Supplementary Material for How Far are AI-generated Videos from Simulating the 3D Visual World: A Learned 3D Evaluation Approach

Chirui Chang[1] Jiahui Liu[1] Zhengzhe Liu[3] Xiaoyang Lyu[1] Yi-Hua Huang[1]
Xin Tao[2] Pengfei Wan[2] Di Zhang[2] Xiaojuan Qi[1*]
[1]The University of Hong Kong [2]Kling Team, Kuaishou Technology [3]Lingnan University

Contents

1. Applications for L3DE

In this section, we mainly demonstrate two downstream applications for our proposed L3DE: 1.) **Fake video detection** and 2.) **Generative video refinement**.

*Corresponding author.

1.1. Fake Video Detection

L3DE is designed to evaluate the 3D real world simulation capabilities of AI-generated videos, enabling it to distinguish low-quality AI-generated videos from real-world ones. Motivated by this capability, we conduct fake video detection experiments to assess how well L3DE performs on this task. This can be achieved by setting a threshold on the L3DE score, allowing us to classify videos as real or fake based on their ability to simulate the real 3D visual world.

Specifically, we use fake videos from our 3D reconstruction verification set together with those from [29] and an equal number of unseen in-the-wild real videos from Pexels [15] to build a fake video detection benchmark. As there is currently no open-source general fake video detector to the best of our knowledge, we adapt fake image detection methods for videos. To do this, we compare L3DE fusion scores with existing fake image detection methods [21, 26, 27] by averaging frame-wise predictions to produce a final prediction for each video. The results are presented in Table 1.

The results indicate that L3DE scores exhibit strong performance in fake video detection, even though L3DE is not specifically designed for this task. Across videos generated by different models, L3DE scores generally achieve higher accuracy than image-based fake detection methods. These results suggest that most synthesized videos still have significant gaps in 3D simulation capabilities. In conclusion, L3DE scores demonstrate strong performance in fake video detection, despite not being specifically designed for this task.

1.2. AI-Generated Video Refinement

In current generative videos with regional artifacts, such artifacts often necessitate discarding the entire video if it does not meet the criteria for downstream tasks. However, with L3DE's ability to identify and localize artifact regions, we can achieve AI-generated video refinement by removing these artifacts in a 3D-consistent manner.

Specifically, we utilize L3DE activation values to local-

Method	Input	MiniMax	Kling 1.5	Runway-Gen3	Luma Dream Machine	CogVideoX	Vidu	Sora	Average
CNNDetection [26]	Image	49.92	50.02	50.00	50.45	50.07	50.00	49.91	50.05
DIRE [27]	Image	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
NPR [21]	Image	60.19	67.91	64.99	54.06	35.79	36.04	**60.82**	54.25
L3DE	Video	**66.51**	**82.52**	**72.19**	**83.38**	**76.73**	**70.01**	56.31	**73.14**

Table 1. Fake video detection performance of L3DE scores and image-based approaches. The reported metric is accuracy with all values presented as percentages.

Figure 1. A qualitative result of generative video refinement. In this example, bounding boxes highlight the regions where artifacts are detected in the original video. After refinement, these artifacts are successfully removed across all frames of the video.

ize the regions of artifacts in the keyframes of the downsampled clip. We then employ SAM-2 [17] to refine and propagate the masks across the entire original generative video. Inspired by [12], we implement a 3D-GS-based multi-view consistent inpainting iteratively using LaMa [20].

We demonstrate our results for video refinement in Figure 1. Based on our findings, the artifact-detection capability of L3DE can effectively guide the post-processing step of video refinement, helping to remove artifacts in generative videos.

2. Data Processing

In this section, we detail our data processing procedures, including raw video processing and video feature extraction.

2.1. Raw Video Processing

We follow the approach introduced in [2] for raw video processing. First, we collect an open-world, in-the-wild long video dataset from Pexels [15], covering a wide range of content with varying aspect ratios, resolutions, and frame rates. Figure 2 showcases the diversity of our dataset. To avoid biases caused by cuts and fades, we apply PySceneDetect [4] to the long videos.

Next, to prepare paired data, we slice these videos into equal-length clips of 4 seconds. For videos that do not match the 16:9 aspect ratio, we apply a center crop and resize them to a resolution of 1024×576 with 25 frames. Additionally, we use the first frame of these processed video clips as image prompts for stable video diffusion [2] to generate paired synthetic samples. Moreover, we provide visualizations of randomly sampled paired videos in Figure 7. As introduced in the main paper, we sample 160,000 paired videos for training the L3DE models.

2.2. Video Feature Extraction

We extract video features using different foundation models following their official implementation: For **appearance features**, we extract frame-wise features from the DINOv2 ViT-G model [13]. For **motion features**, we input adjacent frames into RAFT [25] to obtain the optical flow sequence of the entire video. For **geometry features**, we extract per-frame metric depth using the UniDepth v2 ViT-S model [16].

To align the inputs from different proxies, we use the metric depth and DINOv2 features of the first 24 frames of the video clips, since the optical flow maps are calculated based on adjacent frames. This strategy ensures that L3DE simultaneously captures different modalities of 3D proxies.

2.3. The Impact of Data Diversity

Our goal is to construct a diverse training dataset to improve the robustness of L3DE. Diversity in training data plays a crucial role in enhancing generalization by exposing the model to a broad range of real-world and challenging scenarios. While data diversity can encompass various factors—such as object motion, scene complexity, and environmental variations—we focus on two key aspects in our analysis: (1) the role of object motion (*static vs. mixed static-dynamic scenes*) and (2) the impact of scene diversity (*indoor-only vs. mixed indoor-outdoor scenes*). These controlled experiments illustrate how different types of training data contribute to model performance, reinforcing the importance of a diverse dataset.

Effect of Object Motion To assess the impact of object motion, we train two models using different datasets: one on 10,000 static-scene real and synthetic videos (1:1 ratio), and the other on an equally sized dataset that includes both static and dynamic scenes. Both models are evaluated on a 2,000-sample test set, which consists of an equal number

Figure 2. Our collected real-world, in-the-wild videos encompass a wide range of visual content, from indoor to outdoor scenes, including people, animals, landscapes, food, and more.

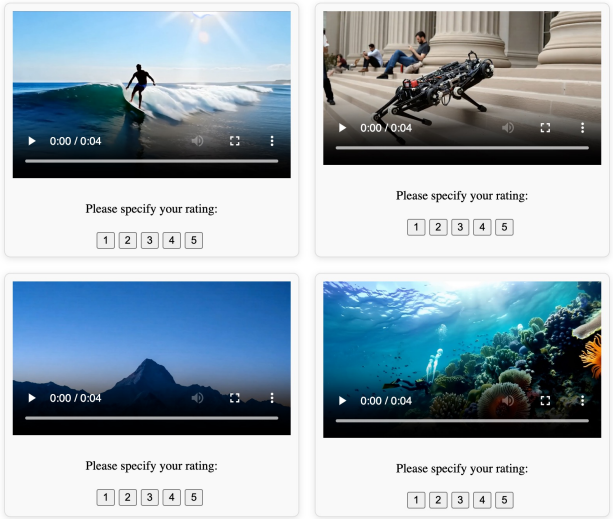


Figure 3. Interface for human rating annotation. Users can provide a rating in the scoring section below after watching the video above.

of real and synthetic videos featuring mixed motion. The synthetic videos are generated using Kling [6]. As shown in Table 2, the model trained solely on static scenes underperforms compared to the one trained with motion diversity (**69.55** *vs.* **77.70**), confirming that incorporating object motion in training significantly improves generalization.

Effect of Scene Diversity To analyze the effect of scene diversity, we train one model using 10,000 indoor real and synthetic videos (1:1 ratio), and another using 10,000 mixed indoor-outdoor videos. Both models are evaluated on a 2,000-sample mixed indoor-outdoor test set, maintaining a 1:1 ratio of real to synthetic videos. As seen in Table 2, the model trained only on indoor data exhibits lower accuracy (**67.60** *vs.* **76.55**), demonstrating that exposure to a wider variety of environments enhances model robustness.

Experiment	Training Data	Test Data	Accuracy
Static-only	Static-scene videos	Mixed-motion videos	69.55
Mixed-motion	Static + dynamic videos	Mixed-motion videos	77.70
Indoor-only	Indoor videos	Mixed indoor-outdoor	67.60
Indoor + Outdoor	Indoor + outdoor videos	Mixed indoor-outdoor	76.55

Table 2. Impact of Data Diversity on Model Performance. Training on diverse data significantly improves accuracy.

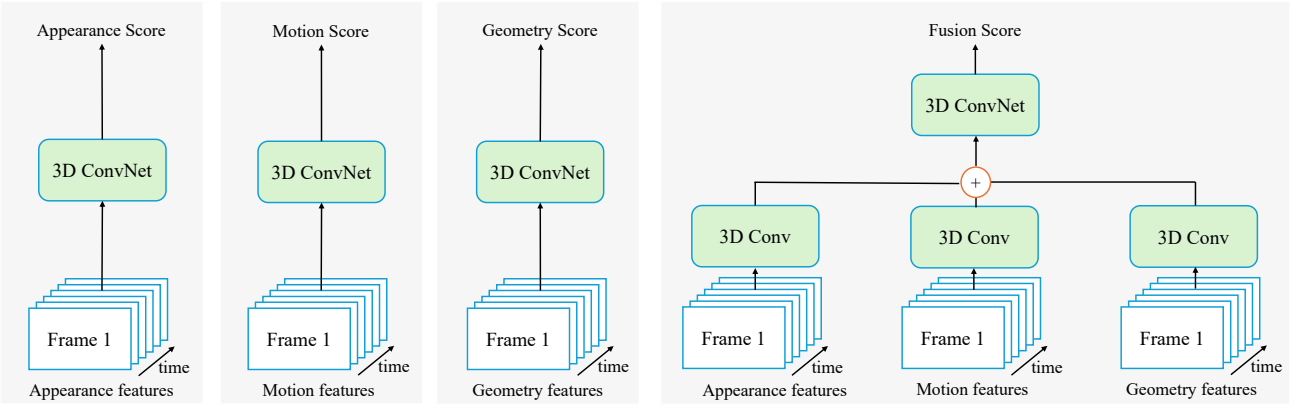
3. L3DE Architecture

In this section, we provide details about the L3DE architecture, including both the single-proxy and fusion versions.

3.1. Single-proxy Network

First, we illustrate our design of the single-proxy version of the L3DE network in Figure 4(a). Given a single aspect proxy, such as frame-wise appearance features of a video as input, the 3D ConvNet produces a corresponding confidence score for the video. Specifically, the single-proxy L3DE is a single-branch 3D convolutional network focusing on capturing spatiotemporal features from a single input modality.

The network begins with sequential 3D convolutional layers that progressively encode high-level representations of the input through non-linear activations and feature refinement. After the convolutional stages, the feature map is flattened into a 1D vector, which is passed through a fully connected layer to reduce dimensionality. The final prediction is performed using another fully connected layer with a sigmoid activation, producing a confidence score.



(a) Single-proxy Network (b) Fusion Network

Figure 4. The design of both single-proxy network shown in part (a), and fusion network illustrated in part (b).

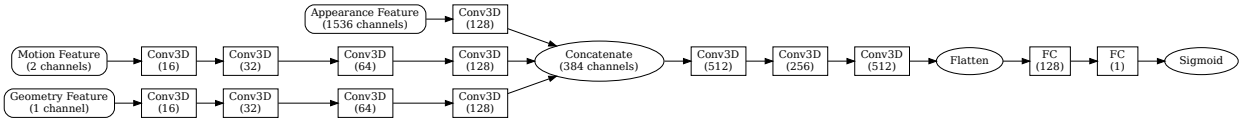


Figure 5. The detailed architecture of fusion network, a 3D convolutional neural network designed for multimodal feature fusion. The network takes three input streams: Appearance Features (1536 channels), Motion Features (2 channels), and Geometry Features (1 channel). Each stream undergoes a series of 3D convolutional layers with ReLU activations before being concatenated into a 384-channel fused representation. The concatenated features are further processed through additional convolutional layers, followed by flattening and fully connected layers.

3.2. Fusion Network

Next, we illustrate the design of the fusion version of L3DE in Figure 4(b). In detail, the fusion network is a 3D ConvNet integrating appearance, motion, and geometric features through a multi-branch architecture. Each input modality—appearance features, motion features, and geometric features—is processed separately using specialized 3D convolutional layers, which hierarchically encode spatiotemporal information through non-linear activations and down-sampling via strided convolutions.

The outputs of the three branches are concatenated along the channel dimension, enabling the model to jointly leverage complementary features from all modalities, in line with prior efforts [7]. The fused representation undergoes further refinement through additional convolutional layers that capture high-level correlations across the integrated features. The network concludes with two fully connected layers and a final sigmoid activation for score prediction.

We also provide the architecture details of the fusion network in Figure 5. Note that each single-branch model adopts the same architecture as its corresponding branch in the fusion network.

3.3. Ablation Study

In this section, we conduct an ablation study to analyze the impact of contrastive loss and feature fusion strategies on distinguishing real and synthetic videos in L3DE. As shown in Table 3, both contrastive loss and fusion strategies play a crucial role in model performance. We compare two feature fusion methods: (1) Element-wise Addition (Add), where features from different sources are summed component-wise; and (2) Feature Concatenation (Concat), where features are stacked along the channel dimension to retain independent information. First, comparing the Add and Concat fusion strategies, we observe that Concat consistently outperforms Add. Without contrastive loss, Concat achieves 68.77%, surpassing Add (66.01%), indicating that concatenation preserves richer feature representations. When contrastive loss is introduced, performance improves significantly in both fusion strategies (+3.25% for Add and +4.37% for Concat), confirming that the loss function enhances feature discrimination. Our L3DE setting (Concat + Contrastive Loss) achieves the highest accuracy (73.14%), as highlighted in Table 3. These results demonstrate that contrastive loss effectively boosts performance by improving the feature separation between real and synthetic videos. Additionally, the superior performance of Concat over Add

Fusion Strategy	Contrastive Loss	Accuracy (%)
Element-wise Addition	✗	66.01
	✓	69.26
Feature Concatenation	✗	68.77
	✓	73.14

Table 3. Ablation study on contrastive loss and feature fusion strategies (Concat vs. Add). The highlighted row represents our setting and results.

suggests that maintaining richer feature representations is beneficial for this task. Thus, we adopt the Concat + Contrastive Loss setting as the default configuration in L3DE.

3.4. Implementation Details

We implement our 3D ConvNet using PyTorch [14]. The models are trained with a learning rate of 1e-4 and a batch size of 20. For video generation with SVD-XT [2] and training of L3DE models, we utilize NVIDIA A100 GPUs. Additionally, NVIDIA 4090 GPUs are used for conducting 3D reconstruction experiments. We follow the official implementation for Grad-CAM [19] visualization.

4. User Study

In this section, we provide detailed descriptions of the user studies mentioned in the main paper.

4.1. User Study for Video Ratings

We conduct a user study involving 15 volunteers, who provide a total of 4,500 annotations on 300 randomly selected generative videos from our dataset. Annotators are recruited via our internal platform. Participants are aged between 20 and 40, come from diverse educational backgrounds, and do not possess specialized computer vision knowledge, ensuring broad representativeness.

To ensure annotation quality, volunteers complete a prelabeling task following previous work [5, 9] and only those showing consistent and accurate judgments qualify for the main study.

Qualified participants receive clear scoring guidelines to ensure consistency. The guidelines explicitly instruct them to evaluate the realism of videos based on 3D visual coherence in appearance, motion, and geometry, rather than semantic content or other unrelated factors. Participants rate each video's realism on a 1 to 5 scale, with clear definitions provided:

- Score **1**: Videos exhibit obvious visual artifacts, severe geometry deformation, unnatural motion, or evident synthetic features.
- Score **2**: Videos have significant artifacts clearly distinguishable from real ones, significantly impacting realism.
- Score **3**: Videos contain noticeable but non-disruptive artifacts, moderately realistic overall.
- Score **4**: Videos closely resemble real-world footage with minor and infrequent artifacts.
- Score **5**: Videos are indistinguishable from real-world footage, exhibiting minimal to no noticeable artifacts or inconsistencies.

Participants rate all 300 videos through our internal annotation interface (Figure 3). After collecting annotations, we then compute the Spearman correlation coefficients between these human ratings and the L3DE scores across different modalities. Moreover, to further verify L3DE's alignment with human perception, we conduct additional human evaluations on the subset "Generated Videos for In-the-wild Scenes." These evaluations comprehensively validate our method's performance on the same dataset, facilitating comparison with the reconstruction-based validation.

4.2. User Study for Grad-CAM Region Ratings

To evaluate the interpretability and effectiveness of the localized regions identified by L3DE (via Grad-CAM), we conduct an additional user study involving 10 qualified volunteers. Participants review 40 randomly selected generative videos from our dataset, each presented alongside visualizations highlighting artifact regions.

Among these 40 videos, for each modality (appearance, motion, and geometry), we randomly select 10 diverse videos. Additionally, we insert 10 videos with randomly generated Grad-CAM highlights serving as a control group to mitigate potential participant biases toward highlighted regions.

Participants view each video along with the corresponding visualization and rate the relevance of highlighted regions to the observed visual artifacts using the following scale:

- Score **1**: Highlighted regions are irrelevant or poorly match the perceived artifacts.
- Score **2**: Highlighted regions slightly match perceived artifacts but miss major inconsistencies.
- Score **3**: Highlighted regions partially match perceived artifacts.
- Score **4**: Highlighted regions generally reflect perceived artifacts with minor discrepancies.
- Score **5**: Highlighted regions accurately reflect major perceived artifacts.

Participants are unaware that 10 of the provided visualizations are randomly highlighted (random baseline) to minimize bias. We specifically evaluate these procedures on the subset "Generated Videos for In-the-wild Scenes" to verify L3DE's effectiveness in localizing artifacts under realistic conditions. Average scores across participants quantify human plausibility, as presented in the main paper. Additionally, 10 participants manually annotate regions they

	Visual Quality	Motion Quality	Temporal Consistency
EvalCrafter	55.4	**45.0**	56.7
Ours	**67.0**	43.6	**58.0**

Table 4. Correlation between L3DE scores and human annotations from the ECTV dataset. Appearance, motion, and fusion scores correspond to visual quality, motion quality, and temporal consistency, respectively.

perceive as unrealistic in 30 unseen videos. This serves as a further validation step for Grad-CAM localization, allowing us to quantitatively evaluate pixel-level correlations between human annotations and Grad-CAM highlighted regions.

5. More Experiments for L3DE

5.1. Additional Comparison with Baselines

To further assess the generalizability of L3DE, we compare its performance against EvalCrafter [9] using correlation metrics on the EvalCrafter Text-to-Video (ECTV) Dataset. EvalCrafter evaluates video quality across multiple dimensions, among which visual quality, motion quality, and temporal consistency are the most relevant to L3DE's evaluation criteria. As shown in Table 4, L3DE achieves a higher correlation with human annotations in terms of visual quality (**+11.6%**) and temporal consistency (**+1.3%**), demonstrating its strong ability to assess both appearance and temporal coherence. L3DE achieves a comparable correlation in motion quality (43.6% vs. 45.0%), indicating its effectiveness in capturing motion fidelity. These results suggest that L3DE provides a more comprehensive and robust evaluation, particularly in aspects that contribute to overall perceptual quality.

5.2. Comparison with External Human Preference Benchmark

To further validate the generalizability and robustness of our L3DE results, we compare the ranking of generative video models obtained by L3DE against the publicly available large-scale human preference leaderboard from Video Arena [1], which aggregates extensive user votes. Although the datasets and specific videos differ, the model rankings obtained by L3DE closely align with those in the Video Arena leaderboard as shown in Table 5. Notably, both assessments consistently identify similar high-performing and lower-performing generative models. This alignment further confirms that L3DE effectively captures general human perceptual judgments regarding video realism, strengthening the validity of our evaluation framework.

5.3. More Qualitative Results

In this section, we provide additional qualitative results of L3DE for reference. Specifically, we illustrate the Grad-CAM results and analyses of L3DE's appearance, motion, and geometry components in Figures 8, 9, and 10, respectively. We further include comprehensive qualitative examples from the Fusion Grad-CAM analysis, highlighting complex artifacts captured by integrating multiple cues. Figure 6 demonstrates cases involving physically implausible interactions, such as abnormal behaviors of liquids interacting with glass and tables, as well as incorrect human scaling. These examples emphasize the enhanced capability of the fusion model to detect high-level inconsistencies beyond individual appearance, motion, or geometry assessments..

Generative Model	L3DE Score ↑	Arena ELO ↑	Ranking (Ours / Arena)
Sora [3]	0.8895	1077	1 / 1
MiniMax [11]	0.7932	1067	2 / 2
Kling 1.5 [6]	0.7518	1058	3 / 3
Runway-Gen3 [18]	0.7162	1017	4 / 4
CogVideoX [28]	0.6104	811	5 / 6
Luma [10]	0.5062	997	6 / 5

Table 5. Comparison of generative model rankings obtained by L3DE and human preference judgments from Video Arena [1]. Rankings only consider models appearing in both our 3D visual simulation benchmark and the Video Arena leaderboard. Although datasets differ and there are minor discrepancies in model versions due to rapid iterations in commercial models, the consistent ranking demonstrates L3DE's alignment with general human perceptual judgments.

Figure 6. Both clips are from Sora [3]. The first row highlights implausible liquid-glass-table interaction (Score: 0.7256), while the second reveals an incorrect human scale (Score: 0.0023).

6. Clarification on Research Scope

L3DE focuses explicitly on **3D visual** coherence, specifically assessing appearance, motion, and geometry, as these dimensions are fundamental prerequisites for realistic simulations. It is important to clarify that our method does not comprehensively evaluate all the aspects related to world simulation such as complex interactions (e.g., accurate physics-based interactions, fluid dynamics). Thus, L3DE provides a targeted assessment specifically related to foundational 3D visual coherence, forming a necessary ba-

sis for further advancements towards comprehensive world simulation.

7. Limitations

Although our study takes a very first step to assess the 3D simulation capabilities of AI-generated videos, several challenges remain: 1.) Dataset Size and Diversity: Currently, we use 160000 video clips to train L3DE model. However, the real-world patterns are very complicated and training on more videos will provide a more general and robust evaluation tool. 2.) Limited Generative Video Length: Due to the constraints of current open-source generative video models, which produce relatively short videos, it is challenging to evaluate long-range coherence and object permanence of the future generative videos. To address these limitations, we plan to continually update L3DE to adapt to the generative videos in the future, and further explore its potential in broader data-centric research [8, 22–24, 30, 31].

References

[1] Artificial Analysis. Video Arena Leaderboard. https://artificialanalysis.ai/text-to-video/arena?tab=Leaderboard. 6

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 5

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 6

[4] Brandon Castellano. Pyscenedetect: Video cut detection and analysis tool. https://github.com/Breakthrough/PySceneDetect. 2

[5] 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. *arXiv preprint arXiv:2311.17982*, 2023. 5

[6] Kuaishou. Kling ai. *https://klingai.kuaishou.com/*, 2024.06. 3, 6

[7] Jiahui Liu, Chirui Chang, Jianhui Liu, Xiaoyang Wu, Lan Ma, and Xiaojuan Qi. Mars3d: A plug-and-play motion-aware model for semantic segmentation on multi-scan 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9372–9381, 2023. 4

[8] Jiahui Liu, Xin Wen, Shizhen Zhao, Yingxian Chen, and Xiaojuan Qi. Can ood object detectors learn from foundation models? In *European Conference on Computer Vision*, pages 213–231. Springer, 2024. 7

[9] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 5, 6

[10] LumaLabs. Dream machine. *https://lumalabs.ai/dream-machine*, 2024.06. 6

[11] MiniMax. Hailuo ai. *https://hailuoai.com/video*, 2024.09. 6

[12] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. 2

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[14] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[15] Pexels. https://www.pexels.com/, 2023. 1, 2

[16] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. *arXiv preprint arXiv:2403.18913*, 2024. 2

[17] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[18] Runway. Gen-3. *https://runwayml.com/*, 2024.06. 6

[19] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016. 5

[20] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2

[21] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 1, 2

[22] Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. In *Neural Information Processing Systems (NeurIPS)*, 2023. 7

[23] Haoru Tan, Sitong Wu, Zhuotao Tian, Yukang Chen, Xiaojuan Qi, and Jiaya Jia. Saco loss: Sample-wise affinity

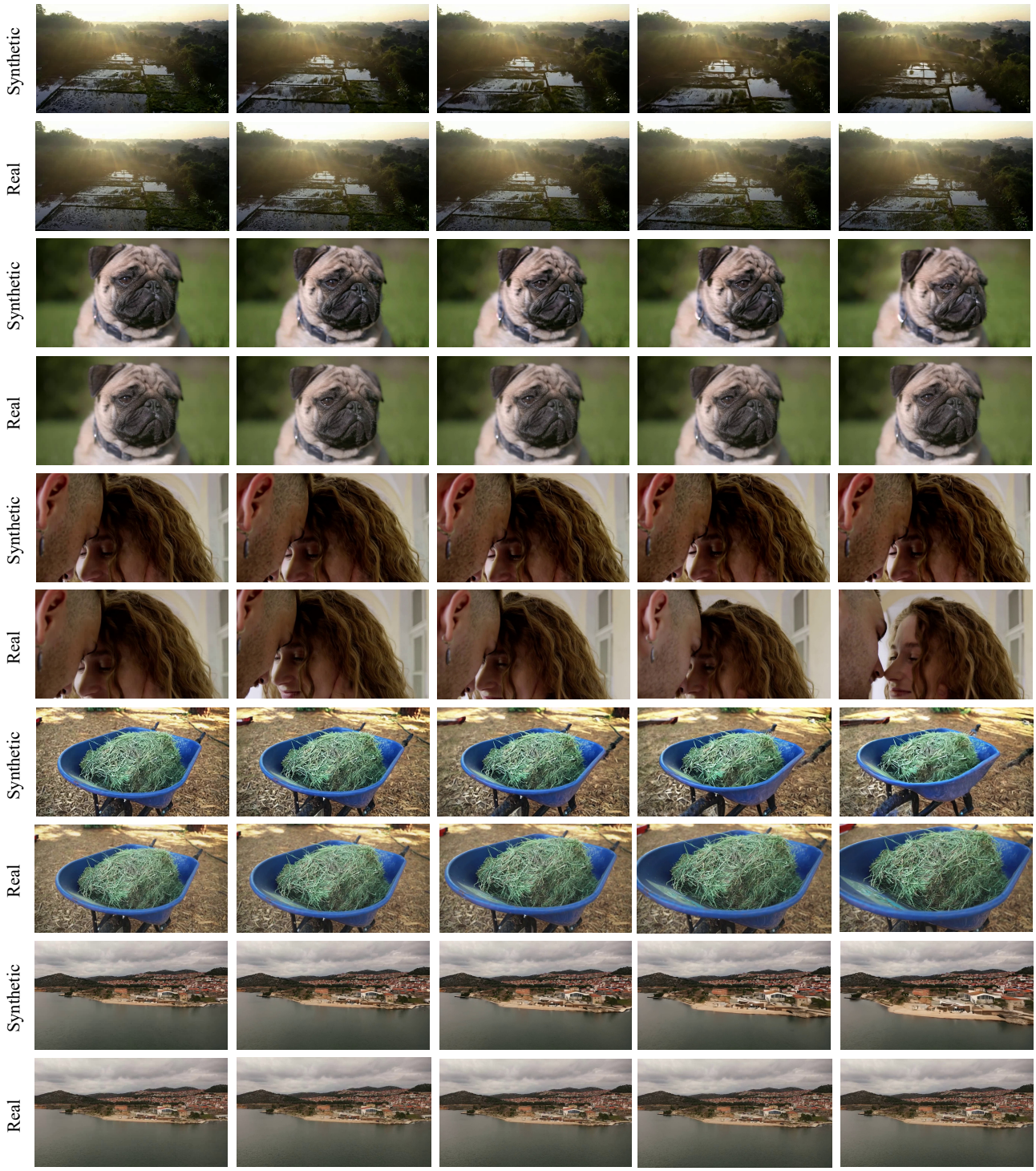


Figure 7. Visualization of randomly sampled paired videos. The images on the left are the image prompts for the generated videos and their first frame. The remaining images show the subsequent frames of the real videos and the generated videos.

consistency for vision-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] Haoru Tan, Sitong Wu, Wei Huang, Shizhen Zhao, and Xiaojuan Qi. Data pruning by information maximization. In *International Conference on Learning Representations (ICLR)*, 2025. 7

[25] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field

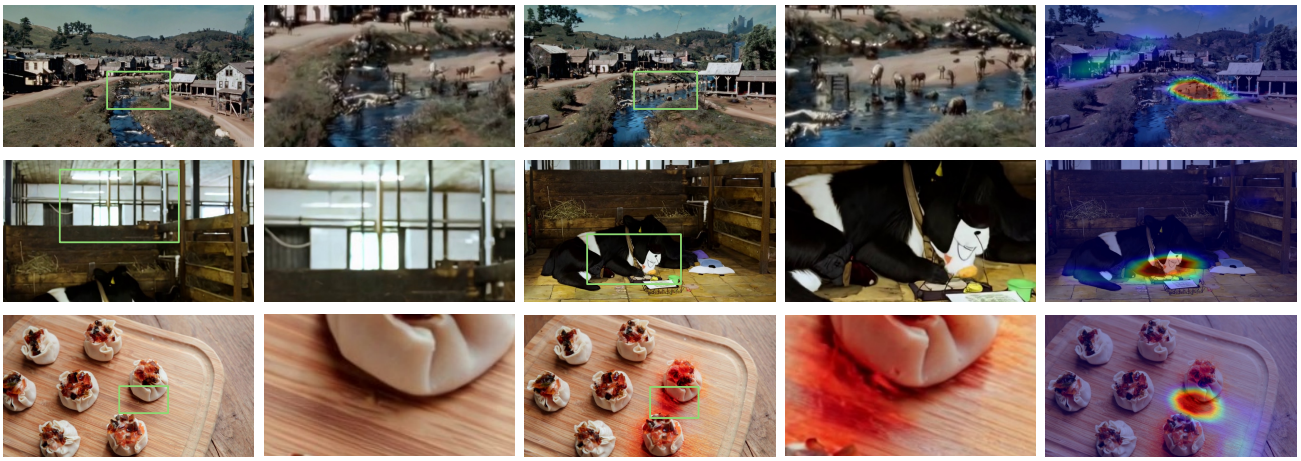


Figure 8. **Appearance Grad-CAM results of L3DE.** For the first video, appearance Grad-CAM detect unstable scene appearances in the connecting regions between the two scenes, such as objects suddenly appearing or disappearing. For the second video, Appearance Grad-CAM detect regions with inconsistent scene appearance styles. Specifically, the first half of the video depicts a realistic cowshed, but it generates cartoon-style cows inside. For the third video, Appearance Grad-CAM detect a sudden change in the texture of the wooden board and food in the video. More specifically, the color of the wooden board and the food in the marked area change significantly between consecutive frames.

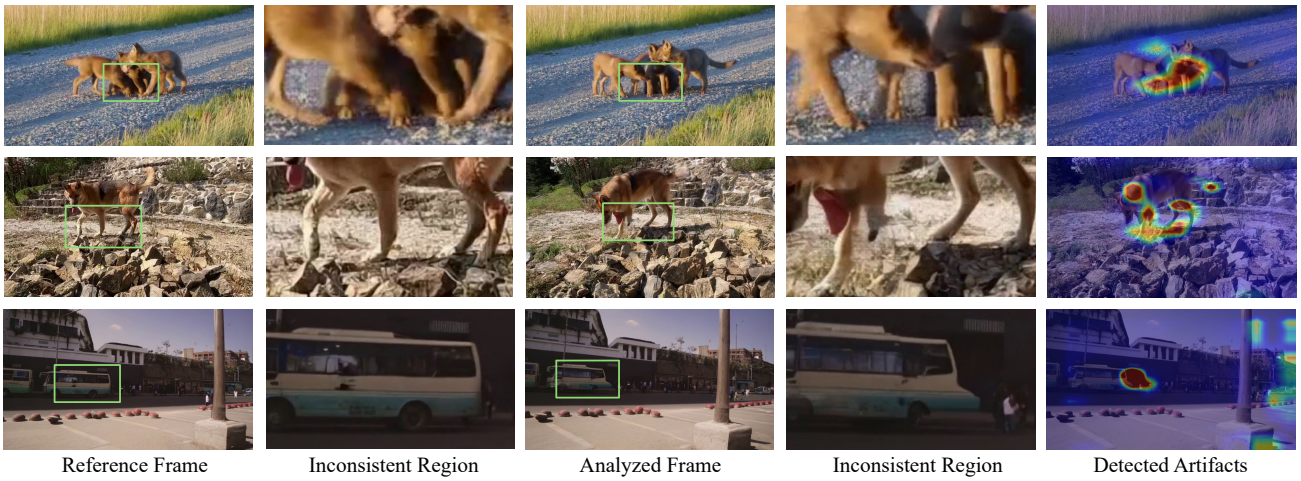


Figure 9. **Motion Grad-CAM results of L3DE.** For the first video, Motion Grad-CAM detect unnatural motion patterns of the wolves. In the video, the movement of the wolves in the marked area is accompanied by an appearance-disappearance phenomenon, which does not conform to real-world motion patterns. For the second video, Grad-CAM detect regions where the wolf exhibits unnatural motion. Specifically, a wolf that appears with normal four legs in the reference frame experiences sudden disappearance of its legs when moving in subsequent frames. Such motion patterns are inconsistent with real-world ones. For the third video, Grad-CAM detect a sudden unnatural 'compression' motion in the bus, which remain stationary in the first half of the video. This does not conform to real-world motion laws.

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. 2

[26] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 1, 2

[27] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 1, 2

[28] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint*

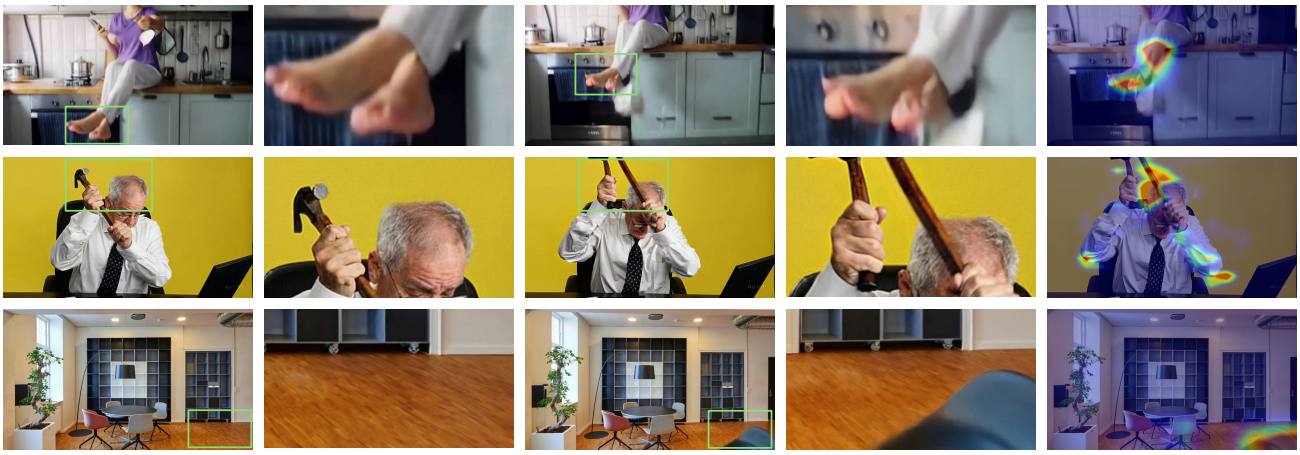


Figure 10. **Geometry Grad-CAM results of L3DE.** For the first video, Grad-CAM detect inconsistent geometric structures in the person's feet, thereby highlighting the corresponding regions. Specifically, the foot region in the analyzed frame differs from that in the reference frame, exhibiting noticeable blurring and distortion. Such degradation of geometric structure does not conform to real-world patterns. For the second video, Grad-CAM detect an abnormal geometric change in the hammer. In the first half of the video, the elderly person holds a single hammer, but in the subsequent frame, the geometry of the hammer suddenly exhibits a 'cloning' effect, splitting into two. Such geometric inconsistency does not conform to real-world geometry rules. For the third video, Grad-CAM detect regions where a chair suddenly appears in the video. Such sudden changes in scene geometry are inconsistent with real-world patterns.

arXiv:2408.06072, 2024. 6

[29] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024. 1

[30] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 7

[31] Shizhen Zhao, Xin Wen, Jiahui Liu, Chuofan Ma, Chunfeng Yuan, and Xiaojuan Qi. Learning from neighbors: Category extrapolation for long-tail learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 7