# MovieBench: A Hierarchical Movie Level Dataset for Long Video Generation

## Supplementary Material

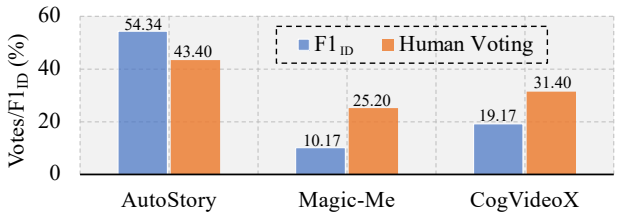



Figure 1. **User Study** $v.s.$ **Automatic Metric.** Automatic metrics and human evaluations show a positive correlation.

Table 1. **Performance for Customized Audio Generation on MovieBench.**

| Dataset | MCD ↓ | WER(%) ↓ | SIM-o ↑ |
|---|---|---|---|
| YourTTS [2] | 8.41 | 0.26 | 0.97 |
| xTTS [3] | 8.31 | 0.28 | 0.98 |
| VALL-E-X [17] | 4.48 | 1.25 | 0.97 |
| F5-TTS [4] | 3.12 | 0.20 | 0.98 |

## 1. User Study

To examine the correlation between automatic evaluation metrics and human judgment, we compare $F1_{ID}$ scores with human voting results across three models: AutoStory, Magic-Me, and CogVideoX, as shown in Figure 1. The results indicate a positive correlation between the two evaluation methods, with models achieving higher $F1_{ID}$ scores also receiving more favorable human votes. Specifically, AutoStory demonstrates the highest alignment, with an $F1_{ID}$ of 54.34% and a human voting score of 43.40%, suggesting strong consistency between automatic and human evaluations. While Magic-Me and CogVideoX also follow this trend, their human evaluation scores exceed their $F1_{ID}$ values by a notable margin, indicating that certain qualitative factors influencing human preference may not be fully captured by automatic metrics.

## 2. Customized Audio Generation

Customized Audio Generation involves creating customized soundtracks for specific characters and emotional cues. We conduct experiments on 6 movies from the test set, splitting the audio of each character into two parts: half as the test set and half as reference audio for evaluation. Following prior works [3, 4], we evaluate performance on a cross-sentence task, where the model synthesizes a reading of a reference text in the style of a given speech prompt.

Table 2. **Quality Evaluation for Portrait Image of Character on Movie Level.** Character bank demonstrates excellent performance in both portrait quality and name relevance.

| Movie | Portrait Quality | Name Relevance |
|---|---|---|
| AS Good As It Gets | 4.56 | 5.00 |
| Clerks | 4.20 | 4.92 |
| Halloween | 4.00 | 4.89 |
| The Hustler | 4.80 | 4.98 |
| Chasing Amy | 4.42 | 4.78 |
| The Help | 4.30 | 5.00 |
| No Reservations | 4.86 | 4.93 |
| An Education | 4.70 | 4.85 |
| Harry Potter and the Chamber of Secrets | 4.73 | 5.00 |
| Seven Pounds | 4.71 | 4.87 |

### 2.1. Metric

Following prior work, three common metrics, namely Word Error Rate (WER), Speaker Similarity (SIM-o), and Mel Cepstral Distortion (MCD), are used to evaluate our dataset. For WER, Whisper-large-v3 [10] is used to transcribe the audio to text, after which word error is calculated at the text level. For SIM-o, a WavLM-large-based speaker verification model [5] is used to extract speaker embeddings, enabling cosine similarity calculation between synthesized and ground truth speech. For MCD, an open-source PyTorch implementation [1] is used to evaluate the similarity between synthesized and real audio. For evaluation, each audio file is converted to a single-channel, 16-bit PCM WAV with a sample rate of 22050 Hz.

### 2.2. Baseline and Analysis

The four audio customization methods—YourTTS [2], xTTS [3], VALL-E-X [17], and F5-TTS [4] were used in MovieBench for evaluation. We performed direct zero-shot testing without any additional training, with F5-TTS achieving the best performance, as shown in Table 1. Notably, each evaluation was conducted individually for each character. However, the real challenge lies in scenes with multi-character interactions, as seen in movies. Generating audio that matches the tone and voice of each character in a way that ensures consistency with the visuals presents a significant challenge, especially in maintaining distinct voices across different audio tracks.

[1] https://github.com/chenqi008/pymcd

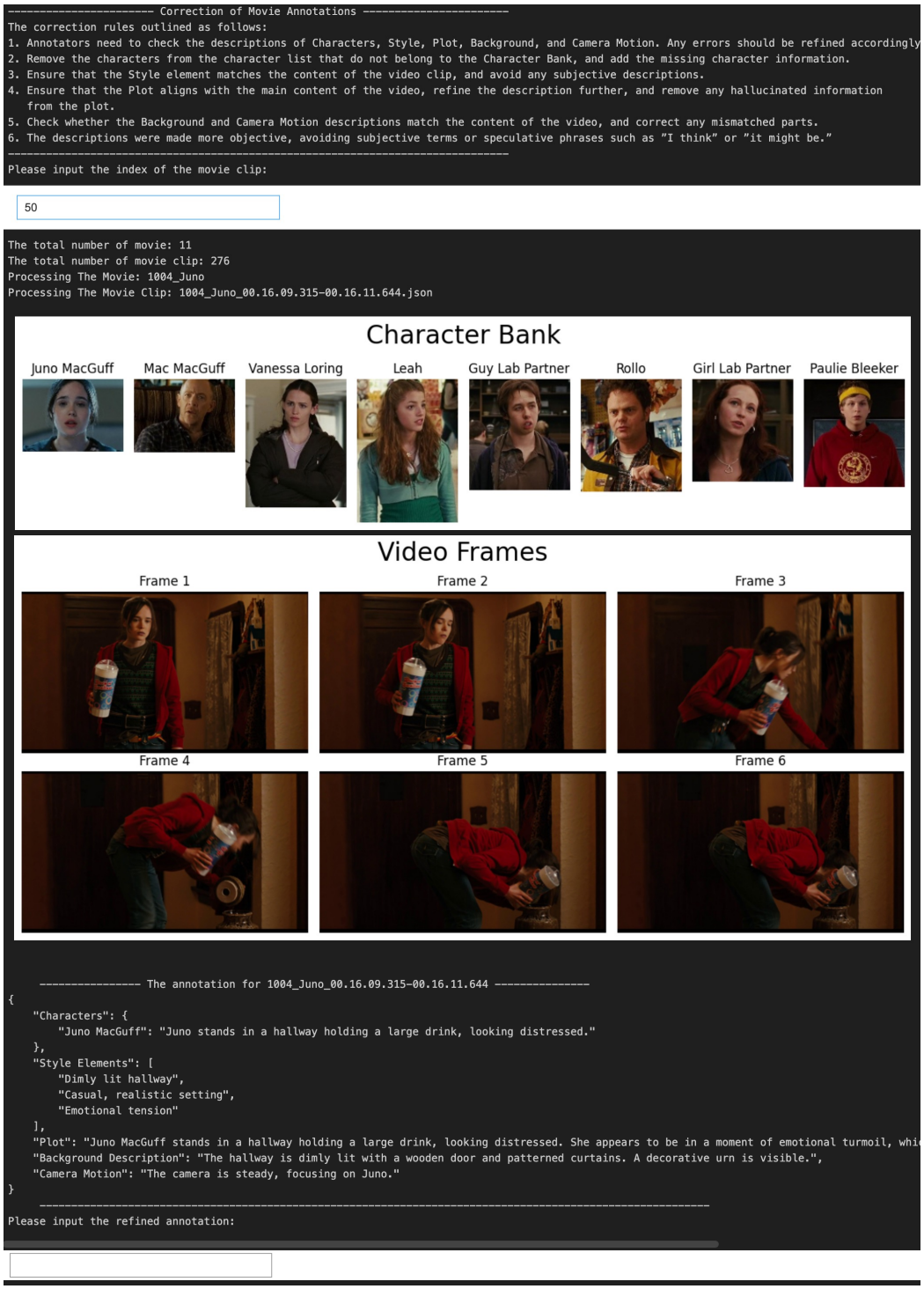



Figure 2. **Manual Correction for Shot-Level Movie Annotations.**

## 3. Quality Evaluation and Correction for Shot-Level Annotation

**Correction for Description-based Annotations.** The main text mentions that we required two annotators to manually correct the shot-level dataset in the test set. The specific correction rules are as follows:

- **Check and Refine Descriptions**: reviewing the descriptions of characters, style, plot, background, and camera motion, correcting any inaccuracies.

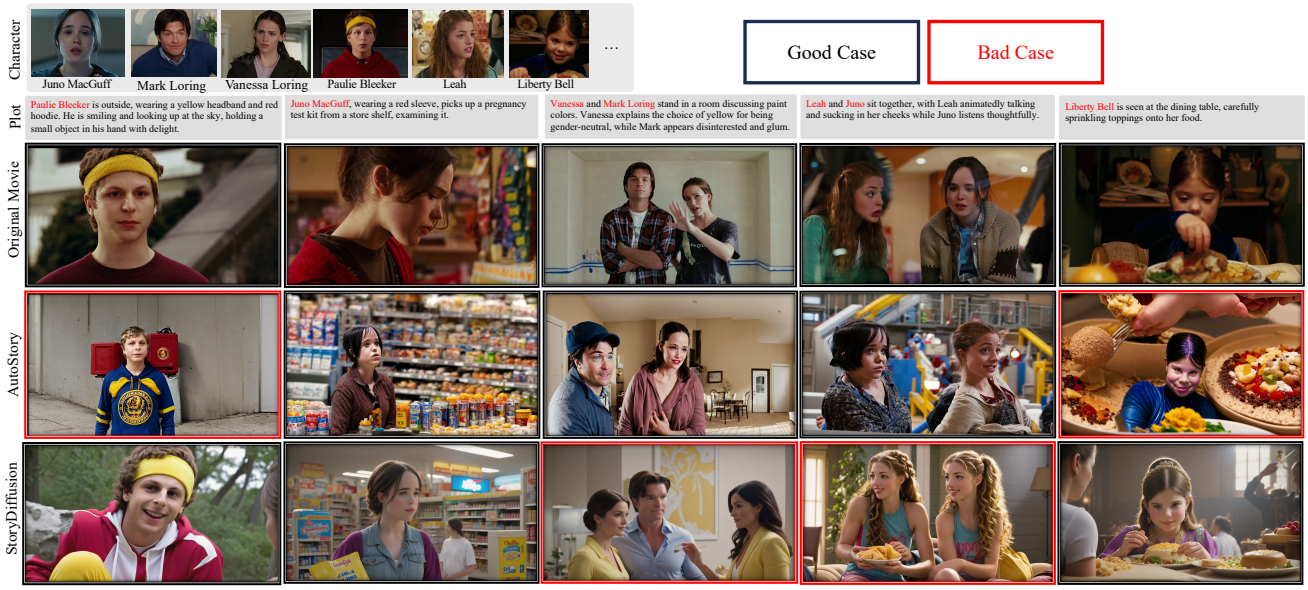
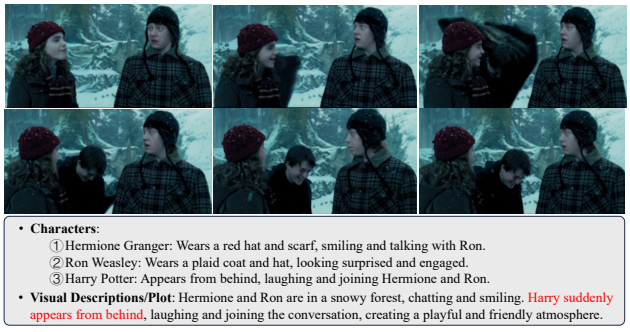
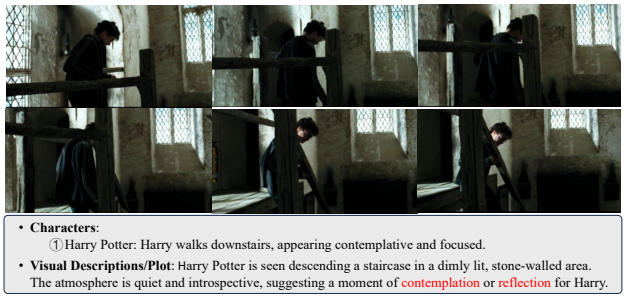



Figure 3. **Visualization Comparison for Movie-Level Keyframe Generation.**



- **Characters**:
  ① Hermione Granger: Wears a red hat and scarf, smiling and talking with Ron.
  ② Ron Weasley: Wears a plaid coat and hat, looking surprised and engaged.
  ③ Harry Potter: Appears from behind, laughing and joining Hermione and Ron.
- **Visual Descriptions/Plot**: Hermione and Ron are in a snowy forest, chatting and smiling. Harry suddenly appears from behind, laughing and joining the conversation, creating a playful and friendly atmosphere.

Figure 4. **Temporal Plot Description.** Shot-level plot descriptions often contain strong temporal information that may not be easily represented by a single key frame.



- **Characters**:
  ① Harry Potter: Harry walks downstairs, appearing contemplative and focused.
- **Visual Descriptions/Plot**: Harry Potter is seen descending a staircase in a dimly lit, stone-walled area. The atmosphere is quiet and introspective, suggesting a moment of contemplation or reflection for Harry.

Figure 5. **Hallucination of Generated Plot.** Descriptions generated by GPT-4-o may still exhibit instances of hallucination.

- **Character Set Adjustments**: Characters not belonging to the Character Bank were removed from the video clip's character set, and any missing characters were added.
- **Style Matching**: Ensure that the style element accurately reflected the video clip's content, avoiding subjective interpretations.
- **Plot Alignment**: The Plot was verified to align with the main content of video, with any hallucinated or irrelevant information removed.
- **Grammatical Accuracy**: Descriptions were refined to ensure grammatical correctness.
- **Objectivity**: The descriptions were made more objective, avoiding subjective terms or speculative phrases such as "I think" or "it might be."

Two annotators were instructed to progressively refine the character set, style, and plot based on the above rules. The refinement interface is shown in Figure 2. The interface provides character photos, names, key frames from the original video, and shot-level annotation details (such as plot, appearing character set, etc.). Annotators use this information to assess the accuracy of the annotations and identify areas for improvement. The refinement process took two annotators approximately one week.

**Quality Evaluation for Shot-level Appearing Character Set.** The main text presents that the character photos in our Character Bank were manually selected by two annotators. After completing the data annotation, we conducted a quality assessment focusing on Portrait Quality and Name-Portrait Relevance. Table 2 shows the relevant experimental results. It can be observed that Portrait-Name Relevance scores significantly higher than Portrait Quality. This is mainly because manually selected images are generally consistent with their names, leaving little room for error in relevance. However, image quality is harder to guarantee, as not all image candidates are of consistently high quality.

Table 3. **Metrics for Evaluation of Model/Dataset.** 'Portrait Quality', 'Portrait-Name Relevance', 'Completeness', and 'Hallucination' are used to assess the quality of MovieBench annotations. Other metrics are primarily used to evaluate model performance.

| Metric | Better | Description |
|---|---|---|
| Portrait Quality | higher | Quality assessment for character portraits, involving human raters scoring the image quality on a scale of 1 to 5, with 5 being the highest score. |
| Portrait-Name Relevance | higher | Portrait-Name relevance score for each character name and portrait pair, with human raters on a scale of 1 to 5, and 5 being the highest score. |
| Completeness | higher | Descriptive completeness score, assessing the extent to which the annotation of textual descriptions (*e.g.,* Plot, Background Description) reflects the completeness of the video content. |
| Hallucination | lower | Fantasy score, assessing the degree of hallucination in textual descriptions of video content (e.g., Plot, Background Description). |
| CLIP Score | higher | The evaluation of semantic alignment between the plot and generated outputs |
| Aesthetic Score(AS) | higher | The evaluation for aesthetic quality of an image by extracting visual features using the CLIP and comparing them with a pre-trained aesthetic model to quantify the score. |
| Frechet Image Distance(FID) | lower | The evaluation for the quality of generated images by comparing the feature distribution of features between real and generated images. |
| Inception Score (IS) | higher | The evaluation for the quality and diversity of generated images by Inception network. |
| False Postive(FP) | lower | The total number of false positives. Formula: $\mathrm{FP} = \sum_{i=1}^{n} \left\lvert \mathcal{C}_i^{\mathrm{pred}} \setminus \mathcal{C}_i^{\mathrm{gt}} \right\rvert$. |
| False Negative(FN) | lower | The total number of false negative. Formula: $\mathrm{FN} = \sum_{i=1}^{n} \left\lvert \mathcal{C}_i^{\mathrm{gt}} \setminus \mathcal{C}_i^{\mathrm{pred}} \right\rvert$. |
| True Postive(TP) | higher | The total number of true positives. Formula: $\mathrm{TP} = \sum_{i=1}^{n} \left\lvert \mathcal{C}_i^{\mathrm{gt}} \cap \mathcal{C}_i^{\mathrm{pred}} \right\rvert$. |
| $\mathrm{Recall}_{\mathrm{ID}}$ | higher | Ratio of correct detections&recognitions to total number of GTs. Formula: $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$ |
| $\mathrm{Precision}_{\mathrm{ID}}$ | higher | Ratio of correct detections&recognitions to total number of predicted detections&recognitions. Formula: $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$ |
| $\mathrm{F1\ Score}_{\mathrm{ID}}$ | higher | $\mathrm{F1\ Score}_{\mathrm{ID}}$ [11]. The ratio of correctly identified detections&recognitions over the average number of ground-truth and computed detections&recognitions. Formula: $\frac{\mathrm{Recall}_{\mathrm{ID}} \times \mathrm{Precision}_{\mathrm{ID}} \times 2}{\mathrm{Recall}_{\mathrm{ID}} + \mathrm{Precision}_{\mathrm{ID}}}$ |
| Subject Consistency | higher | DINO [1] is used to assess whether the appearance remains consistent throughout the entire video. |
| Background Consistency | higher | CLIP feature similarity [9] is used to evaluate the temporal consistency of the background scenes. |
| Motion Smoothness | higher | Video frame interpolation model [7] is used to evaluate the smoothness of generated motions. |
| Dynamic Degree | higher | Optical flow estimation [12] is used to evaluate the degree of dynamics in synthesized videos. |

## 4. Possible Directions? Single-Stage or Two-Stage

Movie/long video generation is typically not done in one go; instead, it is divided into multiple shot clips for individual generation. Currently, there are two main approaches for this task: one-stage and two-stage methods.

**One Stage.** Currently, there are no fully realized one-stage solutions for this task. Most open-source one-stage models [16, 19] focus on text-to-video generation, lacking the ability to maintain character consistency and connect storylines across different video clips. DreamVideo [14] and Magic-Me [8], two commonly used customizable video generation models, are utilized in our paper. The typical workflow involves first creating a script with character-specific details for each shot, generating corresponding video clips for each shot individually, and then stitching these clips together to produce a cohesive long-form video.

**Two Stages.** Directly generating long-form videos is highly challenging. Therefore, the two-stage strategy has become a more practical solution: 1) Firstly, Key frames/Storyboard generation models [13, 15, 18, 20] can be used to generate the key frame for every shot-level video. Figure 3 provides additional visualizations of both successful and challenging cases for AutoStory [13] and StoryDiffusion [20]. It can be observed that AutoStory [13] excels in maintaining consistency across multiple characters but struggles with certain background compatibility. On the other hand, StoryDiffusion [20] performs well in generating natural interactions between characters and backgrounds but has difficulty maintaining consistency across multiple characters. 2) With key frames, image-conditioned video generation models (*e.g.,* SVD), are employed to expand the key frames into full video clips. Finally, the various video clips are stitched together to form a coherent sequence. While this method addresses some issues of video continuity and narrative progression, it still faces difficulties with maintaining a smooth flow across clips and ensur-

ing consistent character representation throughout the film. However, for certain shots with strong temporal dependencies, it is challenging to rely solely on keyframes for representation. Figure 4 shows an example where generating only a single keyframe is clearly insufficient to capture the sequence of Harry's appearance.

## 5. Metric Formulation

As shown in Table 3, we summarize and formulate the evaluation metrics relevant to the tasks discussed in this paper. 'Portrait Quality' and 'Portrait-Name Relevance' assess the accuracy of the Character Bank annotations, specifically evaluating the precision of manual image selection and labeling. 'Completeness' and 'Hallucination' measure the accuracy of description-based annotations (*e.g.,* plot and background descriptions), focusing on the completeness of details and hallucinations from VLM descriptions. The CLIP Score, Aesthetic Score, Frechet Image Distance, and Inception Score evaluate the quality of generated images/videos and their alignment with text descriptions. Additionally, this paper introduces new metrics—$\text{Precision}_{\text{ID}}$, $\text{Recall}_{\text{ID}}$, and $\text{F1}_{\text{ID}}$—to assess character consistency. 'Subject Consistency', 'Background Consistency', 'Motion Smoothness', and 'Dynamic Degree' are recently proposed metrics from VBen [6], aimed at evaluating various aspects of generated video.

## 6. Limitation

**Hallucination of Plot from GPT4-o.** Although GPT-4-o demonstrates high accuracy and rarely makes errors, its generated plot descriptions can still present issues, such as hallucination. Figure 5 provides a clear example: in this video, Harry walks downstairs, yet there is no evidence to conclude that Harry is engaged in contemplation or reflection. However, the summary of GPT-4-o confidently suggests this, introducing a potential misinterpretation. Such hallucinations can reduce data reliability, misleading model training and potentially causing unstable convergence when using this data.

## References


[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4

[2] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022. 1

[3] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*, 2024. 1

[4] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024. 1

[5] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE, 2022. 1

[6] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 5

[7] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 4

[8] Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*, 2024. 4

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 1

[11] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Workshops of ECCV*, pages 17–35, 2016. 4

[12] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4

[13] Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human effort. *arXiv preprint arXiv:2311.11243*, 2023. 4

[14] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 4

[15] Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Paragraph-to-image generation with information-enriched diffusion model. *arXiv preprint arXiv:2311.14284*, 2023. 4

[16] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 4

[17] Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*, 2023. 1

[18] Canyu Zhao, Mingyu Liu, Wen Wang, Jianlong Yuan, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024. 4

[19] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *https://github.com/hpcaitech/Open-Sora*, 2024. 4

[20] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 4