ID-Crafter: VLM-Grounded Online RL for Compositional Multi-Subject Video Generation

Supplementary Material

Appendices Overview

The supplementary material includes the following sections:

- Sec. A: More Implementation Details.
  - Details of Network Architecture (Sec. A.1).
  - Details of Hierarchical Identity-Preserving Attention (Sec. A.1).
  - Details of Reward Calculation (Sec. A.2).
- Sec. B: Dataset Curation Details.
  - Dataset Composition and Statistics (Sec. B.1).
  - Reference Image Construction Strategies (Sec. B.2).
- Sec. C: Evaluation Metric Details.
- Sec. D: Additional Results and Analysis.
- Sec. E: Additional Discussion.
  - Computational Overhead (Sec. E.1).
  - Ethics Statement (Sec. E.2).

A. More Implementation Details

A.1. Network Architecture

Our model's architecture is founded on three synergistic modules: a Vision-Language Model (**VLM**), specifically the Qwen2.5-VL 7B variant [1], for advanced multi-modal comprehension; a Variational Autoencoder (**VAE**) [10] for efficient spatial compression; and a Diffusion Transformer (DiT) [8, 14], which serves as the latent video diffusion backbone. The architectural configurations for each component are detailed in Table S.1.

The 7B-parameter VLM, comprising a 32-layer Vision Transformer (ViT) and a 28-layer Large Language Model (LLM), is responsible for encoding textual prompts and subject images into a unified, semantically rich space. The VAE, with a 54M-parameter encoder and a 73M-parameter decoder, utilizes an $8\times$ spatial downsampling factor, compressing video frames into a compact latent representation that is computationally tractable.

Subject Image Encoder We employ the pre-trained VAE from Wan-Video [14] to transform subject images from the pixel space into the latent space. This process yields a 16-channel latent representation with an $8\times$ downsampling factor. This latent representation is subsequently processed by a 2×2 patch embedding layer, which further reduces its spatial dimensions to align with the hidden dimension of the DiT backbone, ensuring seamless integration between the visual and generative components. The VAE weights are kept frozen during training to preserve its high-fidelity reconstruction capabilities.

Hierarchical Identity-Preserving Attention As detailed in the main paper, our Hierarchical Identity-Preserving Attention module is central to preventing identity leakage in multi-subject generation. It comprises three sequential stages applied within each block of the DiT, designed to progressively refine and integrate identity features. Let Z_l be the video latent tokens at layer l, and $F_{subj} = \{f_1, f_2, \ldots, f_N\}$ be the token sets for the N subject images.

Stage 1: Intra-Subject Attention. To refine the feature representation for each subject independently, we first apply a standard self-attention layer [13] to each subject's token set f_i. This initial stage allows the model to capture fine-grained, identity-specific details and consolidate features for each subject before any cross-subject interaction occurs.

$$f'_i = \text{SelfAttention}(f_i) \quad \forall i \in \{1, \ldots, N\} \tag{S.1}$$

Stage 2: Gated Inter-Subject Attention. To model interactions between subjects while strictly mitigating identity bleed, we employ a novel gated cross-attention mechanism. For each subject f'_i, it queries the features of all other subjects $\{f'_j\}_{j\neq i}$, which then modulates the output of the cross-attention.

$$f''_i = f'_i + \sigma(W_g f'_i) \odot \text{CrossAttention}(Q = f'_i, KV = \{f'_j\}_{j\neq i}) \tag{S.2}$$

Stage 3: Cross-Modal Attention. Finally, the refined subject tokens F''_{subj} and the textual context tokens f_{txt} attend to the main video latent tokens Z_l. This stage injects the rich semantic guidance from the VLM, now conditioned on the robust and disentangled subject identities. While the VLM features [1] are computed only once, their influence is dynamic, as they modulate the video features at each layer of the DiT.

$$Z'_l = Z_l + \text{CrossAttention}(Q = Z_l, KV = [F''_{subj}; f_{txt}]) \tag{S.3}$$

Table S.2 provides a detailed ablation study on this design. These three stages form a single, cohesive block that is inserted within the first 20 layers of the DiT backbone. This design choice is based on the observation that identity features are most critical in the early stages of the generation process.

Table S.1. **Detailed configuration of our model's primary architectural components**. The VLM, VAE, and DiT modules are designed to handle multimodal understanding, spatial compression, and latent diffusion, respectively.

Configuration	VLM		VAE		DiT
	ViT	LLM	Encoder	Decoder	
# Layers	32	28	11	15	30/40
# Num Heads (Q / KV)	16 / 16	28 / 4	-	-	12 / 40
Head Size	80	128	-	-	128/128
Intermediate Size	3,456	18,944	-	-	8,960/13,824
Patch / Scale Factor	14	-	8×8	8×8	-
Channel Size	-	-	16	16	-
# Parameters	**7B**		**54M**	**73M**	**1.3B / 14B**

Table S.2. **Ablation of Hierarchical Attention.** We analyze the contribution of each stage in our hierarchical attention mechanism. The full three-stage model provides the best performance, with **Stage 3** (cross-modal attention to VLM) being the most critical component for semantic alignment.

Method	FaceSim ↑	NexusScore ↑	Total Score ↑
Full Model (3 Stages)	**66.10%**	**45.1%**	**55.16%**
w/o Stage 1 (Intra-Subject)	63.2%	44.2%	54.1%
w/o Stage 2 (Inter-Subject)	64.5%	44.5%	54.5%
w/o Stage 3 (Cross-Modal)	58.9%	38.7%	50.8%
Vanilla Attention (Stage 3 only)	55.1%	39.1%	49.7%

A.2. Details of GRPO-based Post-training

This section provides a more detailed breakdown of the reward calculation and hyperparameter settings for our online post-training phase, which leverages the Groupwise Policy Optimization (GRPO) algorithm [7].

Advantage Normalization To stabilize the training process and reduce variance in the policy gradient updates, we normalize the advantage term. The normalized advantage $\hat{A}_i$ for the i-th sample in a group of size G is calculated from the total trajectory reward r_i:

$$\hat{A}_i = \frac{r_i - \text{Mean}(\{r_j\}_{j=1}^G)}{\text{Std}(\{r_j\}_{j=1}^G) + 10^{-8}} \tag{S.4}$$

where the total reward r_i is computed by averaging the frame-level rewards over the video's duration to produce a single scalar value. The small epsilon (10^{-8}) is added for numerical stability. The constant a for noise injection σ_t was set to 1.0.

Reward Design As defined in the main paper, the objective is to maximize a composite reward function $\mathcal{R}_{\text{total}}$, defined as a weighted sum of fidelity and quality rewards:

$$\mathcal{R}_{\text{total}}(\mathbf{V}) := w_{\text{fid}}\mathcal{R}_{\text{fid}}(\mathbf{V}, \mathcal{I}) + w_{\text{qual}}\mathcal{R}_{\text{qual}}(\mathbf{V}) \tag{S.5}$$

We empirically set the weights to $w_{\text{fid}} = 0.6$ and $w_{\text{qual}} = 0.4$ to prioritize identity preservation while ensuring high quality. The fidelity reward $\mathcal{R}_{\text{fid}}$ is a combination of face-specific and holistic subject consistency scores, while the quality reward $\mathcal{R}_{\text{qual}}$ assesses aesthetic appeal and physical plausibility.

Hyperparameters and Sensitivity The key hyperparameters used during the GRPO training are detailed in Table S.3. We also conducted a sensitivity analysis on critical parameters, with results presented in Table S.4, to demonstrate the robustness of our training setup.

Table S.3. **Reinforcement Learning Hyperparameters.** Key hyperparameters used for the online optimization stage with our GRPO algorithm.

Hyperparameter	Value
Algorithm	Flow-GRPO [4, 7, 11]
Policy Learning Rate	1e-5
Optimizer	AdamW
AdamW βs	(0.9, 0.95)
AdamW ϵ	1e-8
Batch Size	32
Max Gradient Norm	1.0

Training Dynamics and Reward Effectiveness The effectiveness of our multi-faceted reward model is visualized in Figure S.1, which illustrates the evolution of key metrics during the online GRPO phase. The plot clearly shows a positive trend for identity fidelity (e.g., FaceSim), perceptual quality (Aesthetics), and Q-Align [15] over the course of training. This concurrent improvement across diverse axes validates that our reward model successfully guides the generator towards producing videos that are not only more identity-consistent but also more visually appealing and semantically accurate, without suffering from catastrophic forgetting or reward over-optimization on a single metric.

Ablation on Fidelity Reward A crucial component of our total reward is the fidelity term $\mathcal{R}_{\text{fid}}$, which contains a balance factor γ for the face-specific reward $\mathcal{R}_{\text{face}}$. Table S.4 presents an ablation study on the choice of this balance factor. The results indicate that a value of 0.5 provides an optimal balance between average and worst-case identity fidelity, effectively preventing scenarios where one subject's identity collapses while others are preserved.

Table S.4. **Ablation on Fidelity Reward γ.** We study the effect of γ in $\mathcal{R}_{fid}$, which balances average and minimum identity similarity. A value of 0.5 provides the best trade-off.

γ	Avg FaceSim ↑	Min FaceSim ↑	Total Score ↑
0.0 (Average only)	**67.2%**	45.1%	54.8%
0.25	66.8%	50.3%	55.0%
0.5	66.1%	**53.5%**	**55.2%**
0.75	65.2%	52.1%	54.9%
1.0 (Minimum only)	63.9%	51.5%	54.5%

B. Dataset Curation Details

B.1. Dataset Composition and Statistics

Our dataset is meticulously curated from three heterogeneous primary sources to ensure comprehensive coverage and high fidelity. The foundational component is the large-scale OpenS2V-Nexus dataset [16], which offers a wide array of authentic scenes and actions. To ensure the temporal coherence of the clips, we incorporate a video shot detection stage [12], which contributes a substantial volume of 218,230 videos, accompanied by 535,259 masked subject images and 3,098 generated counterparts. To further elevate the dataset's quality, we incorporate a high-fidelity collection of professionally shot videos with detailed annotations, comprising 9,374 pristine videos and an additional 3,000 generated subject images. This synthetically generated data, where subjects rendered by cutting-edge image editing models are placed into novel contexts, systematically increases the diversity of subject-background compositions. Finally, to provide rich priors for appearance synthesis, we integrate a dedicated Subject Image dataset. This static collection contains 14,976 images, which are further categorized into 2,877 human-centric images and 3,458 isolated clothing items. The overall composition is summarized in Table S.5, and the entire data curation pipeline is illustrated in Figure S.2.

B.2. Reference Image Construction Strategies

We analyze various strategies for constructing reference datasets in Figure S.5. Naive data augmentation offers limited visual diversity and is prone to generating occluded subjects, which compromises generalization to complex scenes. Alternatively, employing on-the-shelf image-consistent generators, such as Flux-Kontext [6] or GPT-4o [5], often introduces undesirable appearance artifacts and, most critically, **fails to maintain subject fidelity**. In contrast, our approach, which leverages a tailored in-context learning prompt for the latest Gemini-Flash-Image model (a.k.a., **Nano Banana**) [2], excels at subject decomposition and synthesis, yielding more faithful and varied reference images.

C. Evaluation Metric Details

Our evaluation protocol is principally derived from the OpenS2V-Nexus Benchmark [16]. We adopt their prescribed prompts but substitute the proprietary models with publicly available counterparts to facilitate evaluation and ensure reproducibility.

NexusScore: This metric quantifies the subject consistency between a generated video $\mathbf{V} = \{f_1, \ldots, f_T\}$ and a reference image $\mathbf{I}_{\text{ref}}$. It employs a two-stage pipeline: (1) Grounded-SAM [9] generates a subject mask M_t for each frame f_t. (2) A refined CLIP-based image encoder [3] $\mathcal{E}_{img}$ computes the cosine similarity between the feature embedding of the reference image and that of the masked subject in each frame. This targeted approach ensures the evaluation focuses strictly on identity fidelity by mitigating background interference. The final score is the average similarity over all $\mathbf{T}$ frames: $S_{\text{Nexus}} = \frac{1}{\mathbf{T}} \sum_{t=1}^{\mathbf{T}} \frac{\mathcal{E}_{\text{img}}(\mathbf{I}_{\text{ref}}) \cdot \mathcal{E}_{\text{img}}(f_t \odot M_t)}{\|\mathcal{E}_{img}(\mathbf{I}_{\text{ref}})\| \|\mathcal{E}_{\text{img}}(f_t \odot M_t)\|}$ where $\odot$ denotes element-wise multiplication.

NaturalScore: This metric assesses the perceptual realism and physical plausibility of a generated video $\mathbf{V}$. We employ a state-of-the-art Multimodal Large Language Model (MLLM) to perform a deep semantic analysis. The model is presented with the video and a carefully designed prompt which instructs it to evaluate spatio-temporal consistency, identify violations of physical laws, and detect visual artifacts characteristic of generative models. The model's direct output, a normalized score, serves as the final measure of the video's naturalness.

GmeScore: This metric evaluates the semantic alignment between a generated video $\mathbf{V}$ and its corresponding text prompt $\mathbf{C}$, proposed as an enhanced alternative to conven-

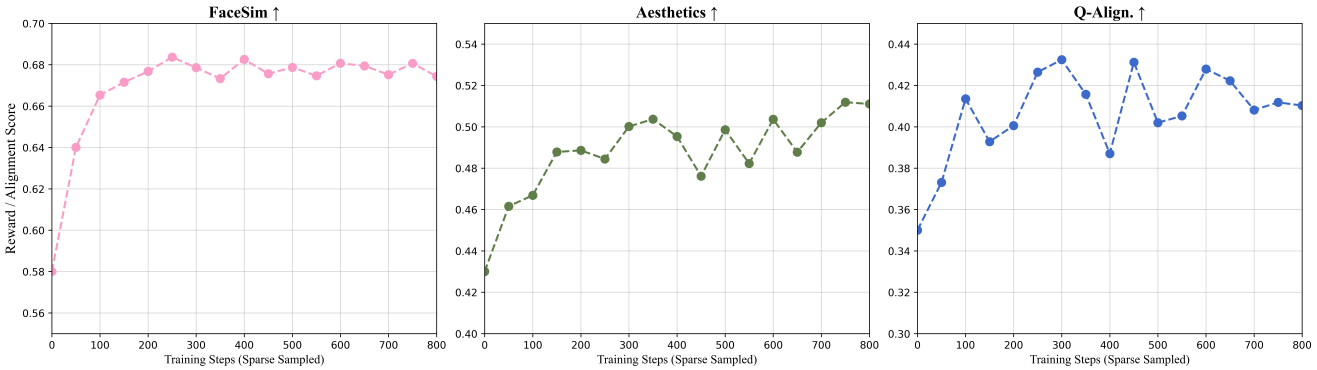

Figure S.1. **Performance Curves of Online GRPO.** We plot the moving average of key evaluation metrics over 800 training steps. Our method demonstrates simultaneous improvement in identity fidelity (FaceSim), perceptual quality (Aesthetics), and text-prompt alignment (e.g., Q-Align [15]), validating the effectiveness of our multi-faceted reward strategy.

Table S.5. **Detailed Composition of the Curated Training Dataset.** This table provides a breakdown of the three primary data sources used to train ID-CRAFTER, detailing the content type, quantity, and its strategic role in the model's training.

Dataset Source	**Content Type**	**Quantity**	**Primary Purpose**
OpenS2V-Nexus [16]	**General Videos**	**218,230**	Broad domain pre-training
	Masked Subjects	**535,259**	Learning subject-background separation
	Generated Subjects	3,098	Initial subject synthesis capability
High-Quality Internal Collection	High-Fidelity Videos	9,374	Fine-tuning for video quality
	Generated Subjects	3,000	Enhancing appearance diversity
Subject Image Dataset	**Total Static Images**	**14,976**	Rich priors for appearance synthesis
	- Human-centric Images	2,877	
	- Isolated Clothing Items	3,458	

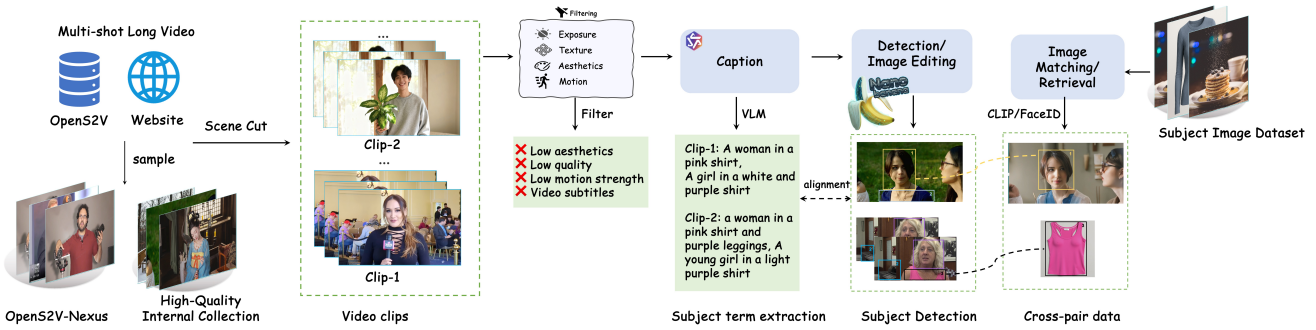

Figure S.2. **Overview of the Data Curation Pipeline.** Our comprehensive data pipeline processes videos and images from diverse sources. It includes subject segmentation, high-quality reference image synthesis via an in-context learning prompt with advanced generative models, and rigorous filtering. This process yields a large-scale, high-fidelity dataset tailored for training identity-preserving video generation models.

tional CLIPScore. To overcome the limitations of prior methods in handling long-form text, GmeScore leverages the advanced text comprehension capabilities of the Qwen2.5-VL [1], which provides a powerful text encoder $\mathcal{E}_{text}$ and a temporally-aware video encoder $\mathcal{E}_{video}$. Unlike methods that average per-frame similarities, GmeScore computes a global feature representation for the entire video, thereby capturing its holistic temporal dynamics. The final score is

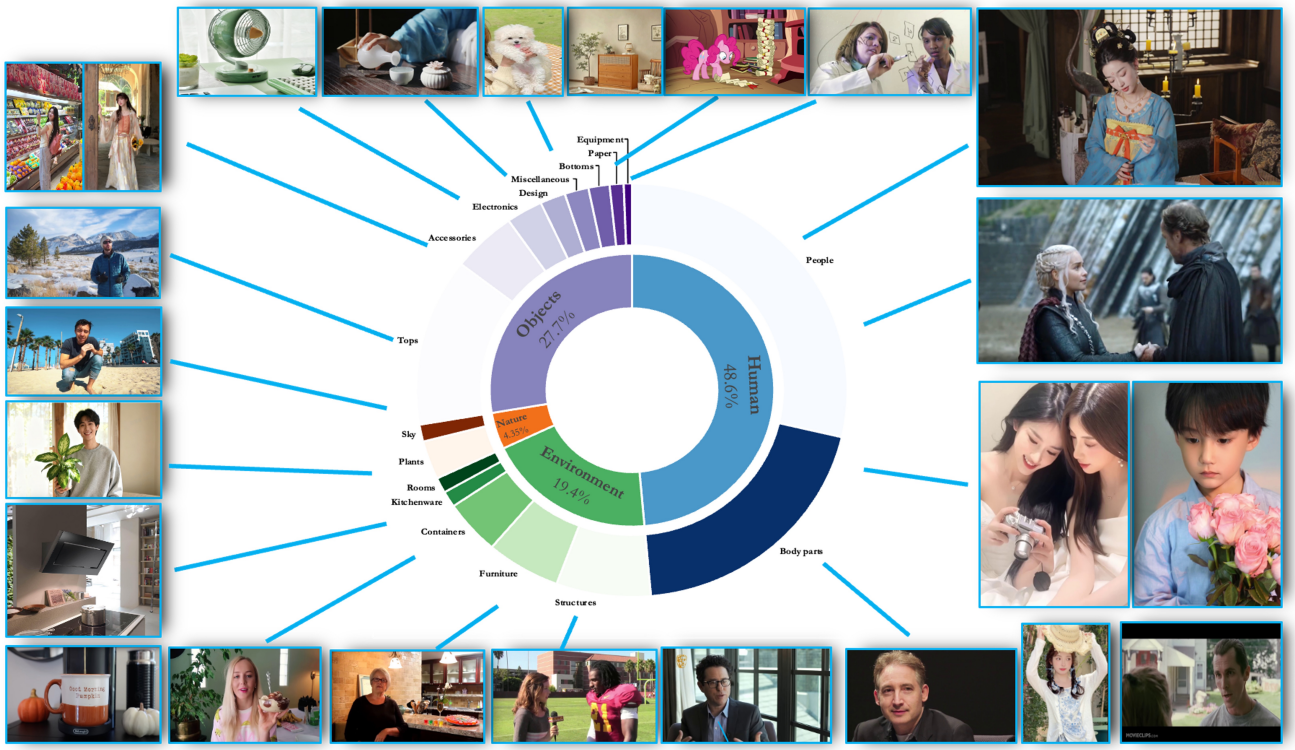

Figure S.3. Statistics of the constructed dataset. The dataset is organized into four primary scenarios: Human, Objects, Environment, and Nature, each containing a variety of subcategories.

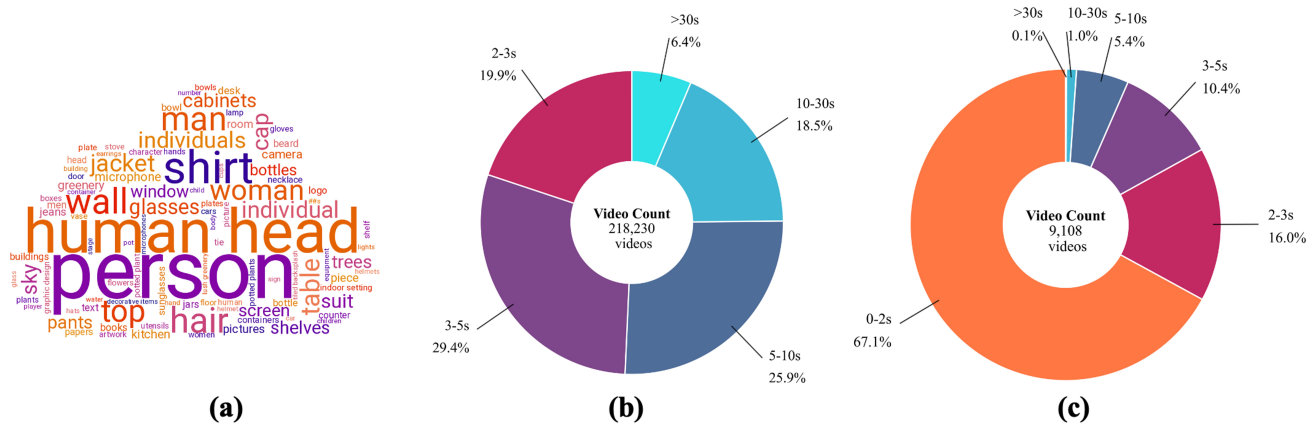

Figure S.4. Statistics of the constructed dataset. (a) A word cloud illustrates the rich linguistic diversity of the reference images. The video duration distributions are shown for the two sources: (b) OpenS2V and (c) our curated videos.

the cosine similarity: $S_{\text{Gme}} = \frac{\mathcal{E}_{\text{video}}(\{f_t\}_{t=1}^{T}) \cdot \mathcal{E}_{\text{text}}(C)}{\|\mathcal{E}_{\text{video}}(\{f_t\}_{t=1}^{T})\| \|\mathcal{E}_{\text{text}}(C)\|}$ This formulation provides a more accurate and comprehensive assessment of text-to-video relevance, especially for complex narratives.

D. Additional Results and Analysis

E. Additional Discussion

E.1. Computational Overhead

The introduction of the Hierarchical Identity-Preserving Attention module adds a marginal computational overhead

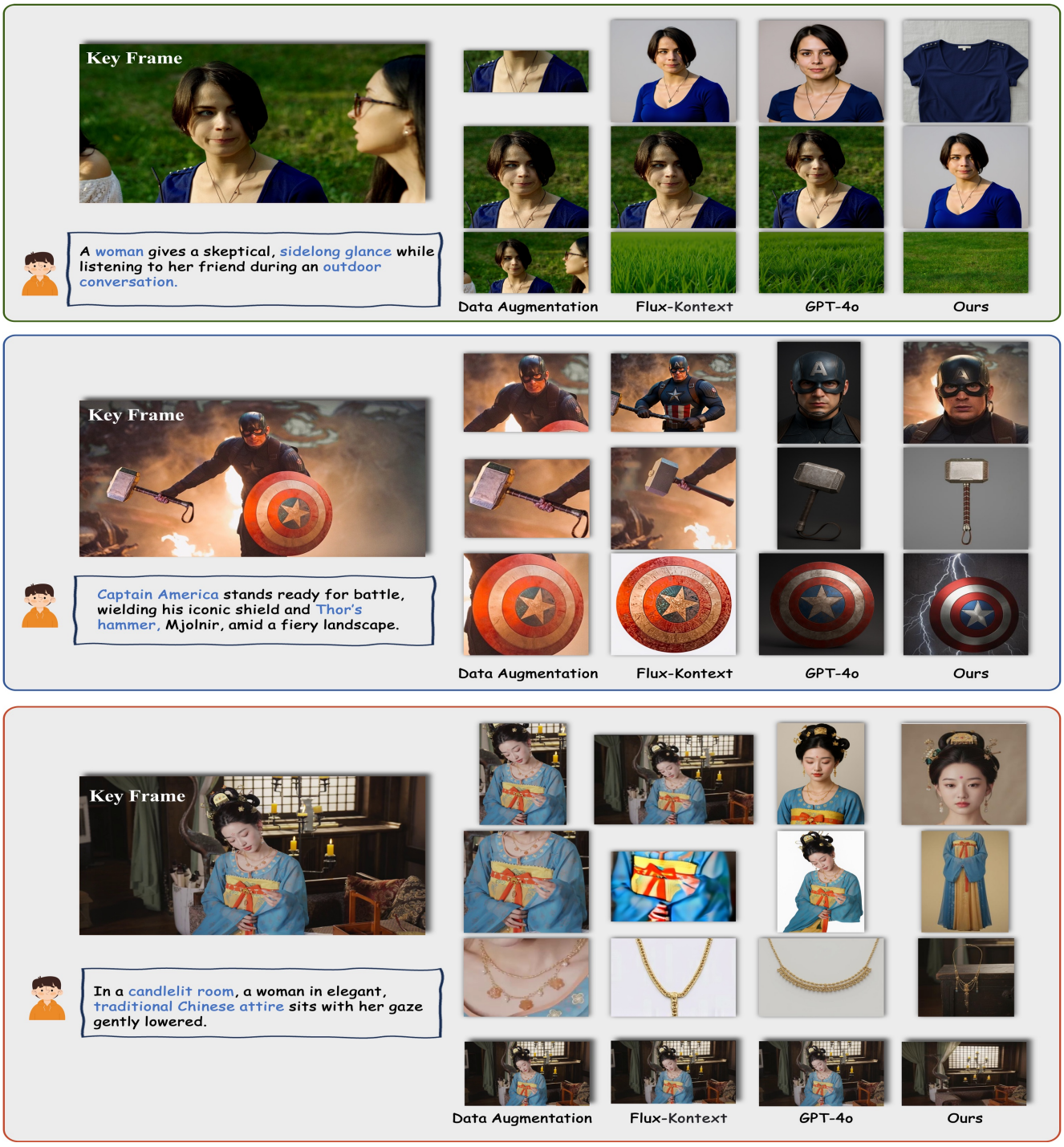

Figure S.5. **Comparison of reference image construction strategies.** In contrast to direct data augmentation or image editing models like Flux-Knotext [6] and GPT-4O [5], our data curation pipeline [1, 2] generates subject images with coherent multi-subject compositions.

compared to the baseline DiT architecture. The primary increase stems from the three additional attention operations per layer for the initial 20 layers. However, since these operations are performed on the relatively small subject token sets, the impact on overall training and inference time is minimal, representing a favorable trade-off for the significant gains in identity preservation. The GRPO post-training phase requires additional computational resources, but it is

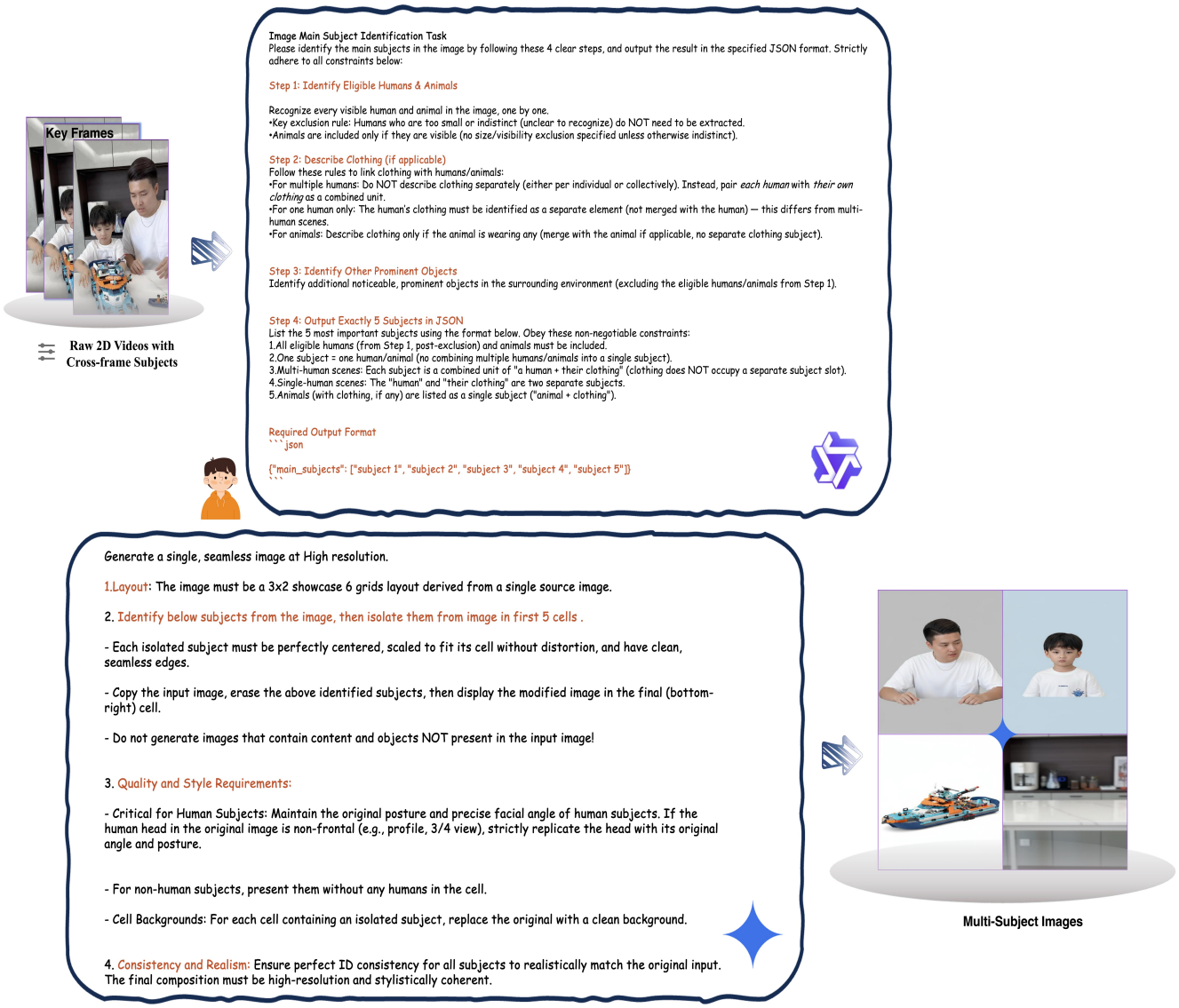

Figure S.6. Prompt for MLLM-based multi-subject images generation.

a one-time cost that substantially enhances model performance.

E.2. Ethics Statement

Our work focuses on advancing controllable video generation, a technology with significant creative potential. We acknowledge the potential for misuse, such as the creation of misleading or harmful content. To mitigate these risks, we have focused our dataset curation on non-realistic, generic subjects and actions, avoiding the use of real-world public figures. The generated videos often contain subtle artifacts that distinguish them from real footage. We advocate for the development of robust detection methods for synthetic media and support the establishment of ethical guidelines for the use of generative models. Our model and dataset will be released to the research community to facilitate further research in this area, and we encourage responsible use.

References

[1] 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. *arXiv preprint arXiv:2502.13923*, 2025. 1, 4, 6

[2] Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing Gemini 2.5 Flash Image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025. 3, 6

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[5] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3, 6

[6] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 3, 6

[7] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advanced in Neural Information Processing Systems (NeurIPS)*, 2025. 2

[8] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[9] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 3

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[11] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2

[12] PySceneDetect Development Team. Pyscenedetect: An open-source video scene detection tool. https://www.scenedetect.com/, 2024. 3

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[14] Team Wan. Wan: Open and advanced large-scale video generative models, 2025. 1

[15] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 3, 4

[16] Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *Advanced in Neural Information Processing Systems (NeurIPS)*, 2025. 3, 4