ConsID-Gen: View-Consistent and Identity-Preserving Image-to-Video Generation

Supplementary Material

1. ConsIDVid Dataset: Additional Details

ConsIDVid is primarily built upon real-world, object-centric videos collected from public sources [22, 34] and is additionally supplemented by proprietary datasets.

1.1. Synthetic Video Construction

Synthetic Video Generation. To significantly enhance the dataset's diversity and coverage, we generate synthetic videos utilizing FramePack [39], a video generator built upon HunyuanVideo [17]. Given that the single-image conditioning used in the standard FramePack pipeline offers limited visual guidance, we extend the framework to support synthesis by conditioning on start and end keyframes.

Controlled Keyframe Selection Strategy. For the synthetic samples derived from MVImgNet2.0 [11], we consciously avoid directly stitching its complete multi-view image sequences. These sequences frequently exhibit rapid camera motion or contain multiple full rotations around the object, which leads to excessive viewpoint shifts. Instead, we employ a controlled strategy where the first frame of each sequence is designated as the starting keyframe, and the ending keyframe is selected from indices 4 through 8 based on the LAION aesthetic predictor [26].

1.2. Hierarchical Video Captioning

In Section 3.3, we adopt a Hierarchical Video Captioning strategy to construct video captions in a structured manner. This process involves generating captions by two distinct levels. The detailed instruction templates used for both levels of caption generation are illustrated in Figure 8 and Figure 9, respectively.

1.3. Comparison with Existing Video Datasets

Recent efforts [6, 19, 32] in video generation primarily focus on collecting large, general-purpose video datasets to train video generative models. However, domain-specific video datasets remain limited in both scale and diversity. As shown in Table 1, many existing domain-focused resources are mostly human-centric (e.g., UCF-101 [28], Taichi-HD [27], FaceForensics++ [24]), making them inadequate for capturing fine-grained object identity or rigid-object motion patterns. While prior approaches like Track4Gen [15] relied on small and minimally curated appearance-preserving datasets, we introduce ConsIDVid. This large-scale, object-centric, identity-preserving video dataset, curated via a scalable pipeline, also includes an appearance-preserving benchmark for standardized evaluation of I2V models.

2. ConsIDVid-Bench

An ideal Image-to-Video (I2V) generator must not only align with the text prompt but, crucially, preserve visual fidelity throughout the temporal dynamics. Accurately quantifying appearance drift and geometric distortion is paramount for fine-grained video generation evaluation. Therefore, in our experiments, we utilize established evaluation metrics from VBench [13, 14] while introducing novel Multi-View Consistency metrics to rigorously measure view and object fidelity.

2.1. VBench Metrics for I2V Evaluation

VBench extends Text-to-Video (T2V) metrics to the I2V domain, focusing on semantic and temporal consistency.

- **I2V Subject**: Cosine similarity between DINO [3] features of the input image and the generated frames, mea-

Table 1. Comparison of existing domain-specific video generation datasets and our ConsIDVid.

Dataset	Year	Scenario	#Videos	Avg. Len (s)	Dur. (h)	Resolution	Caption	Motion Type
UCF-101 [28]	2012	Human	13.3K	7.2	26.7	240p	Short	Text
MSP-Avatar [25]	2015	Human	74	–	3	1080p	N/A	Landmark, Pose
Taichi-HD [27]	2019	Human	3K	–	–	256p	Short	Text
TikTok-v4 [4]	2023	Human	350	–	1	–	N/A	Skeleton
SkyTimelapse [35]	2018	Sky	35K	–	–	360p	N/A	–
FaceForensics++ [24]	2019	Face	1K	–	–	Diverse	N/A	–
CelebV-HQ [40]	2022	Portrait	35K	6.6	68	512p	N/A	–
ChronoMagic [38]	2024	Metamorphic	2K	11.4	7	Diverse	Long	Text
ConsIDVid	2025	Rigid Object	44.3K	8.4	104	Diverse	Hierarchical	Text, Images

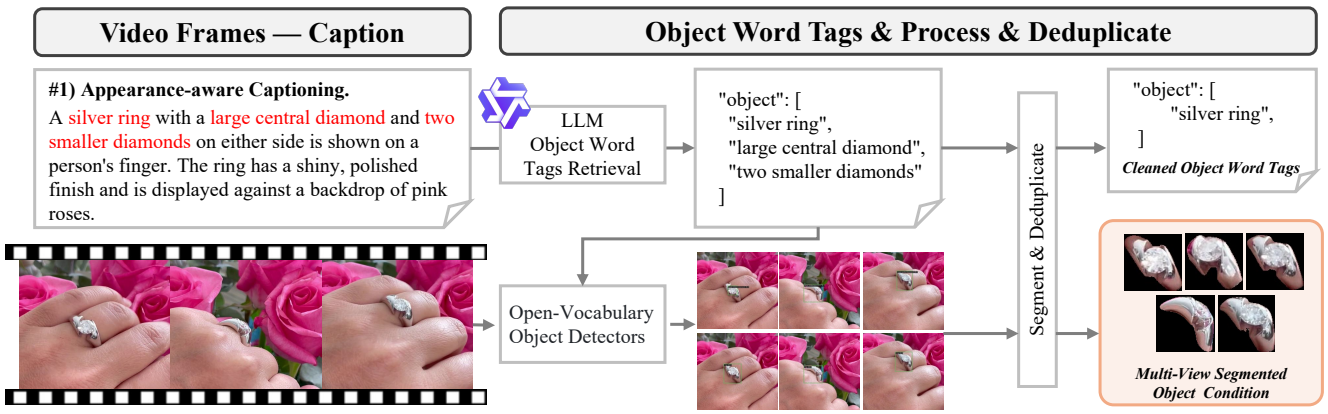

Figure 1. **Overview of object similarity pipeline.** It extracts clean multi-view object segments via caption-based word retrieval, open-vocabulary detection, segmentation, and de-duplication.

suring the preservation of the subject from the input image within the generated video.

- **I2V Background**: DreamSim [8] feature similarity between the input image and generated frames, assessing the visual consistency of the scene/background.
- **Subject Consistency**: Average cosine similarity of DINO features across consecutive frames, evaluating subject appearance consistency throughout the video.
- **Background Consistency**: Average cosine similarity of CLIP [20] features across consecutive frames, measuring the temporal consistency of the background scene.
- **Motion Smoothness**: Motion prior score derived from a video frame interpolation model [18], assessing whether the generated motion remains smooth.
- **Temporal Flickering**: Mean absolute difference between consecutive frames at the pixel level, detecting high-frequency artifacts and local temporal inconsistencies in the generated video.

2.2. Multi-View Metrics for I2V Evaluation

Instead of relying solely on single-frame image-to-video similarity, we further assess video consistency by sampling multi-view (multi-frame) observations from the ground-truth video. This approach allows for a more rigorous measurement of fine-grained identity preservation via the following proposed metrics:

- **Video Similarity**: Average cosine similarity between CLIP features of the ground-truth and generated sampled frames. This metric quantifies overall video realism and content preservation.
- **Object Similarity**: Average cosine similarity between DINO features of the segmented objects in the reference images and the corresponding segments in the generated frames. For rigorous evaluation, multiple reference embeddings per object category are used, and missing objects receive a fixed low similarity penalty. This metric further assesses fine-grained object identity preservation.

2.3. Geometric-aware Metrics for I2V Evaluation

In the context of rigid-object centric I2V generation, the synthesized video can be viewed as a multi/cross-view image sequence derived from a single input image. Crucially, these generated images must exhibit 3D geometric consistency to form a coherent object representation over time. While an I2V generator may produce frames that diverge from the ground-truth, we fundamentally require them to be geometrically consistent with each other.

To address the limitations inherent in ground-truth-dependent geometric evaluation, our key idea is to measure geometric consistency via self-consistency in 3D between the generated multi-view videos. We quantify the geometric fidelity of the I2V output using the following metrics:

- **Chamfer Distance (CD)**: To evaluate this property, we reconstruct 3D point clouds from sampled frames of the generated videos using VGGT [31], followed by point filtering and rigid alignment via ICP (Iterative Closest Point). Our ground-truth point cloud is similarly generated from the true video frames using the same VGGT pipeline. Following prior work on multi-view consistency, we then measure the bidirectional geometric discrepancy between two reconstructed point clouds. This metric captures global shape alignment while penalizing geometric drift or deformation across the synthesized views.
- **MEt3R** [1]: This metric evaluates view consistency by employing DUSt3R [33] to obtain dense 3D reconstructions from image pairs. It measures consistency by projecting DINO + FeatUp [9] features from one view to the other using the reconstructed geometry and calculating the feature similarity among the resulting views. This provides a reliable measure of geometric self-consistency for multi-view coherence in generated images.

Table 2. Quantitative comparison of model performance on ConsIDVid-Bench under two penalty settings (penalty $= 0.1$ and 0.5), evaluated by object similarity. Inference latency is measured on a single NVIDIA A100 GPU. Best and Second-best scores are highlighted.

Model	Params	Latency	Penalty = 0.1		Penalty = 0.5	
			Proprietary	Public	Proprietary	Public
Wan2.1 [29]	1.3B	202 (s)	66.9	69.1	67.1	69.6
SkyReelv2 [5]	1.3B	393 (s)	59.5	68.0	60.0	68.5
ConsistI2V [23]	5.2B	–	62.0	62.4	62.7	63.1
Wan2.2 [30]	5B	359 (s)	68.6	71.6	68.9	72.1
CogVideoX1.5 [37]	5.2B	–	60.1	61.5	60.5	62.1
HunyuanVideo [17]	13B	–	64.3	67.4	64.6	67.6
Wan2.1 [29]	14B	970 (s)	67.9	72.2	68.2	72.8
ConsID-Gen	1.8B	199 (s)	69.2	71.8	69.9	72.3

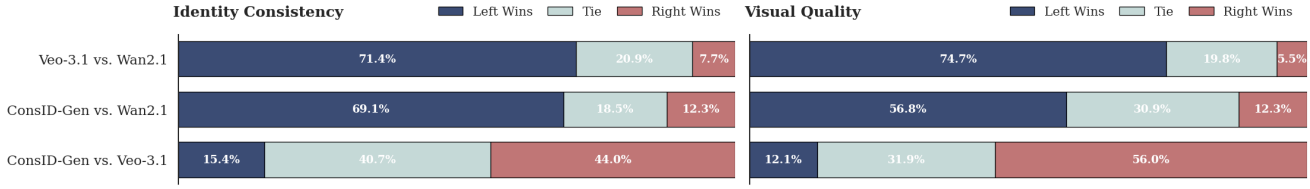

Figure 2. **Human Evaluation** results for Identity Consistency (left) and Visual Quality (right).

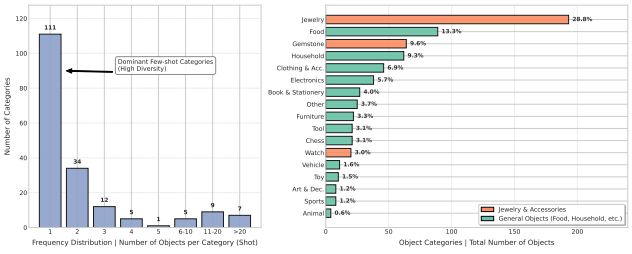

Figure 3. **Statistics of ConsIDVid-Bench.** Left: Frequency distribution of categories; Right: Object category breakdown.

2.4. ConsIDVid-Bench Statistics

As illustrated in Figure 3, ConsIDVid-Bench features a highly diverse distribution of object categories, predominantly encompassing jewelry, food, and household items in few-shot settings.

3. Object Similarity Evaluation

As illustrated in Figure 1, we propose an object similarity evaluation designed to measure fine-grained appearance consistency across generated videos. Our method utilizes multi-view frames rather than relying solely on the first frame, as it is in the I2V Subject, thereby ensuring robustness against background variations. First, we employ a Large Language Model (LLM) [36] for the first stage output of Hierarchical Video Captioning to retrieve object-related word tags. Next, these tags guide an open-vocabulary object detector [10] to localize objects in 5 sampled keyframes, followed by segmentation [21]. Finally, to ensure data reliability, we de-duplicate and consolidate these instances, yielding a reliable set of cleaned object word tags and their corresponding segmented visual counterparts for precise, object-level comparison.

Table 3. Quantitative Comparison with Wan2.1-InP-FL and StepVideo-TI2V.

Method	I2V-Subj	I2V-Back	Subj-Cons.	Back-Cons.	Motion	Temp
Wan2.1-InP-FL	95.98	96.80	90.89	94.43	99.22	98.44
StepVideo-TI2V	97.99	98.33	92.54	94.53	99.38	98.61
ConsID-Gen	**98.31**	**98.66**	**95.30**	**96.10**	**99.52**	**99.24**

4. More Quantitative Comparison Evaluations

Compare with Wan2.1-InP-FL and StepVideo-TI2V. Although our approach does not explicitly model camera trajectories, it preserves identity fidelity through textual guidance alone. To further assess its effectiveness, we compare our method against strong I2V baselines that leverage additional conditioning signals, including Wan2.1-InP (first-last frame) [29] and StepVideo-TI2V [12]. Despite relying only on text-based guidance, our approach achieves superior performance. We also evaluate VACE [16]; however, it yields worse results (e.g., I2V-Subj: 82.66), likely because VACE is primarily designed for video creation and editing, which tends to inadvertently modify the input content.

Object Similarity. As shown in Table 2, ConsID-Gen achieves consistently strong performance across both evaluation settings. Under the default penalty of 0.1, where

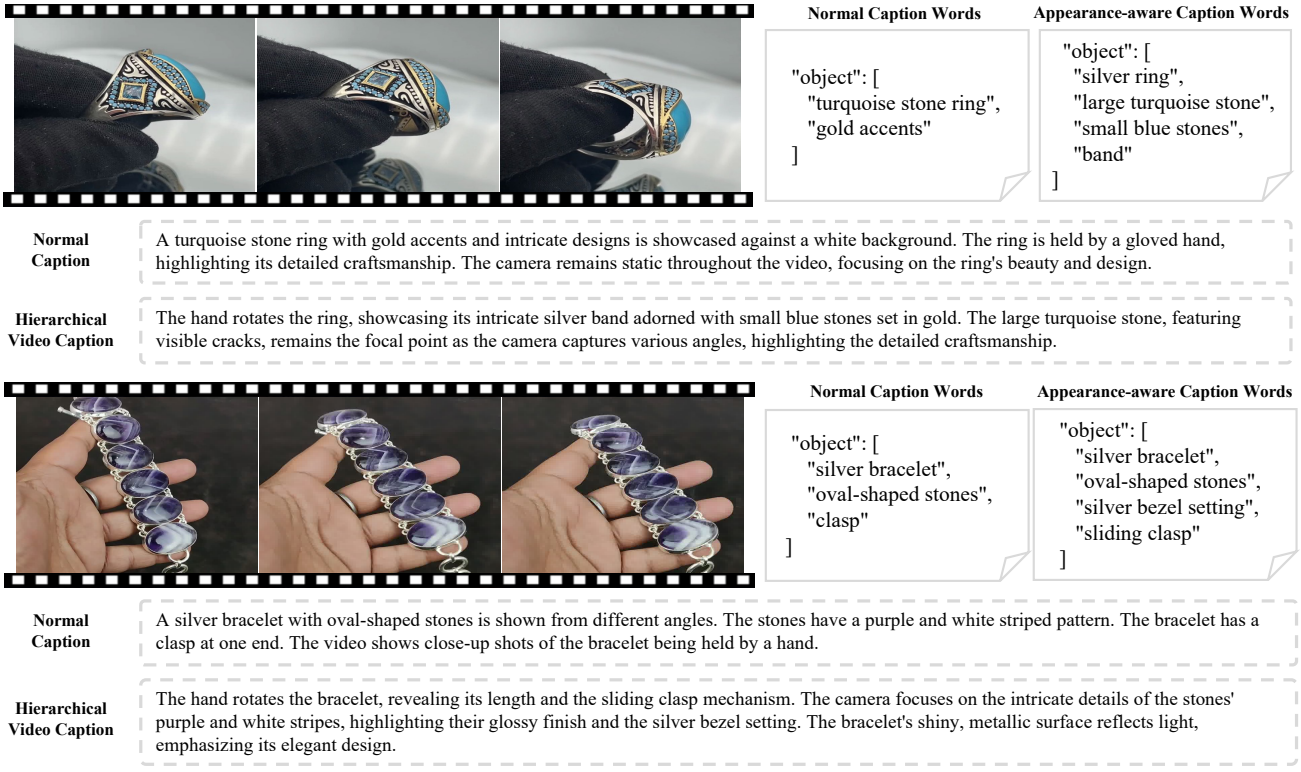

Figure 4. **Qualitative comparison between normal captioning and Hierarchical Video Captioning.** The right side displays the retrieved object word tags from the normal caption and the appearance-aware caption, while the lower section illustrates the captions generated by normal and hierarchical captioning.

video frames with missing objects are assigned a similarity score of 0.1, ConsID-Gen obtains the highest Object Similarity score on the proprietary set, surpassing all competing models of similar or larger scale. A similar trend can be seen when comparing the results for a higher penalty of 0.5. This robust performance indicates that ConsID-Gen effectively maintains stable object visibility and identity fidelity throughout the generated video sequence.

Human Preference Evaluation. We conducted a side-by-side user study to benchmark ConsID-Gen against the open-source Wan2.1 [29] and proprietary Veo-3.1 [7]. Participants were presented with randomized video pairs and asked to express their preference (better or tie) regarding *Identity Consistency* and *Visual Quality*. As shown in Figure 2, our method consistently outperforms Wan2.1 across both metrics. Compared to Veo-3.1, we achieve comparable results in identity consistency.

5. More Ablation Evaluations

Effect of Datasets. To investigate the impact of the data, we fine-tuned Wan2.2-5B via LoRA (rank 64). As shown in Table 4, the resulting Wan2.2-5B-FT showed limited improvement compared to ConsID-Gen, which achieved significantly better results. This underscores that our architectural design is the key factor behind the performance boost.

Table 4. Quantitative ablation of dataset effectiveness. We evaluate model performance using VBench-I2V suite.

Method	I2V-Subj	I2V-Back	Subj-Cons.	Back-Cons.	Motion	Temp
Wan2.2-5B	96.85	97.57	91.99	94.82	98.93	98.10
Wan2.2-5B-FT	97.61	98.17	91.22	94.64	99.21	98.37
ConsID-Gen	**98.31**	**98.66**	**95.30**	**96.10**	**99.52**	**99.24**

Table 5. Effectiveness of Hierarchical Video Captioning. Comparison of object word tag retrieval between normal captioning and our Stage-1 appearance-aware captioning, demonstrating that the latter yields richer and more precise appearance tags.

Method	Avg. Objects / Video	Avg. Word Len.
Normal Caption	3.18	13.69
Appearance-aware Caption	**3.20**	**14.44**

Effective Evaluation on Video Captioning. To assess our Hierarchical Video Captioning strategy, we compared its Stage 1 (Appearance-aware Captioning) against a normal captioning method. Unlike the normal method, which jointly reasons over appearance and temporal dynamics, our Stage 1 processes fewer frames at higher resolution, allowing the VLM [2] to focus exclusively on fine-grained ob-

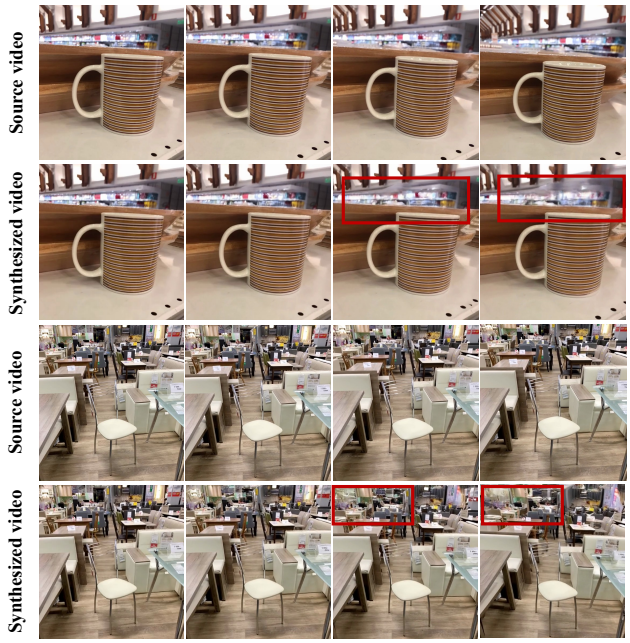

Figure 5. **Visualization of failure cases.**

ject details. Evaluating both on 50 proprietary videos by retrieving object word tags using an LLM [36], we found that Stage 1 yields richer and more precise appearance-centric tags, as presented in Table 5.

6. Miscellaneous Visualization Results

6.1. Effective Evaluation on Video Captioning

As illustrated in Figure 4, Stage-1 appearance-aware captioning produces richer and more precise appearance tags than normal captioning, capturing fine-grained object details such as material, stone texture, and setting structure. Moreover, the hierarchical design enables the model to first ground stable appearance semantics before generating temporal descriptions, yielding more comprehensive and accurate descriptions.

6.2. Additional Comparison with Existing Methods

To complement the quantitative evaluations, we provide additional synthetic samples generated by existing methods. Figure 6 and Figure 7 illustrate these samples on the proprietary and public subsets of ConsIDVid-Bench.

6.3. Visualization of Failure Cases

We present the failure cases in Figure 5. Since our model is built upon a small-scale base model, it inherits certain limitations, particularly in complex scene synthesis. This is notably observed in the public subset, which features more intricate backgrounds. For instance, the model is prone to hallucinations when generating distant or ambiguous details, such as counters and complex furniture arrangements, as highlighted in the red boxes.

7. Limitations

While ConsID-Gen achieves strong appearance preservation, several limitations merit further investigation. First, our method relies on a relatively small baseline due to resource constraints. Although it delivers clear improvements at this scale, adopting larger-capacity models (e.g., 14B) shows promising potential and constitutes an important direction for future work. Furthermore, the baseline is currently restricted to 81-frame sequences. While ConsID-Gen maintains high fidelity within this range, sustaining fine-grained visual consistency across substantially longer horizons remains an open challenge that we aim to address in future research.

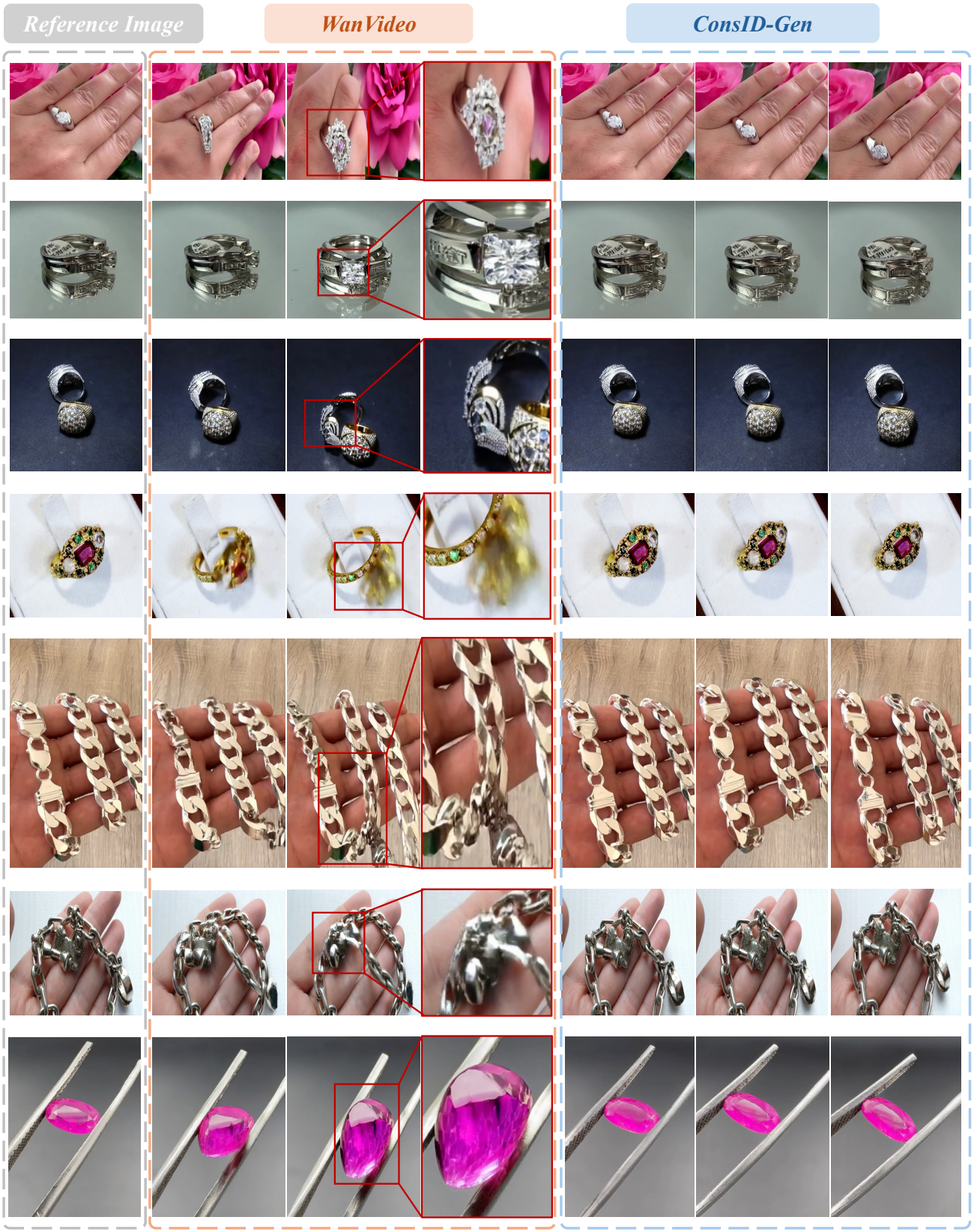

Figure 6. **Additional visual comparisons on the proprietary subset of ConsIDVid-Bench.**

Figure 7. **Additional visual comparisons on the public subset of ConsIDVid-Bench.**

Instruction Template for Appearance-aware Captioning

System Prompt: You are an expert video captioner for object-centric product and item videos. Your task is to write a short descriptive caption focusing only on the visible appearance of the main object.

Rules:

- Write exactly 1–2 sentences under 40 words.
- Describe the main object with 5–7 attributes: category name, color or pattern, material and finish/texture (matte, glossy, brushed, woven), shape or form factor, size or scale cues (in hand, on table), notable parts/features (buttons, seams, zippers, ports, tread), any visible wear or defects, and transcribe any legible text or logos exactly as seen.
- Be strictly factual and concise. Use simple present tense.
- Do not include camera behavior, background context, or usage speculation.
- Focus only on the primary product. Mention hands or props only if they clarify size or material.
- Output only the caption text. No labels, prefixes, hashtags, or emojis.

User Prompt: Write 1–2 factual sentences describing the main item in the video. Include 5–7 visible attributes: category, color or pattern, material and finish/texture, shape or form factor, size or scale cues, notable features like seams or ports, any wear/defects, and exact readable text or logos. Keep it under 40 words, simple present tense, and output only the caption.

Figure 8. **Instruction Template for Appearance-aware Captioning.**

Instruction Template for Temporal-aware Captioning

System Prompt: You are an expert video captioner for object-centric videos. Your task is to generate a single coherent, factual, and detailed caption that integrates the provided appearance description of the main object into a natural temporal-aware observation.

Rules:

- Write 2–3 sentences under 60 words.
- Temporal dynamics must be the main focus: describe camera movement (type, direction, angle, framing), human interaction (holding, rotating, tapping, opening, placing), and object motion or state change (slides, flips, spins, opens, closes).
- Explicitly reuse at least 3–5 key details from the given appearance description (color, material, texture, shape, text, or distinctive features).
- Weave appearance details smoothly into the temporal description so the caption reads as one fluent observation, not separate parts.
- Use objective, factual language in simple present tense. Avoid subjective or aesthetic terms like "beautifully," "showcases," or "highlights."
- Mention background or props only if directly relevant to scale or interaction.
- Output only the caption text. No labels, prefixes, hashtags, or emojis.

User Prompt: The main object's appearance is: {APPEARANCE_DESCRIPTION} Write 1–2 factual and coherent sentences that integrate this appearance information naturally into a temporal-aware description. Explicitly reuse at least 3–5 details from the appearance description (color, material, texture, shape, text, distinctive features). Describe camera movement, angle, and framing, along with any human interaction or object motion, combining them in one fluent observation. Keep it concise, under 60 words, in simple present tense. Output only the caption.

Figure 9. **Instruction Template for Temporal-aware Captioning.**

References

[1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6034–6044, 2025. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 4

[3] 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. 1

[4] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023. 1

[5] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 3

[6] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 1

[7] DeepMind. Veo 3. https://deepmind.google/models/veo, 2025. 4

[8] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 2

[9] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024. 2

[10] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14987–14997, 2025. 3

[11] Xiaoguang Han, Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, and Shuguang Cui. Mvimgnet2. 0: A larger-scale dataset of multi-view images. *arXiv preprint arXiv:2412.01430*, 2024. 1

[12] Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, et al. Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model. *arXiv preprint arXiv:2503.11251*, 2025. 3

[13] 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. 1

[14] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 1

[15] Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7276–7287, 2025. 1

[16] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 3

[17] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 3

[18] 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. 2

[19] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 1

[20] 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. 2

[21] 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. 3

[22] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 1

[23] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhanc-

ing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 3

[24] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1

[25] Najmeh Sadoughi, Yang Liu, and Carlos Busso. Msp-avatar corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2015. 1

[26] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1

[27] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 1

[28] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[29] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 4

[30] Wan Team. Wan2.2: More powerful, more beautiful. https://wan.video/blog/wan2.2, 2025. 3

[31] 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

[32] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 1

[33] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[34] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 1

[35] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2018. 1

[36] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3, 5

[37] 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 arXiv:2408.06072*, 2024. 3

[38] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Rui-Jie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37:21236–21270, 2024. 1

[39] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2(3):5, 2025. 1

[40] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 1