting of film clips with aligned visual and auditory elements. The construction process of our dataset is illustrated in Fig. 2.

Data source. The L-VACOLOR dataset is derived from two public movie datasets: the Condensed Movies Dataset (Bain et al, 2020) and MovieBench (Wu et al, 2025). The Condensed Movies Dataset provides a rich collection of representative clips extracted from 3,605 films with a duration of 1,270 hours, capturing key scenes with synchronized audio and visual content. MovieBench delivers 69 hours of high-quality clips from 91 movies, further offering diverse scenes. By integrating these sources, L-VACOLOR combines abundant movie clips with audio tracks.

Single-shot segmentation. Previous foundational models for video colorization (Chang et al, 2024; Bozic et al, 2024; Li et al, 2024) and video generation (WanTeam, 2025; Chen

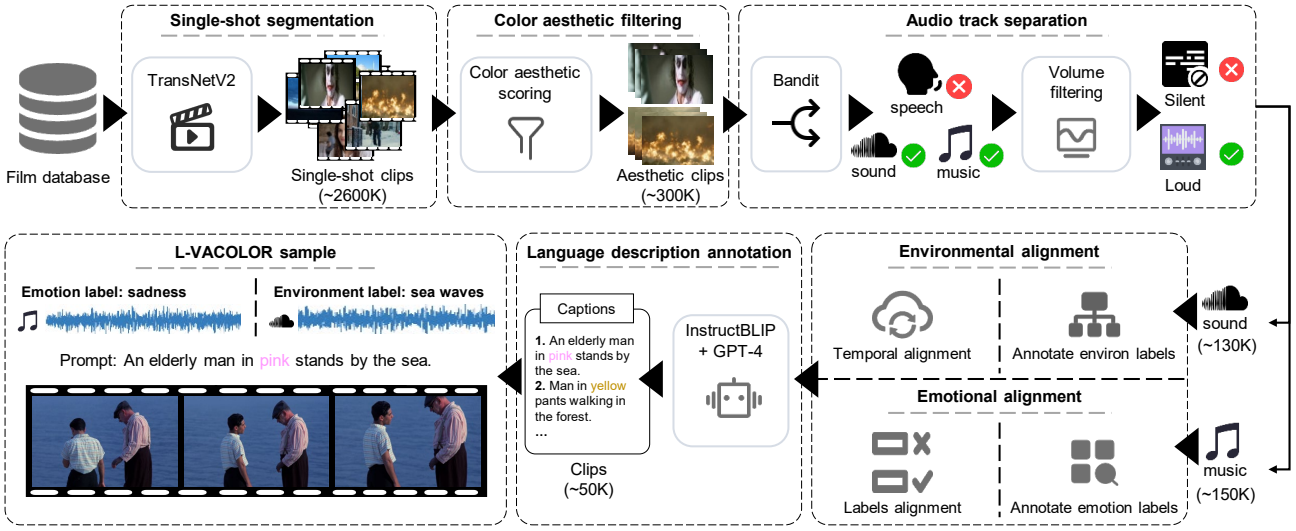

Fig. 2 The processing pipeline for our L-VACOLOR dataset.

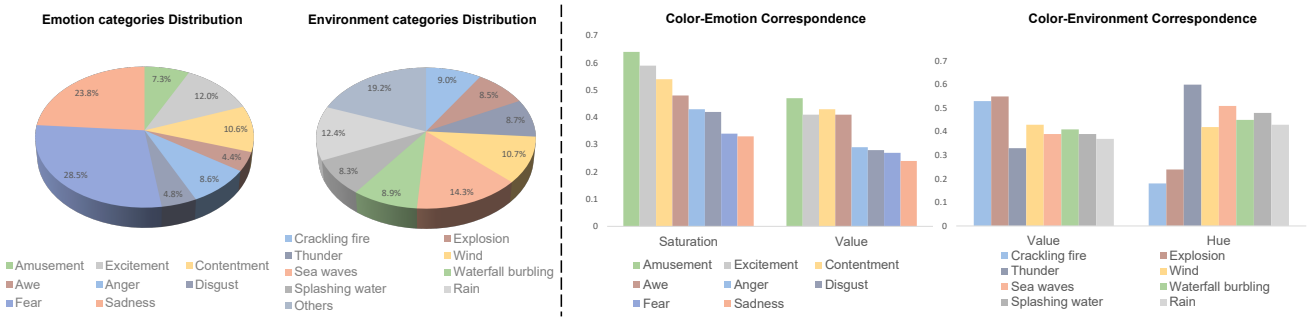

Fig. 3 Statistical analysis of the L-VACOLOR dataset. **Left:** The distribution of emotion and environment categories. **Right:** Illustration of correspondence between audio cues and visual color in collected film clips.

et al, 2024; Blattmann et al, 2023) treat single shots as their fundamental processing unit. Following this practice, we segment movie data into single-shot clips to represent a continuous environmental atmosphere and consistent emotional representation. The pre-trained TransNet (Soucek and Lokoc, 2024) is used to divide videos into single-shot clips with durations ranging from 3 to 15 seconds, with an average duration of approximately 10 seconds.

Color aesthetic filtering. We filter the dataset using the color aesthetic scoring (He et al, 2023), based on dominant colors, color harmony, and color combination. Higher color aesthetic scores indicate more deliberate color design during post-processing, suggesting a stronger artistic expression from filmmakers. We select clips with scores above 6.5, determined empirically.

Audio track separation. Audio tracks of film clips can be formulated as a combination of human speech, background music, and ambient sound effects. We observe that human speech generally lacks a direct correlation with specific color choices. Therefore, we separate the audio tracks of movie clips and discard the human speech track using Bandit (Watcharasupat et al, 2024), which requires approximately 2.1 GB of GPU memory and takes only about 1.16 seconds per sample. Background music is preserved to facilitate emotional tone, while ambient sound effects are preserved to reflect the environmental atmosphere. To remove silent portions, we conduct volume filtering to exclude clips with background music and ambient sound effects tracks falling below -30 dB.

Emotional alignment. To ensure emotional synchronization between film clips and background music, we assign both of them emotion categories based on Mikel's Wheel (Mikels et al, 2005), and then filter out unmatched samples. Specifically, the emotion for each segmented single shot is determined by annotating two key components: its middle frame using the image classifier (He et al, 2016) from EmoGen (Yang et al, 2024a), and its background music track using Qwen2-Audio (Chu et al, 2024). Only data with consistent emotion categories across these modalities are preserved for subsequent processing.

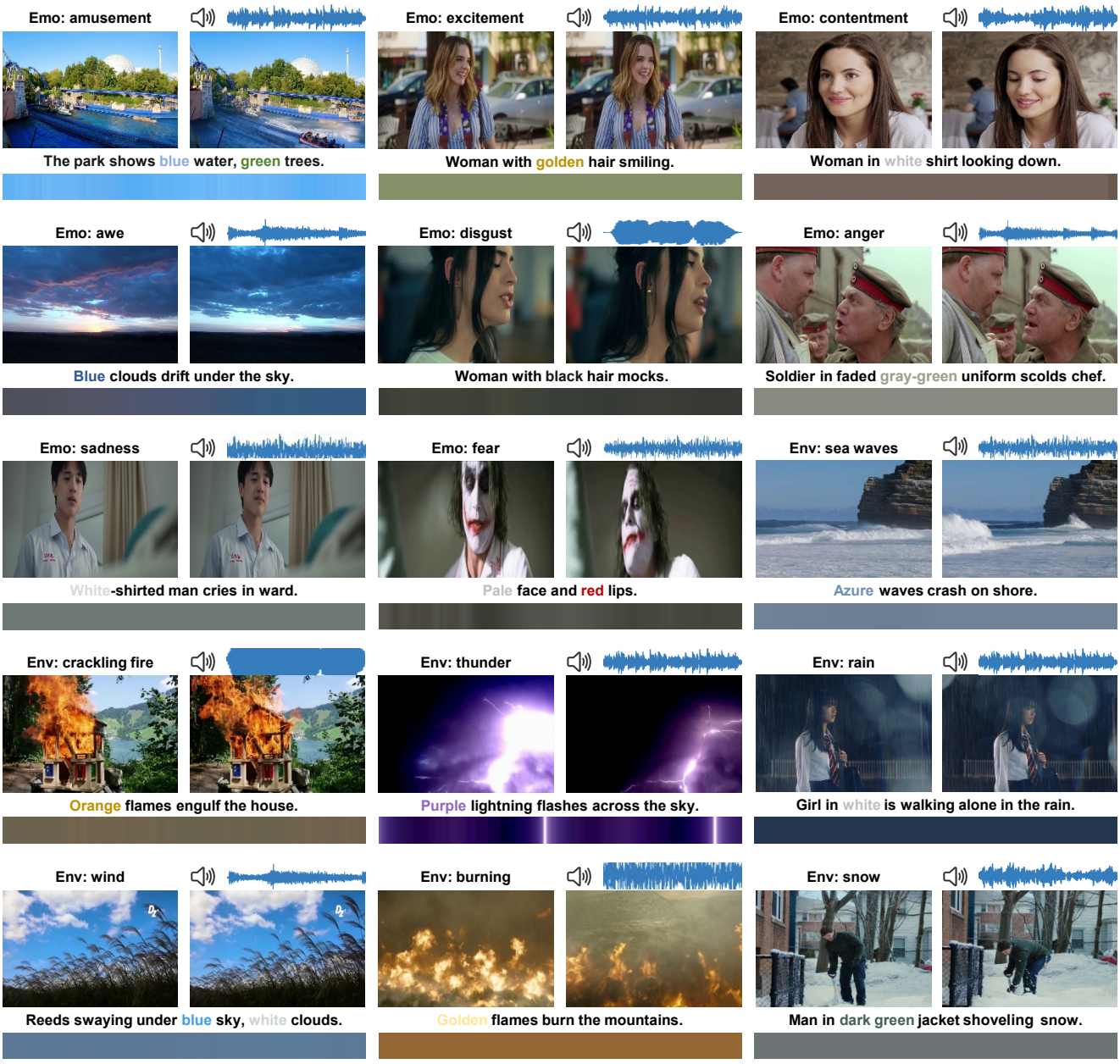

Fig. 4 Samples from the L-VACOLOR dataset, demonstrating that our collected film clips effectively align the visual and auditory elements.

Environmental alignment. To align the film clips with ambient sound effects, we discard data with unclear environmental semantics or out-of-synchronization issues. Specifically, we use labels from the ESC-50 dataset (Piczak, 2015) and the Landscape dataset (Lee et al, 2022a) (*e.g.*, "rain" and "sea waves") to construct language descriptions representing ambient sound effects, including a total of 59 distinct categories that cover a wide range of cinematic environments. Then, we utilize CLAP (Elizalde et al, 2023) to calculate the similarity between ambient sound effects and language descriptions. To filter out samples with unclear environmental semantics, we only preserve those with a similarity score greater than 0.7. Finally, we apply ASVA (Zhang et al, 2024) to calculate the similarity between the ambient sound effect track and the video, filtering out scores lower than 0.8, thereby further ensuring temporal synchronization.

Language description annotation. Following InternVid (Wang et al, 2023), we further annotate the colors of central objects in each clip using language descriptions. Specifically, we generate captions for film clips at 20-frame intervals using InstructBLIP (Dai et al, 2023). After that, we integrate these captions into coherent descriptions using GPT-4 (Achiam

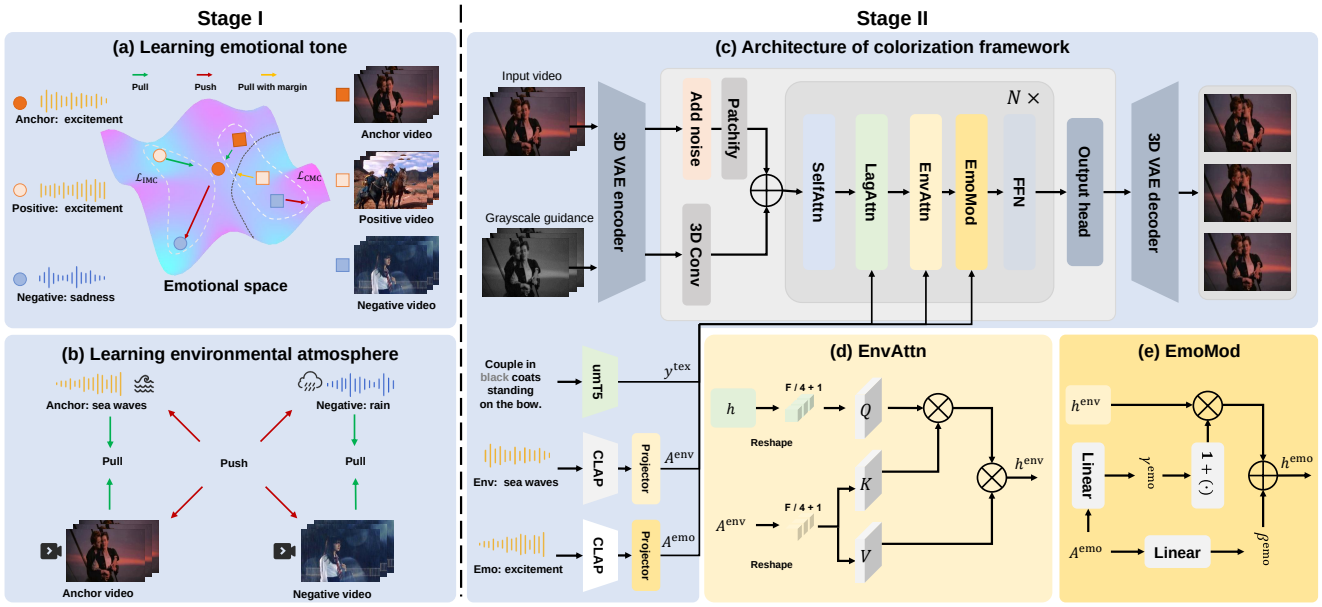

Fig. 5 The pipeline of L-VOCAL. (a) For learning emotional tone, we adopt the Intra-Modality Contrastive (IMC) loss to construct the emotionally-aware feature space and the Cross-Modality Contrastive (CMC) loss to align audio features with the emotional tone in a contrastive learning manner. (b) For learning environmental atmosphere, we introduce the Attentional Cross-modality Contrastive (ACC) loss to align environmentally-aware features with the corresponding sound-producing objects. (c) Our colorization framework builds upon a pre-trained text-to-video generation model, including a self-attention (SelfAttn), a language cross-attention (LagAttn), and feed-forward network (FFN). To further supplement the context not explicitly provided by language descriptions, we use these aligned audio features with (d) the environment cross-attention (EnvAttn) to render a realistic surrounding environmental atmosphere and (e) the emotion modulation (EmoMod) to control the global style of the generated video.

et al, 2023) with pre-defined prompts. Additionally, we employ human annotators to provide captions describing the colors of central objects in the evaluation film clips.

Dataset statistics and analysis. Our processed dataset includes approximately 50K videos for training and 50 videos for testing, each with a resolution of 480×832 pixels. Ambient sound effects, background music, and language descriptions of central objects are carefully annotated for these videos. To clarify the category distribution, we present the distribution of emotion and environment categories in Fig. 3 left. To further illustrate the correspondence between audio cues and visual color in these film clips, quantitative results are presented in Fig. 3 right, where average saturation, value, and hue for each category are calculated. Note that hue is a cyclical value from 0 to 1, where 0 and 1 represent red, 0.33 corresponds to green, and 0.67 to blue. Analysis of the correspondence between background music and color attributes reveals that positive emotions (*i.e.*, "amusement", "excitement", "contentment", and "awe") tend to exhibit higher saturation and brightness, whereas negative emotions (*i.e.*, "anger", "disgust", "fear", and "sadness") correspond to lower saturation and brightness. Moreover, fire-related sounds (*i.e.*, "crackling fires" and "explosions") are generally associated with higher brightness and warm colors while water-related sounds (*i.e.*, "sea waves" and "waterfall burbling") tend to correspond to lower brightness and cool colors. We present samples of our L-VACOLOR dataset in Fig. 4.

4 Method

In this section, we introduce the proposed framework of L-VOCAL, as shown in Fig. 5. Firstly, we build the alignment model to learn the correspondence between color and audio through two key aspects: emotional tone and environmental atmosphere (Sec. 4.1). Subsequently, we develop our colorization model, which leverages language descriptions to specify the colors of central objects while incorporating audio features to enhance the emotional tone and refine the color of the surrounding environment (Sec. 4.2). Finally, we describe the training details (Sec. 4.3).

4.1 Alignment between audio and color

Following previous approaches (Radford et al, 2021; Elizalde et al, 2023; Girdhar et al, 2023), we design the alignment model to align color and audio, enabling the accurate extraction of aligned emotionally-aware and environmentally-aware audio features.

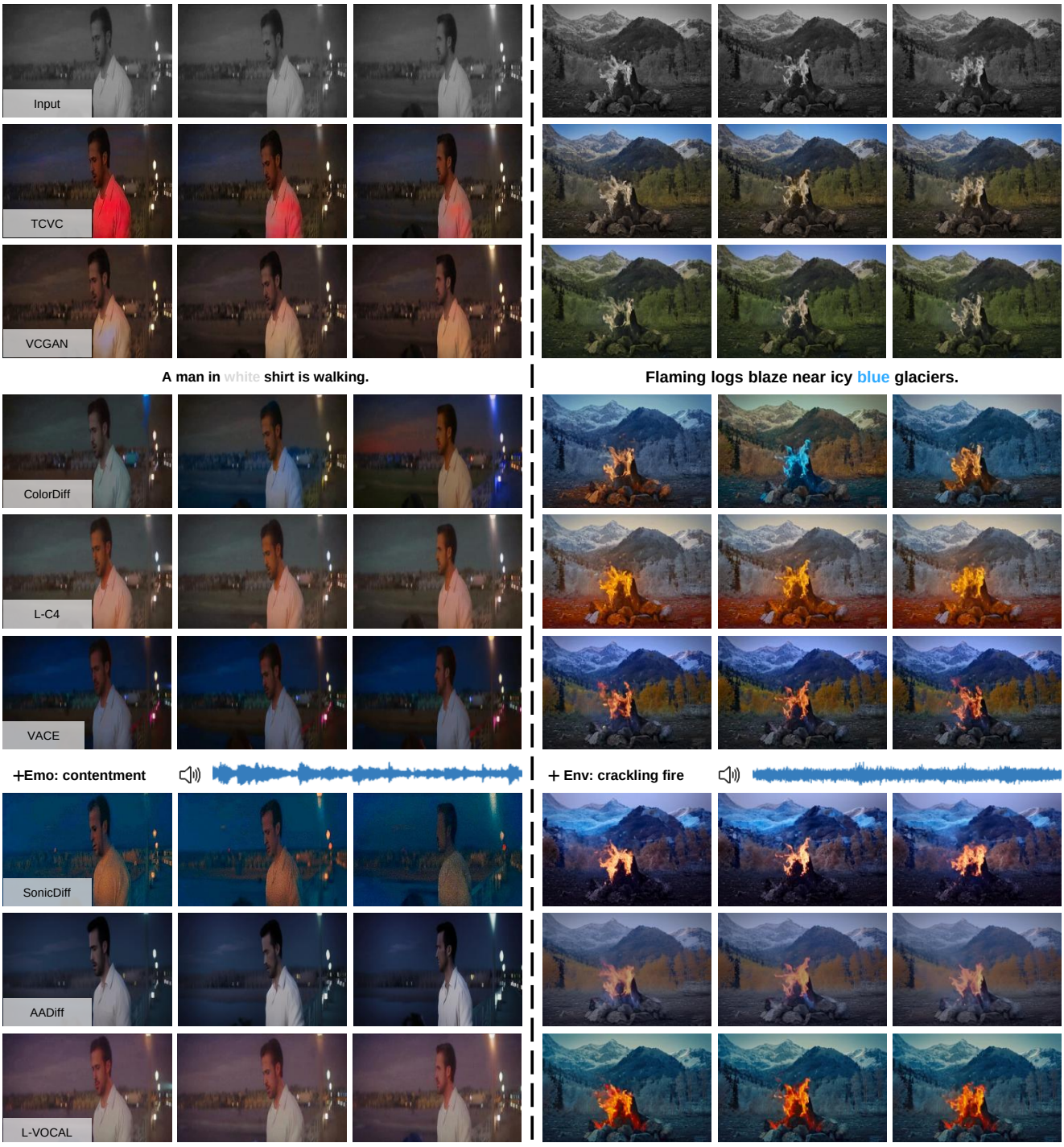

Fig. 6 Visual quality comparison results with relevant methods, including automatic colorization (*i.e.*, TCVC (Liu et al, 2024b) and VCGAN (Zhao et al, 2023)), language-based colorization (*i.e.*, L-C4 (Chang et al, 2024), ColorDiff (Liu et al, 2023), and VACE (Jiang et al, 2025)), and audio-driven approaches (*i.e.*, SonicDiff (Biner et al, 2024) and AADiff (Lee et al, 2023a)).

4.1.1 Multi-modal feature extraction

Auditory features. We leverage the background music and ambient sound effects provided by the L-VACOLOR dataset (Sec. 3) to learn the correspondence with visual colors. Specifically, since the background music of film clips contains abundant emotional cues and the emotion representation within a single shot typically remains consistent, we utilize the CLAP (Elizalde et al, 2023) audio encoder to extract emotionally-aware features from it. This converts the background music into a global vector, which is subsequently mapped into an emotion space using an emotional projec-

tor, represented as $A^{\text{emo}} \in \mathbb{R}^D$, where D is the embedding dimension. Furthermore, since dynamic environmental atmosphere often indicates local scene variations, we extract environmentally-aware features from the ambient sound effect track. The CLAP audio encoder is then employed to generate a temporal feature sequence. This sequence is subsequently projected into an environment space through an environmental projector,represented as $A^{\text{env}} \in \mathbb{R}^{M\times D}$, where M is the length of the sequence.

Visual features. Given a colored video $X \in \mathbb{R}^{F\times H\times W\times 3}$, we employ a pre-trained VAE from the generative model framework (WanTeam, 2025), including a stack of 3D convolutional and flatten operations, to extract a visual semantic representation $V \in \mathbb{R}^{N\times D}$, where $N = (F/4+1)*(H/16)*(W/16)$ represents the sequence length of visual features. To obtain a global representation reflecting the overall scene context, we perform average pooling on the visual representation V across both the temporal and spatial dimensions. This pooled representation is then projected into a shared emotion space using a linear layer, resulting in a vector $V^{\text{emo}} \in \mathbb{R}^D$. Another linear layer is applied to map visual representation into a shared environment space, denoted as $V^{\text{env}} \in \mathbb{R}^{N\times D}$.

4.1.2 Learning emotional tone

We develop intra-modal and cross-modal contrastive losses for effective emotional representation.

Intra-modal loss. This loss is designed to construct the emotionally-aware feature space for audios according to emotion categories defined by Mikel's Wheel (Mikels et al, 2005). Specifically, we sample paired triplets based on their emotion categories, which comprise an anchor feature, a positive feature, and a negative feature, denoted as $[A^{\text{emo}}_{\text{anc}}, A^{\text{emo}}_{\text{pos}}, A^{\text{emo}}_{\text{neg}}]$ for audio features. Here, positive samples match the anchor's emotion category, while negative samples have a different one. Consequently, the intra-modal contrastive loss is formulated as:

$$\mathcal{L}_{\text{IMC}} = \max\big(d(A^{\text{emo}}_{\text{anc}}, A^{\text{emo}}_{\text{pos}}) - d(A^{\text{emo}}_{\text{anc}}, A^{\text{emo}}_{\text{neg}}) + \alpha_1, 0\big), \quad (1)$$

where the margin parameter $\alpha_1 = 0.01$ and $d(\cdot,\cdot)$ represents the cosine distance.

Cross-modal loss. This loss serves to align the audio features A^{emo} with their corresponding visual feature V^{emo} in a shared emotion space. Specifically, we further sample paired triplets for emotional visual features, denoted as $[V^{\text{emo}}_{\text{anc}}, V^{\text{emo}}_{\text{pos}}, V^{\text{emo}}_{\text{neg}}]$. After that, we establish a hierarchical correspondence, where paired audio and visual features have the highest similarity, followed by features sharing the same emotion category, while features with different emotion categories show the greatest distance. As a result, the cross-modal contrastive loss is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{CMC}} = &\max\big(d(A^{\text{emo}}_{\text{anc}}, V^{\text{emo}}_{\text{anc}}) - d(A^{\text{emo}}_{\text{anc}}, V^{\text{emo}}_{\text{pos}}) + \alpha_2, 0\big) + \\ &\max\big(d(A^{\text{emo}}_{\text{anc}}, V^{\text{emo}}_{\text{pos}}) - d(A^{\text{emo}}_{\text{anc}}, V^{\text{emo}}_{\text{neg}}) + \alpha_3, 0\big), \quad (2)\end{aligned}$$

where we set $\alpha_2 = 0.02$ and $\alpha_3 = 0.01$ to control margins between different correspondences.

4.1.3 Learning environmental atmosphere

To enhance the rendered atmosphere of film clips, we further align the environmentally-aware audio features with the corresponding sound-producing objects. Specifically, we calculate the dot product similarity between the i-th local audio feature $a_i \in \mathbb{R}^D$ from the environmentally-aware audio features A^{env} and the j-th local video feature $v_j \in \mathbb{R}^D$ from the environmental visual features V^{env} as $\mathcal{S}(a_i, v_j) = a_i \cdot v_j$. Next, we perform a weighted sum based on these similarity scores to produce the visual representation corresponding to the i-th audio feature, formulated as:

$$\tilde{v}_i = \sum_{j=1}^{N} w_j v_j, \quad \text{where } w_j = \frac{\exp(\mathcal{S}(a_i, v_j))}{\sum_{k=1}^{N}\exp(\mathcal{S}(a_i, v_k))}. \quad (3)$$

As a result, the similarity between audio and visual features can be formulated as:

$$\mathcal{R}(A^{\text{env}}, V^{\text{env}}) = \log(\sum_{i=1}^{M}\exp(\mathcal{S}(a_i, \tilde{v}_i))). \quad (4)$$

To ensure that corresponding video and audio pairs have higher similarity within the fed batch of size B, we calculate the attentional cross-modal contrastive loss:

$$\mathcal{L}_{\text{ACC}} = -\sum_{i=1}^{B}\log\left(\frac{\exp(\mathcal{R}(A^{\text{env}}_i, V^{\text{env}}_i)/\tau)}{\sum_{j=1}^{B}\exp(\mathcal{R}(A^{\text{env}}_i, V^{\text{env}}_j)/\tau)}\right), \quad (5)$$

where the temperature hyperparameter $\tau = 0.1$ controls the smoothness of the distribution. Finally, these losses are combined to train the alignment model:

$$\mathcal{L}_{\text{ALI}} = \lambda_1\mathcal{L}_{\text{IMC}} + \lambda_2\mathcal{L}_{\text{CMC}} + \mathcal{L}_{\text{ACC}}, \quad (6)$$

where $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ are loss function weights.

4.2 Language-based colorization with audio alignment

We build our colorization model upon a text-to-video generation framework (WanTeam, 2025). Leveraging its robust language understanding capabilities, our model precisely specifies colors for central objects. Additionally, we incorporate aligned audio features to guide the generation of emotional tone and the surrounding environment.

Table 1 Quantitative results on two test sets. Throughout this paper, ↑ (↓) means higher (lower) is better. Best scores are highlighted in **bold**. CDC is reported with a scale of 1000 for better readability.

Method	L-VACOLOR test								Black-and-white test			
	Color. ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	CDC ↓	Temp-S ↑	CAM-S ↑	Color. ↑	CDC ↓	Temp-S ↑	CAM-S ↑
VCGAN (Zhao et al, 2023)	19.78	24.77	0.863	0.202	2457.81	2.524	0.233	0.468	19.66	6.457	0.222	0.380
TCVC (Liu et al, 2024b)	25.48	24.79	0.885	0.209	2825.24	2.917	0.225	0.478	25.79	7.753	0.221	0.382
ColorDiff (Liu et al, 2023)	26.88	22.84	0.895	0.211	2247.36	4.216	0.234	0.488	24.16	6.955	0.257	0.452
L-C4 (Chang et al, 2024)	23.90	25.17	0.912	0.202	1919.56	1.458	0.237	0.486	21.65	5.432	0.257	0.454
VACE (Jiang et al, 2025)	25.37	23.86	0.897	0.304	2968.25	3.545	0.225	0.443	25.26	8.123	0.249	0.468
AADiff (Lee et al, 2023a)	19.87	22.24	0.848	0.252	2563.44	2.477	0.241	0.468	18.91	7.568	0.245	0.414
SonicDiff (Biner et al, 2024)	19.71	22.97	0.852	0.266	2245.99	2.747	0.232	0.466	19.33	6.985	0.246	0.424
W/ RTD	26.23	24.92	0.928	0.197	1985.47	1.656	0.245	0.492	25.77	7.342	0.237	0.451
W/o AC	23.11	24.77	0.924	0.206	2256.33	1.771	0.247	0.480	21.77	7.868	0.251	0.458
W/o ETG	24.55	24.88	0.910	0.212	2274.19	1.601	0.245	0.494	24.77	6.274	0.257	0.454
W/o EAG	27.06	24.18	0.930	0.221	2436.87	1.567	0.250	0.488	25.99	5.787	0.249	0.462
W/o ETA	26.94	23.75	0.918	0.200	2147.26	1.878	0.239	0.504	25.22	8.023	0.247	0.456
W/o EAA	27.07	23.10	0.907	0.215	1988.93	1.784	0.249	0.488	24.51	7.651	0.254	0.458
Ours (L-VOCAL)	**27.81**	**25.59**	**0.929**	**0.184**	**1654.54**	**1.341**	**0.260**	**0.510**	**26.87**	**5.146**	**0.258**	**0.472**

Table 2 Qualitative comparison with relevant methods in model efficiency and complexity.

Method	VCGAN	TCVC	L-C4	ColorDiff	AADiff	SonicDiff	VACE	Ours (L-VOCAL)
Trainable params	106.02 M	198.38 M	1.43 B	1.11 B	116.01 M	109.47 M	1.95 B	1.52 B
GPU memory usage	4.61 GB	3.43 GB	12.23 GB	7.27 GB	13.71 GB	8.23 GB	25.94 GB	7.06 GB
Inference time	12.3s	15.7s	334.6s	539.7s	457.5s	427.1s	313.7s	179.9s

4.2.1 Architecture of colorization framework

Our video colorization model integrates a variational autoencoder (VAE) to compress high-dimensional video data into a compact latent representation. This representation is then processed by a diffusion transformer, which converts random noise into structured latent code based on language descriptions and aligned audio conditions. To preserve the original grayscale video structure, we incorporate grayscale guidance. The diffusion model is trained using rectified flow loss.

Variational autoencoder. To enable scalable and efficient training, a 3D causal variational autoencoder (VAE) is used to extract compact latent representations from high-dimensional video data. Specifically, the VAE encoder compresses a colored video $X \in \mathbb{R}^{F\times H\times W\times 3}$ into a latent code Z with a shape of $[1 + F/4, H/8, W/8, 16]$. Then, the VAE decoder with a symmetric structure reconstructs the video from this compressed latent code.

Diffusion transformer. The diffusion transformer is designed to convert noise into a structured latent code. It primarily comprises three components: a patchify module, transformer blocks, and an output head. The patchify module uses a 3D convolutional layer to further spatially compress the noisy latent code Z_t by a factor of 2. We then flatten this output to obtain $h_{\text{in}} \in \mathbb{R}^{N\times D}$, which serves as the input for the transformer blocks, where $N = (F/4+1)*(H/16)*(W/16)$ represents the sequence length of visual features. Within each transformer block, a self-attention mechanism is first integrated to model complex dependencies within the visual sequence. After that, language descriptions are encoded by umT5 (Chung et al, 2023) and injected to guide the generation process via a language cross-attention. Following this, emotionally-aware A^{emo} and environmentally-aware A^{env} audio features are injected into the latent code to control the generation of the emotional tone and the surrounding environment, respectively. Finally, a feed-forward network is adopted to further extract high-dimensional features. The output head then projects the output of the transformer block into the dimensions of the original latent code Z.

Grayscale guidance. Since the VAE encoder requires a 3-channel input, we first repeat the single color channel of the grayscale video to form a 3-channel input, resulting in $X^{\text{g}} \in \mathbb{R}^{F\times H\times W\times 3}$. The VAE encoder then processes this grayscale input to produce a latent representation Z^{g}. A 3D convolution further compresses and flattens Z^{g} to match the shape of h_{in}. This grayscale feature is added to h_{in} via element-wise summation, thereby injecting structural information from the grayscale video into the transformer's input.

Rectified flow loss. Our model is formulated as a rectified flow model. Compared to traditional diffusion models, this formulation offers improved stability and robustness during training. Its forward process is formulated as a linear interpolation between the latent code Z and random noise $\epsilon \sim \mathcal{N}(0, 1)$, expressed as $Z_t = tZ + (1 - t)\epsilon$. The backward process involves predicting a velocity u_θ to transform the noise into structured data $\mathrm{d}Z/\mathrm{d}t =$

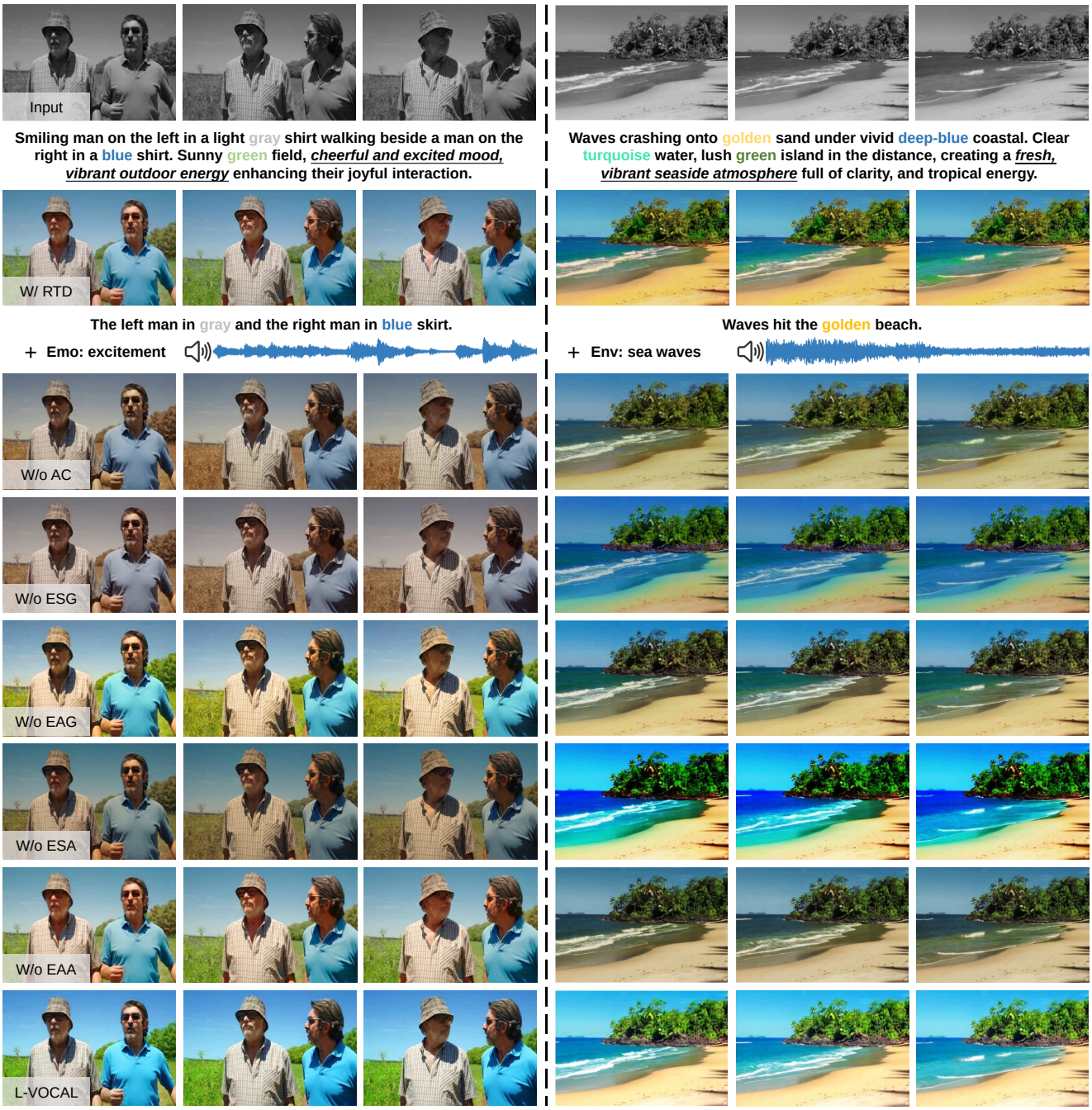

Fig. 7 Ablation study results, including discarding audio conditions (*i.e.*, *W/* RTD and *W/o* AC), removing audio guidance modules (*i.e.*, *W/o* ETG and *W/o* EAG), and disabling audio alignment modules (*i.e.*, *W/o* ETA and *W/o* EAA).

$u_\theta(Z_t, Z^{\mathrm{g}}, y^{\mathrm{tex}}, A^{\mathrm{env}}, A^{\mathrm{emo}}, t)$, where u_θ represents the predicted velocity conditioned on the latent code Z_t, grayscale video Z^{g}, language description y^{txt}, and time t. The ground truth velocity is defined as $v_t = Z - \epsilon$. Thus, the loss function can be formulated as the mean squared error (MSE) between the model output and v_t:

$$\mathcal{L}_{\mathrm{MSE}} = \mathbb{E}_{Z,\epsilon,t}||u_\theta(Z_t, Z^{\mathrm{g}}, y^{\mathrm{tex}}, A^{\mathrm{env}}, A^{\mathrm{emo}}, t) - v_t||^2. \quad (7)$$

4.2.2 Colorization with aligned audio guidance

Environmental guidance. To render a realistic surrounding environmental atmosphere, we introduce environmentally-aware audio features A^{env} to guide the colorization process via an environment cross-attention. Specifically, this cross-attention is performed frame-wise to learn the temporal correspondence between audio and video. We first divide the environmentally-aware audio features A^{env} into $(F/4+1)$

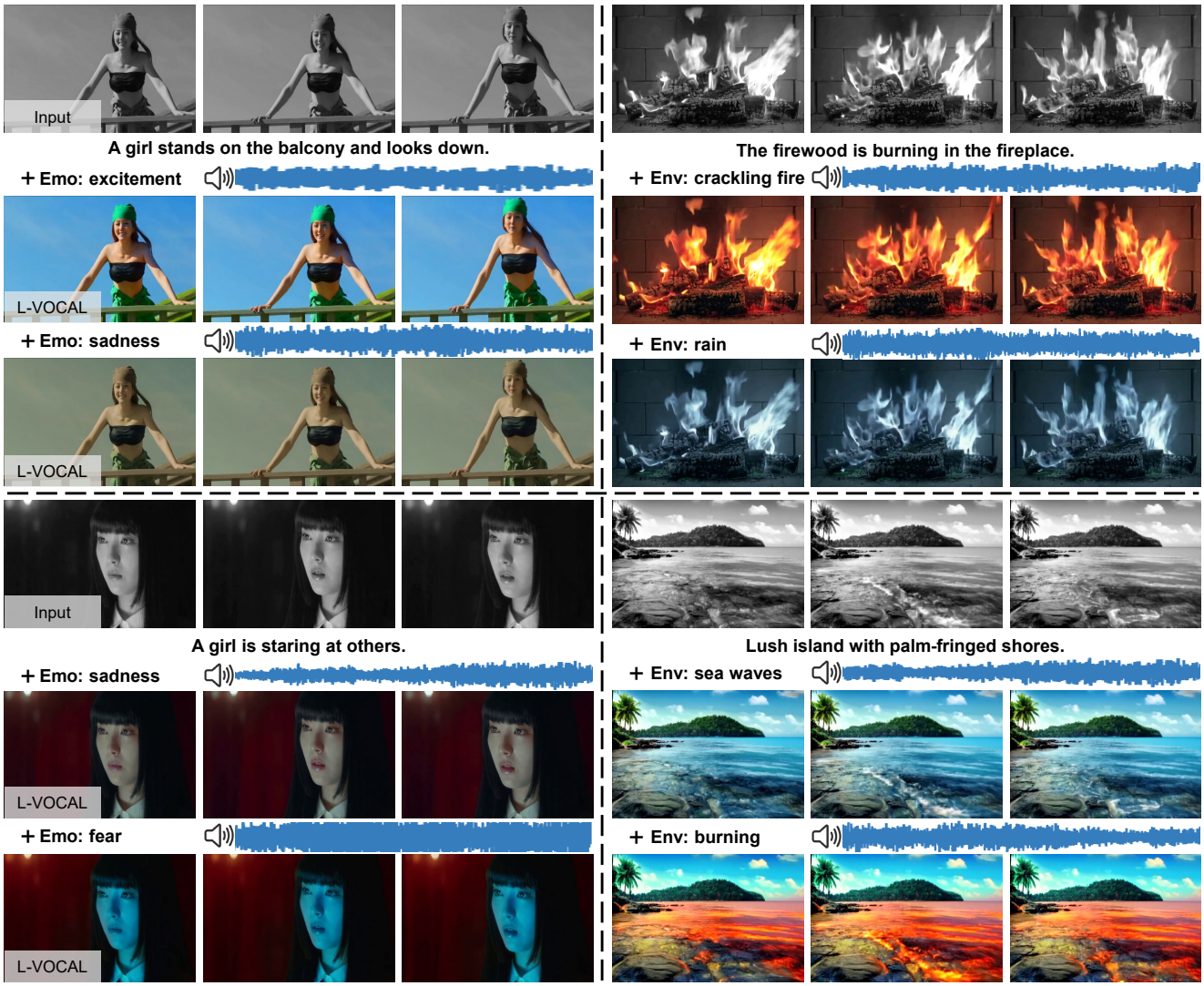

Fig. 8 Qualitative colorization results demonstrating audio condition controllability. Given the same language descriptions, L-VOCAL produces diverse colorization results that align with the emotional tones and environmental atmospheres of the provided audio conditions.

segments, matching the number of latent video frames. Each segment is then injected into the corresponding frame representation:

$$h_f^{\text{env}} = \text{CrossAttn}(h_f, A_f^{\text{env}}) + h_f, \tag{8}$$

where h_f represents the f-th latent frame and A_f^{env} represents the corresponding audio segment. $\text{CrossAttn}(\cdot,\cdot)$ refers to a cross-attention module where the latent frame serves as the query, and the audio feature segment as the key and value.

Emotional guidance. To ensure emotional synchronization between colorization results and the provided background music, we incorporate emotionally-aware audio features A^{emo} into the colorization process. Inspired by the effectiveness of feature modulation (Huang and Belongie, 2017) in controlling the global style of generated results, we use an emotion modulation to integrate emotional audio features. Specifically, we process the emotionally-aware audio feature A^{emo} through two linear layers to produce the modulation parameters γ^{emo} and β^{emo}:

$$\gamma^{\text{emo}} = A^{\text{emo}} \cdot W^{\gamma}, \quad \beta^{\text{emo}} = A^{\text{emo}} \cdot W^{\beta}, \tag{9}$$

where the linear layer parameters W^{γ} and W^{β} are initialized to zero to ensure that the output remains unchanged at the beginning of training. These parameters then modulate the latent feature h^{env} through scaling and shifting, ensuring the colorization captures nuanced emotional tones aligned with the accompanying audio:

$$h^{\text{emo}} = h^{\text{env}} \odot (\gamma^{\text{emo}} + 1) + \beta^{\text{emo}}. \tag{10}$$

4.3 Training and evaluation details

We train the alignment model with a batch size of 16 and the colorization model with a batch size of 2, requiring approx-

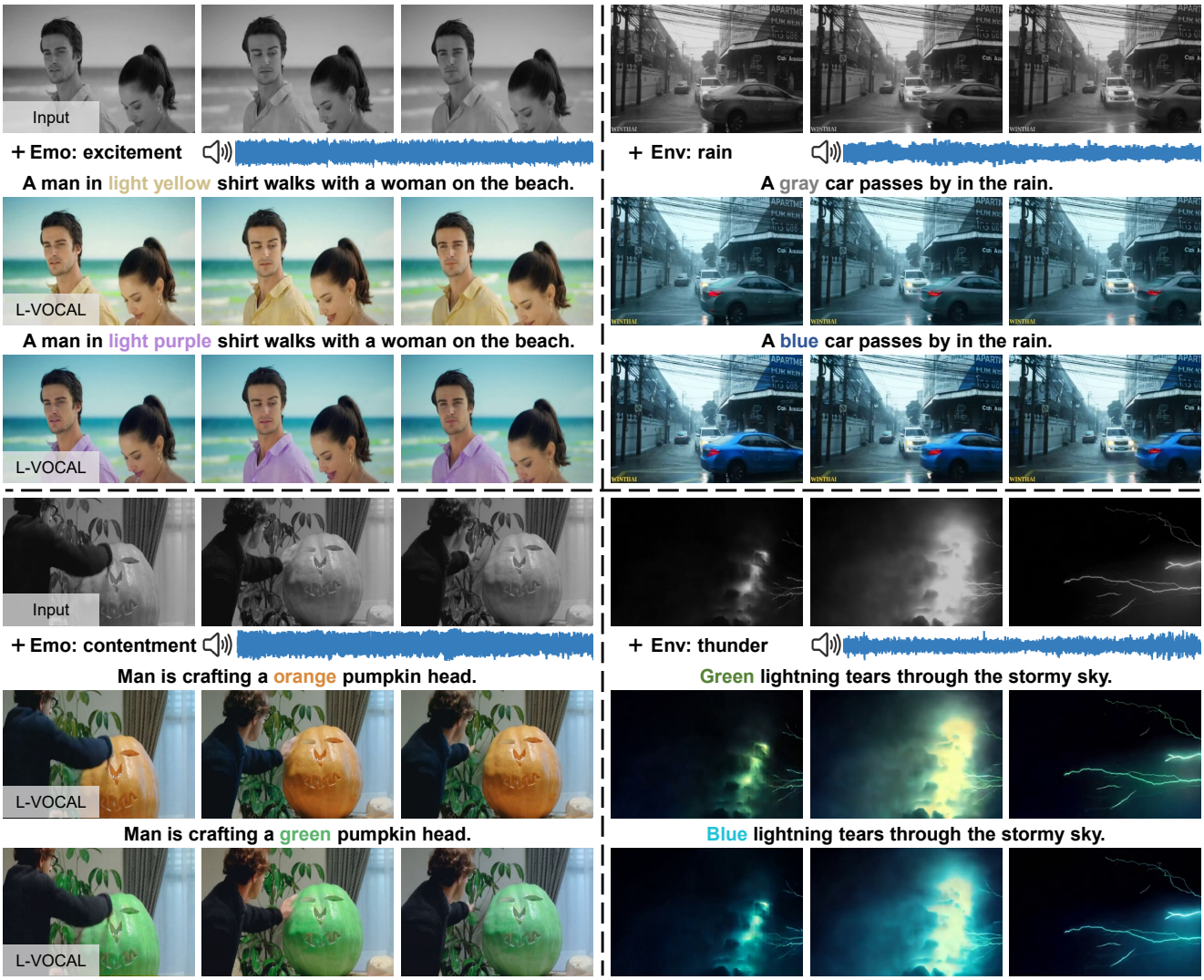

Fig. 9 Qualitative colorization results demonstrating language description controllability. Given the same audio conditions, L-VOCAL produces diverse results that specify central object colors by modifying text descriptions.

imately 36 hours and 72 hours, respectively. AdamW optimizer (Loshchilov and Hutter, 2019) is used with a learning rate of 1×10^{-5} and momentum parameters $\beta_1 = 0.99$ and $\beta_2 = 0.999$. The clip length for training is set to $F = 81$. All experiments are conducted on 4 NVIDIA A100 40G graphics cards. During inference, L-VOCAL takes approximately 179.9 seconds and consumes 7.06 GB of GPU memory to colorize an 81-frame monochrome video (480×832) on a single NVIDIA A100 GPU.

5 Experiments

Metrics. We adopt eight metrics to evaluate our proposed colorization model comprehensively. Following previous video colorization methods (Liu et al, 2024b; Zhao et al, 2023; Bozic et al, 2024; Li et al, 2024; Yang et al, 2024b; Liu et al, 2023), we assess the video quality using: *(i)* the **Fréchet Video Distance** (FVD) (Unterthiner et al, 2019) for perceptual realism; *(ii)* the **Colorfulness score** (Color.) (Hasler and Suesstrunk, 2003) for color vibrancy; *(iii)* the **PSNR** (Huynh-Thu and Ghanbari, 2008), **SSIM** (Wang et al, 2004), and **LPIPS** (Zhang et al, 2018) for perceptual quality; and *(iv)* **Color Distribution Consistency** (CDC) (Liu et al, 2024b) and **Temporal Score** (Temp-S) (Shen et al, 2024) for temporal consistency. Additionally, we introduce the **Color-Audio Matching Score** (CAM-S), to assess the alignment between colorization results and the corresponding audio. Specifically, we leverage the Qwen2.5-Omni (Xu et al, 2025) 7B model to assign a relevance score ranging from 0 to 10, quantifying the correlation strength between visual chromatic attributes

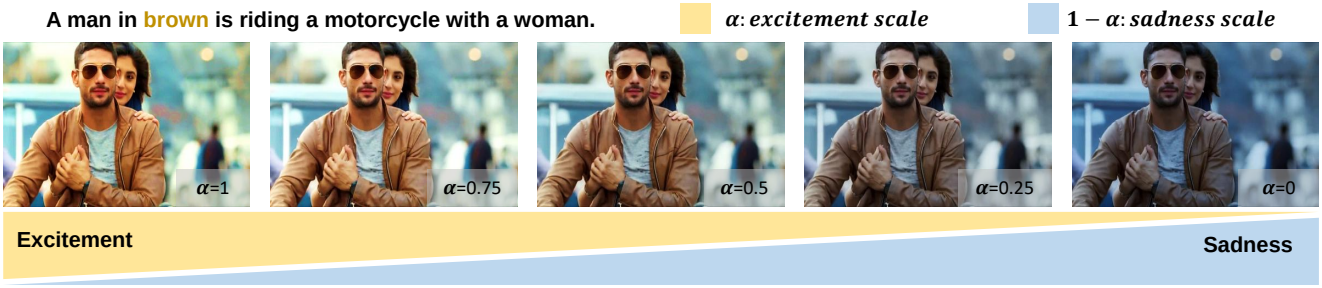

Fig. 10 Qualitative results for linearly interpolated emotional audio features. As features shift from excitement to sadness, the model enables fine-grained controllability over the emotional tone.

Fig. 11 Colorization results under conflicting text and audio conditions. When replacing the original audio with a conflicting one (different emotional tone), L-VOCAL adjusts the global style accordingly while preserving the semantic object colors defined by the text.

and emotionally-aware and environmentally-aware audio features. Higher scores indicate stronger perceptual coherence.

Evaluation dataset. Following the test set sizes of the previous works (Yang et al, 2024b; Liu et al, 2024b; Zhao et al, 2023; Li et al, 2024), we construct two test sets: *(i)* **L-VACOLOR test**: We randomly select 50 color film clips from the L-VACOLOR dataset, ensuring no overlap with its training set; *(ii)* **Black-and-white test**: We collect 50 black-and-white film clips to evaluate the model's generalization and real-world applicability. As ground truth color is unavailable for these clips, we manually annotate the colors of salient objects and compute only a subset of the metrics. We utilize Bandit (Watcharasupat et al, 2024) for automated track separation, which requires approximately 2.1 GB of GPU memory and takes only about 1.16 seconds per sample. This demonstrates that processing monaural audio does not impose a significant computational burden on our model in real-world application scenarios.

5.1 Comparison

Qualitative comparisons. As shown in Fig. 6, we conduct qualitative comparisons with relevant video editing methods. Compared to automatic colorization approaches (*i.e.*, TCVC (Liu et al, 2024b) and VCGAN (Zhao et al, 2023)), our L-VOCAL effectively addresses the inherent ambiguity in color assignment (*e.g.*, the man's white clothes in Fig. 6 left). Compared to language-based colorization approaches (*i.e.*, L-C4 (Chang et al, 2024), ColorDiff (Liu et al, 2023), and VACE (Jiang et al, 2025)), our L-VOCAL leverages environmentally-aware audio features to supplement the environmental atmosphere (*e.g.*, the crackling fire in Fig. 6 right). Furthermore, in contrast to the audio-driven approaches (*i.e.*, SonicDiff (Biner et al, 2024) and AADiff (Lee et al, 2023a), adapted by fine-tuning on L-VACOLOR), our method specifically establishes the correspondence between emotional tone and background music, thereby better reflecting the filmmaker's original expression (*e.g.*, the contentment tone in Fig. 6 left).

Quantitative comparisons. We further quantitatively compare our L-VOCAL against the relevant automatic, language-based, and audio-driven video colorization approaches previously discussed. As shown in Table 1, our L-VOCAL achieves the best scores across all evaluation metrics on both test sets, demonstrating superior colorization quality and temporal consistency. The top CAM-S further highlights the strong alignment between the colorization results and the accompanying audio.

Efficiency comparison. We additionally compare the efficiency and complexity of L-VOCAL against relevant methods, reporting the number of trainable parameters, GPU memory footprint, and average inference time per frame. To ensure a fair comparison, all models are evaluated on a single NVIDIA A100 GPU at a resolution of 480×832. As presented in Table 2, L-VOCAL demonstrates a significant inference speed advantage over comparable diffusion-based methods (*e.g.*, L-C4 (Chang et al, 2024), ColorDiffuser (Liu et al, 2023), and VACE (Jiang et al, 2025)). While some

non-diffusion models (*e.g.*, TCVC (Liu et al, 2024b) and VCGAN (Zhao et al, 2023)) still offer faster inference speeds, this efficiency comes at the cost of colorization quality, as reflected in the lower performance on generative metrics (*e.g.*, FVD and CDC) shown in Table 1.

5.2 Ablation

To evaluate the effectiveness of our proposed modules, we conduct five ablation studies. Quantitative and qualitative results are respectively presented in Fig. 7 and Table 1.

***W/* Rich Text Descriptions (RTD).** We discard all audio modules and instead provide the model with rich text descriptions that explicitly describe the emotional tone and environmental atmosphere. This leads to noticeably inferior emotional and environmental rendering. As shown in Fig. 7 (left), the model generates results with insufficient saturation to express the excited emotional tone specified in the text.

***W/o* Audio Condition (AC).** We discard audio conditions and their corresponding injection modules from the denoising network. This leads our model to degrade into a language-based video colorization model without audio alignment. As shown in Fig. 7 (left), the produced colorization results tend to be neutral and lack an accurate emotional tone and environmental atmosphere.

***W/o* Emotional Tone Guidance (ETG).** We remove the emotional modulation (EmoMod) module from the colorization framework, thereby discarding emotional guidance from the background music. This results in the unmatched emotional tones of colorization. As shown in Fig. 7 (left), the colorization results cannot effectively express the tone of excitement.

***W/o* Environmental Atmosphere Guidance (EAG).** We remove the environmental cross-attention (EnVAttn) module from the colorization framework, preventing the colorization model from receiving guidance from ambient sound effects. This leads to an inaccurate environmental atmosphere. As illustrated in Fig. 7 (right), the model failed to reproduce the vivid color of the seawater.

***W/o* Emotional Tone Alignment (ETA).** We disable the alignment module for learning emotional tone. This results in injected emotionally-aware audio features being unaligned, and the colorization model loses pre-trained correspondence with the emotional tone. As shown in Fig. 7 (left), the overall tone becomes dull, mismatching exciting music.

***W/o* Environmental Atmosphere Alignment (EAA).** We disable the alignment module for learning the environmental atmosphere. Therefore, the environmentally-aware audio features used for cross-attention are unaligned, leading the colorization results to be less sensitive to ambient sound effects. As shown in Fig. 7 (right), the color contrast between the beach and the sea is reduced.

Table 3 User study results. Our L-VOCAL produces a higher score than relevant approaches.

CVE	TCVC	VCGAN	L-C4	ColorDiffuser	L-VOCAL
	10.2 %	13.8 %	15.4 %	14.4 %	**46.2 %**
AAE	AADiff	SonicDiff	L-C4	ColorDiffuser	L-VOCAL
	12.4 %	10.6 %	20.2 %	13.6 %	**43.2 %**

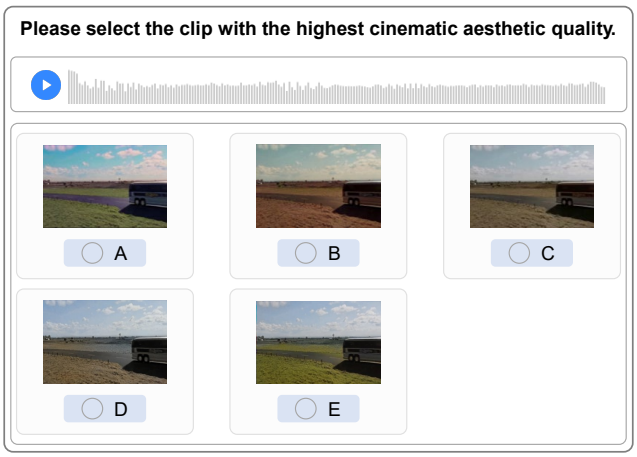

Fig. 12 The interface used for the user study, where audio and video colorization results are played synchronously. Users are required to select their preference according to the given requirement.

5.3 Controllability study

Audio control. We demonstrate audio controllability by replacing the provided audio condition. As a result, given the same grayscale video and language descriptions, L-VOCAL's colorization results align with the emotional tones and environmental atmosphere of the provided audio features. As shown in the first row of Fig. 8 (left), L-VOCAL renders a vibrant color for the excitement tone, while the sadness tone is well applied even when the character has a happy expression. As shown in the second row of Fig. 8 (right), sea wave sounds produce cool colors, whereas the burning sound renders the sea surface in vibrant red hues. These results demonstrate the effectiveness of audio condition controllability.

Language control. We demonstrate language controllability by replacing the color in language descriptions of the central object. As shown in Fig. 9, the modified language description effectively alters the color of the specified instances, while preserving emotional tone and environmental atmosphere alignment with the provided audio condition.

Emotion control. We demonstrate the fine-grained controllability of emotional guidance by linearly interpolating emotional audio features between two distinct emotions. As shown in Fig. 10, when the features shift from excitement to sadness, colorization results demonstrate a smooth transition from vibrant and warm tones to desaturated and cool

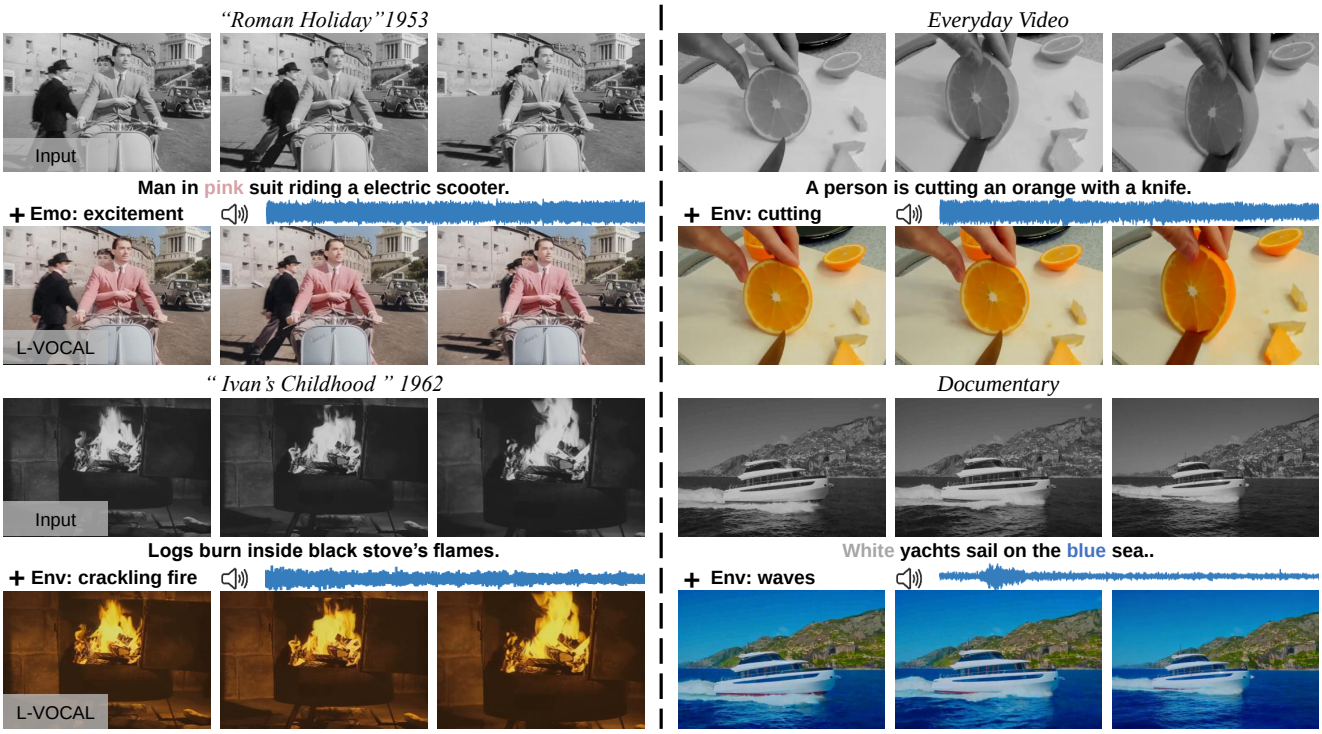

Fig. 13 Qualitative colorization results for diverse real-world scenarios, including realistic black-and-white sound films (left), along with everyday videos and documentaries (right).

Fig. 14 Colorization results for long video clips. L-VOCAL colorizes monochrome videos exceeding 200 frames using a sliding window strategy, maintaining temporal consistency without degradation.

tones. This confirms that our model achieves fine-grained controllability over the global style.

Condition disentanglement. We demonstrate that our model effectively disentangles conflicting control signals from text and audio by visualizing a potential conflict case in Fig. 11. Specifically, we fix the text description to present a "happy" emotion for both samples, while introducing corresponding "amusement audio" and conflicting "sad audio" in the first and second rows, respectively. As a result, the colorization results with sad audio present reduced brightness and saturation to render the global emotional style compared to the case with amusement audio, while preserving the semantic colors of the central objects as defined by the text.

5.4 User study

In addition to qualitative and quantitative comparisons, we conduct two user studies to evaluate human observer preferences: *(i)* **Cinematic Visual Effects (CVE).** Participants are shown colorization results produced by L-VOCAL alongside relevant comparison methods and are asked to select the one with the highest cinematic aesthetic quality. *(ii)* **Audio Alignment Evaluation (AAE).** Participants are provided with the original film audio and colorization results from L-VOCAL and relevant comparison methods. They are asked to select the colorization results most aligned with the audio condition. We conduct experiments on Amazon Mechanical Turk (AMT). Each experiment uses 20 randomly selected samples

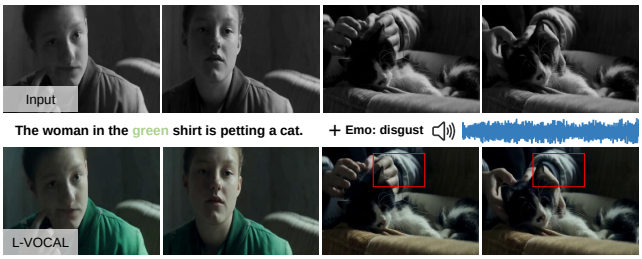

Fig. 15 Visualization of failure cases. L-VOCAL fails to track the identity of characters across shots in multi-shot videos.

from the L-VACOLOR test set, and outcomes are independently evaluated by 25 volunteers. As shown in Table 3, our model achieves highest preference scores in both experiments. The experimental interface is presented in Fig. 12.

5.5 Application

Generalization ability. To demonstrate that the learned color priors generalize well to diverse real-world scenarios, we apply L-VOCAL to colorize realistic black-and-white sound films, everyday videos, and documentaries. As shown in Fig. 13, L-VOCAL achieves robust colorization results that capture the appropriate emotional tone and enhance the environmental atmosphere on these samples, validating its strong cross-domain generalization capability.

Long video colorization. We extend L-VOCAL to colorize longer monochrome videos via a standard sliding window inference strategy. By processing overlapping segments across the temporal dimension and fusing the results to ensure smooth transitions, L-VOCAL can generate temporally consistent colorization results for videos exceeding 200 frames without performance degradation, as shown in Fig. 14.

6 Conclusion

In this paper, we propose L-VOCAL, a novel language-based video colorization with audio alignment to supplement context not explicitly provided by language descriptions. By establishing the correspondence between color and audio for emotional tone and environmental atmosphere, we obtain aligned audio features to guide the colorization process through designed condition injection modules. Additionally, we contribute the L-VACOLOR dataset, specifically tailored for this task, which provides abundant film clips with diverse audio and color tones for both training and evaluation. Experimental results demonstrate that L-VOCAL produces colorization results that faithfully capture filmmakers' artistic expression, achieving emotionally-aware and environmentally-aware colorization.

Limitations. While our model performs robustly within single continuous video shots, it faces challenges in maintaining semantic consistency across abrupt shot transitions in real-world films. As visualized in Fig. 15, the model fails to track the identity of the character across shots, resulting in the woman's sleeve incorrectly shifting from a dark color to white in the subsequent shot. This issue could be mitigated by designing a cross-shot identity correspondence mechanism, which we leave for future work.

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grant No. 62136001, PKU Kunpeng&Ascend Center of Excellence, and BUPT Excellent Ph.D. Students Foundation No. CX20242080.

Data availability. Datasets used in this study are available from co-first authors on reasonable request.

References

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al (2023) GPT-4 technical report. arXiv preprint arXiv:230308774

Bain M, Nagrani A, Brown A, Zisserman A (2020) Condensed movies: Story based retrieval with contextual embeddings. 2005.04208

Barbiere JM, Vidal A, Zellner DA (2007) The color of music: Correspondence through emotion. Empirical Studies of the Arts 25:193–208

Biner BC, Sofian FM, Karakaş UB, Ceylan D, Erdem E, Erdem A (2024) Sonicdiffusion: Audio-driven image generation and editing with pretrained diffusion models. 2405.00878

Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, Levi Y, English Z, Voleti V, Letts A, Jampani V, Rombach R (2023) Stable video diffusion: Scaling latent video diffusion models to large datasets. 2311.15127

Bozic V, Djelouah A, Zhang Y, Timofte R, Gross M, Schroers C (2024) Versatile vision foundation model for image and video colorization. In: ACM SIGGRAPH

Brooks T, Holynski A, Efros AA (2023) Instructpix2pix: Learning to follow image editing instructions. In: CVPR, pp 18392–18402

Chang Z, Weng S, Zhang P, Li Y, Li S, Shi B (2023) L-CAD: Language-based colorization with any-level descriptions using diffusion priors. In: NeurIPS, vol 36, pp 77174–77186

Chang Z, Weng S, Ouyang H, Li Y, Li S, Shi B (2024) L-C4: Language-based video colorization for creative and consistent color. 2410.04972

Chen H, Xie W, Vedaldi A, Zisserman A (2020) Vggsound: A large-scale audio-visual dataset. In: ICASSP, pp 721–725

Chen H, Zhang Y, Cun X, Xia M, Wang X, Weng C, Shan Y (2024) VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR, pp 7310–7320

Cheng Z, Yang Q, Sheng B (2015) Deep colorization. In: ICCV, pp 415–423

Chu Y, Xu J, Yang Q, Wei H, Wei X, Guo Z, Leng Y, Lv Y, He J, Lin J, Zhou C, Zhou J (2024) Qwen2-audio technical report. arXiv preprint arXiv:240710759

Chung HW, Constant N, Garcia X, Roberts A, Tay Y, Narang S, Firat O (2023) Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. 2304.09151

Dai W, Li J, Li D, Tiong AMH, Zhao J, Wang W, Li B, Fung P, Hoi S (2023) InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS, vol 36, pp 49250–49267

Elizalde B, Deshmukh S, Al Ismail M, Wang H (2023) Clap learning audio concepts from natural language supervision. In: ICASSP, pp 1–5

Feng R, Weng W, Wang Y, Yuan Y, Bao J, Luo C, Chen Z, Guo B (2024) Ccedit: Creative and controllable video editing via diffusion models. In: CVPR, pp 6712–6722

Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, Misra I (2023) Imagebind: One embedding space to bind them all. In: CVPR, pp 15180–15190

Hasler D, Suesstrunk SE (2003) Measuring colorfulness in natural images. In: Human vision and electronic imaging VIII, pp 87–95

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, pp 770–778

He S, Ming A, Yaqi L, Jinyuan S, ShunTian Z, Huadong M (2023) Thinking image color aesthetics assessment: Models, datasets and benchmarks. In: ICCV, pp 21781–21790

Huang X, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV, pp 1510–1519

Huynh-Thu Q, Ghanbari M (2008) Scope of validity of PSNR in image/video quality assessment. Electronics letters 44(13):800–801

Iizuka S, Simo-Serra E (2019) Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. ACM TOG 38(6):1–13

Jeong Y, Ryoo W, Lee S, Seo D, Byeon W, Kim S, Kim J (2023) The power of sound (tpos): Audio reactive video generation with stable diffusion. In: ICCV, pp 7822–7832

Jiang Z, Han Z, Mao C, Zhang J, Pan Y, Liu Y (2025) Vace: All-in-one video creation and editing. arXiv preprint arXiv:250307598

Lee S, Kong C, Jeon D, Kwak N (2023a) Aadiff: Audio-aligned video synthesis with text-to-image diffusion. 2305.04001

Lee SH, Oh G, Byeon W, Kim C, Ryoo WJ, Yoon SH, Cho H, Bae J, Kim J, Kim S (2022a) Sound-guided semantic video generation. 2204.09273

Lee SH, Roh W, Byeon W, Yoon SH, Kim C, Kim J, Kim S (2022b) Sound-guided semantic image manipulation. In: CVPR, pp 3367–3376

Lee SH, Kim S, Yoo I, Yang F, Cho D, Kim Y, Chang H, Kim J, Kim S (2023b) Soundini: Sound-guided diffusion for natural video editing. 2304.06818

Lei C, Chen Q (2019) Fully automatic video colorization with self-regularization and diversity. In: CVPR, pp 3753–3761

Li J, Zhao H, Wang Y, Lin J (2024) Towards photorealistic video colorization via gated color-guided image diffusion models. In: ACM MM, p 10891–10900

Liu H, Xie M, Xing J, Li C, Wong TT (2023) Video colorization with pre-trained text-to-image diffusion models. 2306.01732

Liu S, Zhang Y, Li W, Lin Z, Jia J (2024a) Video-p2p: Video editing with cross-attention control. In: CVPR, pp 8599–8608

Liu X, Wan L, Qu Y, Wong TT, Lin S, Leung CS, Heng PA (2008) Intrinsic colorization. ACM TOG 27:152:1–152:9

Liu Y, Zhao H, Chan KC, Wang X, Loy CC, Qiao Y, Dong C (2024b) Temporally consistent video colorization with deep feature propagation and self-regularization learning. CVM 10(2):375–395

Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: ICLR

Mikels J, Fredrickson B, Samanez-Larkin G, Lindberg C, Maglio S, Reuter-Lorenz P (2005) Emotional category data on images from the international affective picture system. Behavior research methods 37:626–630

Palmer SE, Schloss KB, Xu Z, Prado-León LR (2013) Music-color associations are mediated by emotion. Proceedings of the National Academy of Sciences 110:8836–8841

Piczak KJ (2015) ESC: Dataset for Environmental Sound Classification. In: ACM MM, p 1015–1018

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: ICML, pp 8748–8763

Shen K, Quan R, Zhu L, Xiao J, Yang Y (2024) Audioscenic: Audio-driven video scene editing. 2404.16581

Soucek T, Lokoc J (2024) Transnet v2: An effective deep network architecture for fast shot transition detection. In: ACM MM, pp 11218–11221

Stewart S, Avramidis* K, Feng* T, Narayanan S (2024) Emotion-aligned contrastive learning between images and music. In: ICASSP, pp 8135–8139

Tsiamas I, Pascual S, Yeh C, Serrà J (2024) Sequential contrastive audio-visual learning. 2407.05782

Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S (2019) FVD: A new metric for video generation. In: ICLR Workshop

Wan Z, Zhang B, Chen D, Liao J (2022) Bringing old films back to life. In: CVPR, pp 17694–17703

Wang Y, He Y, Li Y, Li K, Yu J, Ma X, Li X, Chen G, Chen X, Wang Y, et al (2023) InternVid: A large-scale video-text dataset for multimodal understanding and generation. In: ICLR

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: From error visibility to structural similarity. TIP 13(4):600–612

WanTeam (2025) Wan: Open and advanced large-scale video generative models. 2503.20314

Watcharasupat KN, Wu CW, Ding Y, Orife I, Hipple AJ, Williams PA, Kramer S, Lerch A, Wolcott W (2024) A generalized bandsplit neural network for cinematic audio source separation. IEEE Open Journal of Signal Processing pp 73–81

Welsh T, Ashikhmin M, Mueller K (2002) Transferring color to greyscale images. ACM TOG 21:277–280

Wu W, Liu M, Zhu Z, Xia X, Feng H, Wang W, Lin KQ, Shen C, Shou MZ (2025) Moviebench: A hierarchical movie level dataset for long video generation. In: CVPR, pp 28984–28994

Xu J, Guo Z, He J, Hu H, He T, Bai S, Chen K, Wang J, Fan Y, Dang K, Zhang B, Wang X, Chu Y, Lin J (2025) Qwen2.5-omni technical report. 2503.20215

Yang J, Feng J, Huang H (2024a) Emogen: Emotional image content generation with text-to-image diffusion models. In: CVPR, pp 6358–6368

Yang Y, Pan J, Peng Z, Du X, Tao Z, Tang J (2024b) BiSTNet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. IEEE TPAMI 46(8):5612–5624

Zhang B, He M, Liao J, Sander PV, Yuan L, Bermak A, Chen D (2019) Deep exemplar-based video colorization. In: CVPR, pp 8052–8061

Zhang L, Mo S, Zhang Y, Morgado P (2024) Audio-synchronized visual animation. In: ECCV, pp 1–18

Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR, pp 586–595

Zhao Y, Po LM, Yu WY, Rehman YAU, Liu M, Zhang Y, Ou W (2023) VCGAN: Video colorization with hybrid generative adversarial network. IEEE TMM 25:3017–3032