CETCAM: Camera-Controllable Video Generation via Consistent and Extensible Tokenization

Supplementary Material

6. Phase I Training Data Filtering Details

This section provides the full details of the filtering pipeline used to construct the Phase I training dataset. Our goal is to obtain a large, diverse collection of raw Internet videos that exhibit meaningful camera motion, sufficient aesthetic quality, and reliable geometry estimates from VGGT [23]. Starting from an initial pool of approximately 330K crawled videos, the final filtered dataset contains about 100K clips. We obtain the captions of the training data from [30] and use them as prompts during Phase I training. We describe each filtering stage below.

6.1. Source Collection and Initial Preprocessing

We gather publicly accessible Internet videos covering a broad range of content types, including indoor and outdoor natural scenes, human–object interactions, cinematic B-roll, landscape footage, and casual handheld recordings. All videos are used in full accordance with their respective licenses, which permit research activities such as model training. We further verify that the dataset excludes all offensive, violent, or otherwise inappropriate material. No identity-specific content is intentionally collected. Each raw video undergoes the following preprocessing steps:

- **Frame extraction:** Frames are uniformly re-sampled at 16 FPS.
- **Resolution normalization:** The longer image side is resized to 720 pixels while preserving aspect ratio.
- **Clip segmentation:** Videos longer than 45 seconds are divided into 8–12 second clips via shot-boundary detection using HSV-gradient and color-histogram thresholds.
- **Discard rules:** Clips with effective resolution below 360p, severe compression artifacts, or fewer than 81 usable frames are removed.

This produces roughly 330K candidate clips for further filtering.

6.2. Camera Motion Detection with VGGT

Phase I training requires videos with visible camera motion rather than static-camera footage. For each clip, we apply VGGT [23] to estimate per-frame extrinsic parameters $\{\mathbf{E}_t\}_{t=1}^{T}$, where $\mathbf{E}_t = [\mathbf{R}_t|\mathbf{T}_t]$.

Translation and Rotation Magnitudes. We compute the frame-to-frame translation:

$$\Delta_T(t) = \|\mathbf{T}_{t+1} - \mathbf{T}_t\|_2, \tag{9}$$

and the quaternion-based rotation:

$$\Delta_R(t) = 2\arccos\left(|\langle \mathbf{q}_{t+1}, \mathbf{q}_t \rangle|\right). \tag{10}$$

We use the mean values μ_T and μ_R as indicators of camera motion.

Static-Camera Rejection. A clip is removed if the motion magnitudes satisfy:

$$\mu_T < 0.002 \quad \text{and} \quad \mu_R < 0.5^\circ. \tag{11}$$

We additionally discard clips with discontinuous, unstable, or invalid pose predictions (e.g., quaternion flips, NaNs). After this step, approximately 160K clips remain.

6.3. Aesthetic Quality Filtering

To ensure the model does not learn low-quality photometric statistics, we evaluate each clip using the VBench [10] aesthetic quality predictor. For each frame:

$$s_{\text{aesthetic}} = \frac{1}{T}\sum_{t=1}^{T} f_{\text{VBench}}(I_t). \tag{12}$$

We discard clips with $s_{\text{aesthetic}} < 0.20$. Heuristic rules remove additional low-quality clips exhibiting excessive compression, over/underexposure, or artificial slideshow-like pan-zoom effects. This step removes approximately 60K clips.

6.4. Filtering Summary

Stage	Videos	Reduction
Raw crawled videos	330,421	–
Motion filtering	160,002	-52%
Aesthetic filtering	100,021	-38%

Table 3. Phase I training data filtering summary.

Table 3 summarizes the filtering pipeline. The final Phase I dataset thus offers stable and meaningful camera motion and high aesthetic quality, providing a strong foundation for learning robust and generalizable camera control.

7. HoIHQ Benchmark Collection

This section provides details of how we construct the HoIHQ benchmark used in Phase II training and all evaluations involving human–object interactions (HoI) and

high-fidelity indoor scenes. The goal of HoIHQ is to create a high-quality, diverse, and geometry-rich dataset that complements CameraBench [15] and Uni3C-OOD-Challenging [4]. All videos in HoIHQ are generated using the state-of-the-art commercial video foundation model Kling 2.5 [21], paired with carefully curated source images. The content focuses primarily on shopping-related scenarios with high commercial value. No real identity-specific content is included. Below, we describe each component in detail.

7.1. Source Image Collection

Our pipeline starts with constructing a diverse image corpus emphasizing human–object interactions (HoI) and indoor settings. All images are sourced from publicly accessible datasets and copyright-free photography repositories that permit research usage. To ensure strong geometric cues and compatibility with controllable video generation, we apply the following preprocessing and filtering steps:

- **Resolution requirement:** Images must have a minimum longer-side resolution of 720 px.
- **Foreground quality:** The main subject (human, object, or both) must be clearly visible, not heavily occluded, and free of motion blur.
- **Scene type diversity:** Images are selected across a broad set of categories including kitchen, living room, office, workshop, retail stores, and studio environments.
- **Safety filtering:** Images containing identifiable private individuals, minors, copyrighted characters, or sensitive content are manually removed.

A total of **500** high-quality HoI-focused images are preserved after filtering.

7.2. Prompt Generation

Before generating videos, we first construct a text prompt for each input image. Each prompt contains two components: a content description that captures the scene semantics, and a camera-movement instruction that specifies the intended motion. Both components are generated jointly using a Tarsier-2 [30]. We feed the image into Tarsier-2 and, through carefully designed system instructions, ask the model to produce a prompt that (1) accurately reflects the visual content—especially human–object interaction, (2) includes plausible rich motion movements, (3) includes explicit, large-scale camera-movement directives. This design encourages the model to learn strong and diverse camera motions in the resulting training videos. Here we provide some sample prompts used to generate videos:

- **Sample Prompt 1.** Camera: wide orbit left with extreme dolly out. Outdoors, captured in a medium shot, a woman stands near the ocean, holding a phone with a stylish case in her hand. The background features soft waves and a cloudy sky, creating a serene coastal vibe. She gently raises the phone to eye level, showcasing the case's design while maintaining a natural and relaxed posture. The overall atmosphere is calm and breezy, complementing the coastal setting.
- **Sample Prompt 2.** Camera: wide orbit right with extreme dolly in. Outdoors, captured in a bright, sunny medium shot, a woman sits poolside holding a sunscreen spray bottle. She gently sprays the product onto her leg and spreads it evenly, highlighting the ease of application and smooth texture. The scene conveys a carefree, summery mood, enhanced by the sparkling pool in the background.
- **Sample Prompt 3.** Camera: extreme dolly out with sweeping pan left. Indoors, captured in a medium shot, a child wearing a bright yellow sweater sits at a table with a glass of chocolate drink in front of them. The child picks up the glass with both hands and takes a sip, emphasizing the drink's creamy texture. The vibrant blue background and cheerful setting reinforce a playful and inviting atmosphere.

7.3. Video Generation Using Kling 2.5 [21]

For each source image and each prompt with camera, we generate a 5-second video using the Kling 2.5 video generation model. Kling 2.5 is chosen due to its strong scene coherence, high-resolution synthesis, and stable handling of indoor scenes. Each video cost about five US dollars. The generation protocol is standardized to maintain reproducibility:

- **Input:** See Section 7.1.
- **Prompt:** See Section 7.2.
- **Model Settings:** 16 FPS, 5-second duration, default Kling sampling schedule.

We generated a dataset of 500 videos. The dataset is randomly split into two parts, where 100 videos are left for evaluation, while the remaining videos are used for training.

8. Quantitative Comparison against VACE

To evaluate the extensibility of CETCAM beyond camera control, we conduct a detailed quantitative comparison against the unified controllable video generation framework VACE [11]. We benchmark both methods across four representative controllable-generation tasks: inpainting, gray-to-color, scribble-to-video, and reference-image conditioning. Source videos, prompts and videos are from the VACE benchmark. For each task, we report VBench scores (overall, consistency, aesthetic quality, imaging quality, temporal stability, and motion smoothness), 3D camera metrics (ATE/RPE/RRE), and two-part human evaluation: *Human-Gen* for overall perceptual quality and *Human-Cam* for perceived camera-following accuracy and smoothness. When grading *Human-Cam*, human participants are given a de-

Method	Overall	Subject	Background	Aesthetic	Imaging	Temporal	Motion	ATE↓	RPE↓	RRE↓	Human-Gen↑	Human-Cam↑
Task: Inpainting												
VACE [11]	84.91	89.34	92.17	**61.42**	71.33	**98.14**	97.28	3.912	1.942	14.21	91.2	12.3
CETCAM (Ours)	**86.73**	**91.26**	**93.51**	59.31	**73.28**	97.93	**99.41**	**0.812**	**0.301**	**3.674**	**92.8**	**89.6**
Task: Gray-to-Color												
VACE [11]	85.77	90.54	92.83	**60.74**	70.08	97.56	96.95	3.504	1.721	12.92	92.0	10.7
CETCAM (Ours)	**86.12**	**91.04**	**93.08**	58.91	**72.74**	**97.85**	**99.44**	**0.793**	**0.286**	**3.622**	**91.9**	**90.4**
Task: Scribble-to-Video												
VACE [11]	84.62	89.03	91.04	57.64	69.08	**97.61**	96.22	4.221	2.104	15.01	90.7	11.8
CETCAM (Ours)	**85.57**	**90.52**	**92.64**	**57.83**	**71.30**	97.44	**99.10**	**0.864**	**0.319**	**3.811**	**91.4**	**88.7**
Task: Reference Image												
VACE [11]	86.92	91.71	**94.33**	**61.12**	71.21	97.86	97.44	3.712	1.863	14.83	93.1	13.9
CETCAM (Ours)	**87.24**	**92.81**	93.57	58.62	**76.02**	**98.06**	**99.58**	**0.768**	**0.259**	**3.414**	**93.4**	**92.7**

Table 4. **Extensibility comparison between VACE and CETCAM across four compositional control tasks.** VACE achieves competitive or even stronger VBench scores in a few dimensions (e.g., aesthetic quality, temporal stability, or background consistency). However, VACE completely fails at camera-following, resulting in extremely poor pose metrics and Human-Cam scores. CETCAM maintains competitive generation quality while delivering accurate and smooth 3D-consistent camera control across all tasks.

scription of the input camera trajectory. Other settings are same as specified before in Section 4.3.

Table 4 shows that VACE achieves competitive—and occasionally higher—VBench scores on appearance-centric metrics such as aesthetic quality, temporal stability, or background consistency. This reflects its strong underlying video diffusion backbone and its ability to perform high-quality appearance editing under various conditioning modalities. However, VACE fundamentally lacks camera-following capability. Since its architecture does not incorporate explicit geometry grounding or pose-dependent conditioning, VACE produces videos that visually resemble the controlled input but do not track the prescribed camera trajectory. This results in extremely poor pose accuracy across all tasks, with ATE, RPE, and RRE one to two orders of magnitude worse than CETCAM. The Human-Cam evaluation confirms this failure: participants consistently rated VACE's camera adherence as very low ($\sim$10–14), often reporting "static-camera" or "incorrect motion" artifacts. In contrast, CETCAM maintains strong and consistent performance across both appearance and geometry domains. Overall, CETCAM offers the best of both worlds: competitive generation quality and reliable camera-following behavior. Meanwhile, VACE remains effective as a general-purpose controllable generator but is fundamentally unsuitable for camera-based control due to its absence of 3D-consistent mechanisms.

9. Model Parallelization Details.

To save GPU memory, we implement two essential strategies: Fully Sharded Data Parallel (FSDP) [33] and Sequence Parallelism (SP) [14]. These two techniques are applied on the base Wan DiT blocks, not on the CETCAM Context Blocks or VACE Context Blocks. FSDP [33] partitions model parameters, gradients, and optimizer states across multiple GPUs, substantially reducing per-device memory usage while preserving scalability. We enable mixed precision with `bf16` computation to further reduce activation memory. SP [14] further distributes long sequence computations among devices, allowing efficient handling of large spatiotemporal token sequences during training. We use a SP size of two during our training.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *CVPR*, pages 22875–22889, 2025. 3

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from a single video, 2025. 2

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. 3

[4] Chenjie Cao, Jingkai Zhou, shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. *arXiv preprint arXiv:2504.14899*, 2025. 2, 3, 4, 6, 7, 8

[5] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pages 7310–7320. IEEE Computer Society, 2024. 3

[6] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2025. 3, 6, 7

[7] Chen Hou and Zhibo Chen. Training-free camera control for video generation. In *ICLR*, 2025. 2, 3

[8] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh, 2025. 3

[9] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. 3

[10] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 5, 7, 8, 1

[11] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing, 2025. 2, 3, 4, 5, 7, 8

[12] Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention, 2025. 3

[13] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 3

[14] Shenggui Li, Fuzhao Xue, Yongbin Li, and Yang You. Sequence parallelism: Making 4d parallelism possible. *CoRR*, abs/2105.13120, 2021. 6, 3

[15] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video. 2025. 2, 5, 7, 8

[16] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10258–10268, 2024. 2, 3

[17] Stefan Popov, Amit Raj, Michael Krainin, Yuanzhen Li, William T. Freeman, and Michael Rubinstein. Camctrl3d: Single-image scene exploration with precise 3d camera control. In *2025 International Conference on 3D Vision (3DV)*, pages 649–658, 2025. 2, 3

[18] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, 2025. 2, 3

[19] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pages 501–518, Cham, 2016. Springer International Publishing. 3

[20] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *NeurIPS*, 37:80220–80243, 2024. 2, 3

[21] Kuaishou Technology. Kling ai: Text-to-video generation from kuaishou. https://klingai.com, 2024. Accessed: 2025-11-05. 5, 2

[22] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 4, 5

[23] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt:

Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3, 4, 5, 8, 1

[24] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Advances in Neural Information Processing Systems*, pages 7594–7611. Curran Associates, Inc., 2023. 3

[25] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 3

[26] Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Epic: Efficient video camera control learning with precise anchor-video guidance, 2025. 3, 7

[27] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 2, 3

[28] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2, 3

[29] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding, 2025. 6

[30] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025. 1, 2

[31] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, XIAOPENG ZHANG, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *ICLR*, 2024. 3

[32] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *ECCV*, pages 273–290. Springer, 2024. 3

[33] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel, 2023. 6, 3

[34] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint*, 2025. 3, 6, 7

[35] Zhenghong Zhou, Jie An, and Jiebo Luo. Latent-reframe: Enabling camera control for video diffusion models without training. In *ICCV*, pages 12779–12789, 2025. 2, 3