Character Mixing for Video Generation

Tingting Liao **Chongjian Ge** **Guangyi Liu** **Hao Li** **Yi Zhou**
Institute of Foundation Models, Mohamed bin Zayed University of Artificial Intelligence
Pinscreen

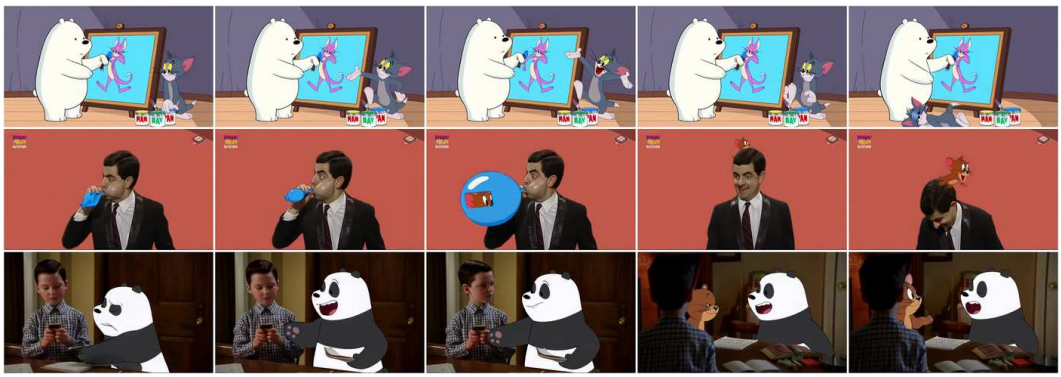


Figure 1: **Multi-character Mixing.** Our method preserves character identity, behavior, and original style while generating plausible interactions between characters that have never coexisted: from cartoons (*We Bare Bears*, *Tom and Jerry*) to realistic humans (*Mr. Bean*, *Young Sheldon*).

Abstract

Imagine Mr. Bean stepping into *Tom and Jerry*—can a video model generate interactions that stay faithful to each character's identity, behavior, and style across worlds? We present **MiMix**, a unified framework for multi-character text-to-video generation that goes beyond visual appearance to capture character-specific personas generalizing to unseen scenes and partners. To evaluate this, we introduce the **Cross-Character Generalization Benchmark**, a *social stress test* placing each character in novel scenes with previously uncoexistent partners. MiMix combines **Cross-Character Embedding (CCE)**, which disentangles identity and behavior via text-aligned annotations, with **Cross-Character Augmentation (CCA)**, which synthesizes cross-style training data while preserving native appearance. Experiments, including a blind user study, show consistent gains in identity fidelity, interaction quality, and style consistency over prior personalized, multi-reference, and general video models. Additional results and videos are available on our project page: https://tingtingliao.github.io/mimix.

1 Introduction

In an era where films and iconic characters are just a click away, a natural question arises: what if we could unite these beloved characters together, merging their roles and interactions into a single story?

Since the release of Sora [OpenAI, 2024b] by OpenAI, fundamental text-to-video (T2V) generation models [Yang et al., 2024, OpenAI, 2024b, Team Wan, 2025, Team, 2024b, DeepMind, 2024, Kong et al., 2024, kli, 2024] have achieved substantial progress in general video synthesis. For producing videos centered on specific or customized characters, a common approach is to leverage a reference

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

image as input. Such personalized video generation methods [He et al., 2024, Liang et al., 2025, Jiang et al., 2024, Wei et al., 2024, Fei et al., 2025, Chen et al., 2025b] allow users to create customized content with their own images. However, reference images alone provide visual similarity but fail to capture motion characteristics such as the character's distinctive behaviors and interactions with the environment, and others. Thus, while preserving the identity, these approaches often fail to faithfully capture the character's complex behavior. To this end, we want to develop a method that could allow the model to learn from all the footages and scripts of the characters of interest that can be collected, so as to learn not only their appearance but also motion idiosyncrasies and personality, and to generate vivid videos that maximally match their behavior traits and motion patterns.

To faithfully generate videos for a single character, one can fine-tune a text-to-video model on footage of that character. However, moving beyond single-character settings brings two major challenges.

The first is the **non-coexistence challenge**: characters from different shows never co-occur in any training video, leaving no paired data to model their joint interactions. To address this, we explicitly encode each character's identity and behavior into text by annotating their names and actions in the captions. This disentangles character-specific behavior embeddings from the underlying training videos, enabling us to fine-tune one foundation model on each character's individual footages while still allowing them to co-exist and interact naturally at inference time.

The second is the **style delusion challenge**: characters often originate from domains with drastically different visual styles, such as live-action sitcoms and cartoons, which never naturally co-exist in the same videos. Directly training on mixed styled data leads to unstable character styles, as shown in Figure 2, where the model tries to render all characters in a uniform style matching the training data.

We tackle this by introducing a style-aware data augmentation strategy that composites characters from different domains into the same video while preserving their native appearances. We find that even a small proportion of such augmented data substantially improves style preservation in cross-domain character mixing. For background ambiguities, we introduce an extra prompt for background style.



Figure 2: **Style delusion examples.** When mixing different style characters, their styles may shift undesirably. For instance, Mr. Bean looks cartoonish (top row), while Ice Bear appears realistic (bottom row).

To validate our approach, we curate a dataset of 374 episodes (22,478 video clips) featuring two cartoons (*Tom and Jerry*, *We Bare Bears*) and two realistic shows (*Young Sheldon*, *Mr. Bean*). Each clip is annotated with explicit character names and style information, supporting fine-grained control during training and inference. We further establish the first benchmark for multi-character video generation, evaluating identity preservation, motion fidelity, interaction realism, and style consistency.

Our contributions are summarized as follows:

- We proposed the first video generation framework for multi-character mixing that addresses both the non-coexistence challenge and the style delusion challenge.
- We curated a dataset of 374 episodes (22,478 video clips) with character- and style-annotated captions, enabling controllable multi-character video synthesis across domains.
- We conducted extensive experiments and introduced a benchmark, showing that our method significantly improves identity preservation, motion consistency, and interaction quality compared to prior art.

2 Related Works

Video Generation The advent of diffusion models has transformed video generation, advancing from early text-to-video systems such as ImagenVideo [Ho et al., 2022], Make-A-Video [Singer et al., 2022], and VideoLDM [Blattmann et al., 2023], to large-scale architectures like Sora [OpenAI, 2024b], Goku [Chen et al., 2025a], Wan2.1 [Team Wan, 2025] and HunyuanVideo [Kong et al., 2024], which achieve state-of-the-art general video synthesis. However, these models remain limited in generating content with specific identities or custom subjects, motivating the emerging direction of



Figure 3: **Diverse character interactions generated by our method.** Per-row composition and per-example prompts are listed in the supplementary.

personalized video generation, where reference visual signals guide the synthesis of videos with consistent appearances and motion dynamics.

Single-Concept Personalization Early methods such as DreamVideo [Wei et al., 2024], Magic-Me [Ma et al., 2024], and PersonalVideo [Li et al., 2024] customize videos with per-subject tuning, achieving identity preservation but requiring costly optimization. More recent zero-shot approaches like ID-Animator [He et al., 2024] leverage facial adapters to enable identity-consistent generation from a single reference image without fine-tuning. Despite these advances, existing methods largely emphasize visual similarity while overlooking motion-related aspects such as unique behaviors and environment interactions.

Multi-Concept Customization Compared to single-concept personalization, multi-concept customization is more challenging due to identity blending, where multiple characters risk being fused into a composite scene. Video Alchemist [Chen et al., 2025b] addresses this through cross-attention–based fusion of text and image representations for open-set subject and background control, while Movie Weaver [Liang et al., 2025] employs tuning-free anchored prompts to preserve distinct identities. Other approaches, such as CustomVideo [Wang et al., 2024] and Custom Diffusion [Kumari et al., 2023], explore parameter-efficient fine-tuning and joint optimization for multi-subject composition. Recent large-scale reference-to-video models, including VACE [Jiang et al., 2025], Phantom [Liu et al., 2025], SkyReels-A2 [Fei et al., 2025], and OmniVCus [Cai et al., 2025], condition a video diffusion transformer on multiple subject images to compose several identities in one clip. However, existing image-guided methods cannot leverage video and textual data during training, limiting their ability to model realistic interactions and dynamics across characters.

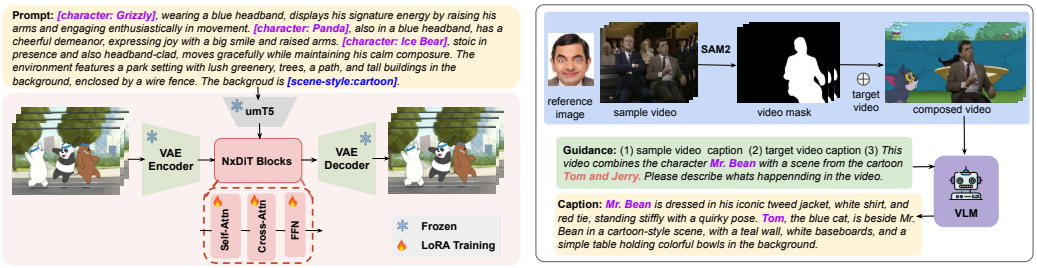


Figure 4: **Model finetuning architecture** (left) and **data augmentation pipeline** (right).

3 Methods

The goal of our method, which we call **MiMix**, is to learn the essence of characters from large collections of video data and enable them to interact seamlessly in new, mixed contexts. Given the abundance of video series—spanning cartoons and live-action shows, our approach seeks to (1) capture characters' unique identity and behavioral traits, and (2) enable flexible mixing of characters across styles and universes.

Our curated dataset (Section 3.3) consists of TV shows and animations. We pair each video clip with a structured caption that explicitly grounds character identity and disentangles it from scene context, providing crucial cues about each character's appearance and behavioral logic. Each domain features one or multiple central characters, sometimes appearing in ensembles (e.g., Tom and Jerry), other times in isolation (e.g., Mr. Bean).

In this section, we develop a novel training scheme that introduces **Cross-Character Embedding (CCE)** and **Cross-Character Augmentation (CCA)** to achieve robust identity modeling, behavior preservation, and style-controllable mixing.

Preliminaries. MiMix is instantiated on Wan2.1-T2V-14B [Team Wan, 2025], a latent diffusion-transformer (DiT) [Peebles and Xie, 2023] text-to-video model with three components. (i) A causal 3D VAE compresses an RGB video by $4\times$ temporally and 8×8 spatially into a 16-channel latent sequence; each 10-second, 161-frame training clip at 16 fps therefore maps to 41 latent frames. (ii) A umT5 [Chung et al., 2023] text encoder maps the prompt to a token-embedding sequence that conditions generation through the cross-attention layer of every DiT block. (iii) The DiT is trained with a flow-matching [Lipman et al., 2023] denoising objective in this latent space. MiMix leaves the architecture unchanged: all of our contributions enter through the conditioning and data pathway. Stage 1 (CCE) fully fine-tunes the DiT under the structured character–action captions; Stage 2 (CCA) attaches rank-32 LoRA [Hu et al., 2022] adapters to the query/key/value/output projections of the attention layers and the first feed-forward projection of each DiT block, trained on a mixture of original clips and cross-style composites (Section 3.3).

3.1 Cross-Character Embedding (CCE)

Faithfully reproducing a character requires learning from dynamic data that captures not only appearance but also behavior, idiosyncratic motion patterns, and contextual habits. However, images are insufficient because they omit the motion and interaction cues that define a character's identity.

We therefore design a framework to learn the character concept embeddings across different domains. We need to tackle character disentanglement in multi-character shows (e.g., *Tom and Jerry*, *We Bare Bears*), where multiple identities must be separated within the same clip. We also face the challenge of **non-coexistence** of characters from different universes, who never appear together in the training data but must interact coherently at inference.

Character–Action Prompting. Our key insight is to design a character–action captioning format that explicitly grounds each character's identity while separating it from scene context. Unlike standard captions that describe only visual events, our captions use a tagged structure that pairs each character with their in-clip appearance and reports the scene-level action separately:

[character]<name>:<appearance>**[/character]**
[character]<name>:<appearance>**[/character]**...
[action]<scene action>**[/action]**

The per-character *appearance* field captures clothing, posture, and held objects *as visible in this clip* (e.g., *wearing a striped shirt and bow tie, holding a suitcase*), keeping a stable canonical name as the identity anchor while letting visual details vary across clips. The *action* field describes what all visible characters are doing at the scene level, including the setting.

This design ensures that embeddings encode characters as independent entities with disentangled actions and identities. During inference, the same prompting scheme enables coherent composition of characters who never coexisted in training—for example, Mr. Bean interacting with Tom and Jerry. Although Mr. Bean and Jerry never meet each other in the dataset, the model has observed how each interacts with others in their respective domains, allowing it to generalize to cross-universe interactions. A more detailed illustration is provided in Figure 4.

Prompt Generation. We employ GPT-4o [OpenAI, 2024a] to automatically generate captions. For each clip we provide: 1) 8 frames uniformly sampled across the clip span as visual context, 2) source metadata (animated cartoon vs. live-action sitcom), and 3) a per-show character identification list mapping each canonical name to a short physical descriptor (e.g., *"Sheldon Cooper: a young boy with short brown hair"*). The character list constrains GPT-4o to output only canonical names—eliminating vague phrasings such as "a man" or "a cat"—and is paired with an alias table that normalizes informal references (e.g., *Grizz* → *Grizzly*, *George* → *George Sr.*) to a single canonical form. This setup enables GPT-4o to reliably identify character names while minimizing hallucinations. The resulting process yields 22,478 clips and 21,226 valid captions across 374 episodes; the remainder are GPT-4o safety refusals or empty-character shots and are dropped before training. Each character mention is annotated with `[character]name:appearance[/character]` tags, which act as identity anchors and support controllable character-level generation. More details are provided in the supplemental.

Model Adaptation. We fine-tune Wan2.1-T2V-14B [Team Wan, 2025] on the curated text–video pairs via full-parameter fine-tuning (training recipe in Section 3.3). Our captioning and augmentation strategies introduce no backbone-specific architectural module and are therefore potentially portable to other text-to-video models. As a feasibility check, the character-tagged captioning recipe also produces cross-series character mixes on the LTX-2.3 audio–video backbone [HaCohen et al., 2026] (supplementary, Section D.3); validating the complete two-stage pipeline across diverse backbones remains future work. The curated text–video pairs capture each character's actions, emotions, and behaviors, enabling the learned embeddings to serve as flexible building blocks for multi-character generation.

3.2 Cross-Character Augmentation (CCA)

While CCE ensures characters act authentically, training across mixed domains (cartoon and live-action) introduces a **style delusion problem**: characters may drift into unintended visual styles (e.g., *Mr. Bean* rendered as a cartoon, or Ice Bear appearing overly realistic), as shown in Figure 2.

To preserve original styles while allowing cross-style interactions, we introduce **Cross-Character Augmentation (CCA)**. This tackles a second non-coexistence challenge: in the training data, cartoon and real characters never appear together, and neither do their backgrounds.

Synthetic Cross-Domain Compositing. Our intuition is that even imperfect synthetic co-occurrences can effectively guide the model toward more consistent, style-preserving generation. We therefore create augmented training clips by segmenting characters from source videos and pasting them into backgrounds from the opposite style domain. For example, Mr.Bean (live-action) may be placed in a cartoon *Tom and Jerry*scene as shown in the right part of Figure 4.

Characters are segmented using SAM2 [Ravi et al., 2024], which handles both live-action and animation content. To ensure the quality and relevance of the synthetic compositions, we apply different filtering strategies. For live-action characters such as *Mr. Bean*and *Young Sheldon*, we provide one or more curated reference images per target identity and extract normalized face embeddings with InsightFace [Deng et al., 2019]. Candidate clips are retained only when a sampled frame contains a face whose embedding matches the reference by cosine similarity. To initialize segmentation, a

YOLO person proposal is verified against the reference identity, and only the matched box is passed to SAM2 for mask propagation. For cartoon characters, where face embeddings are unreliable, we use Gemini [Team, 2024a] for character-presence detection and filtering; segmentation remains detector-guided SAM2.

The composited clips are then captioned with GPT-4o, which receives both the background source and the inserted character identities. Each caption is further enriched with an explicit style tag (`[scene-style]cartoon[/scene-style]` or `[scene-style]real[/scene-style]`), providing the model with clear supervision for style control. The complete caption format becomes:

[character]<name>:<appearance>**[/character]**
[character]<name>:<appearance>**[/character]**...
[action]<scene action>**[/action]**
[scene-style]<style>**[/scene-style]**

Empirical Findings. We observe that a **small proportion** of such augmented clips suffices to unlock robust cross-style composition. Excessive synthetic data, however, degrades realism and harms overall video quality. A detailed analysis is provided in the experiments section.

3.3 Training and Data

We adopt a two-stage fine-tuning recipe. **Stage 1 (Cross-Character Embedding).** We fully fine-tune Wan2.1-T2V-14B [Team Wan, 2025] on the four-show curated text–video pairs for $5{,}000$ steps with Adam (learning rate 10^{-5}, batch size 64) on 72 NVIDIA A100 GPUs ($\approx$ 57 hours). This stage grounds each character as a distinct concept under the character–action prompting format. **Stage 2 (Cross-Character Augmentation).** Starting from the Stage-1 checkpoint, we apply rank-32 LoRA fine-tuning [Hu et al., 2022] on a mixture of the original four-show data and the synthetic cross-style composites (Section 3.2) for $1{,}000$ steps with Adam (learning rate 2×10^{-5}, batch size 64) on 64 A100 GPUs ($\approx$ 12 hours). Mixed-precision (FP16) training and gradient clipping are used throughout both stages.

Scenes and segments. We curate a dataset comprising two cartoons and two live-action shows: 78 episodes of *Tom and Jerry*, 137 episodes of *We Bare Bears*, 18 episodes of *Mr. Bean*, and 141 episodes of *Young Sheldon*, for a total of 374 episodes. Raw videos are first preprocessed to crop out bottom text overlays (e.g., subtitles, network bugs) so the model is not biased by spurious language cues. We then run a three-stage curation pipeline:

(i) Scene detection. We apply PySceneDetect's `ContentDetector` (threshold 40, minimum scene length 10 s) to every episode to obtain shot boundaries.

(ii) Intro/outro filtering. For each show, we automatically detect the repeated opening and closing credit segments by sampling one frame per second across all episodes, computing perceptual (dHash) signatures, and trimming time offsets where a dominant visual cluster is shared by at least half of the episodes. For shows with variable cold opens (e.g., *Young Sheldon*), we additionally use GPT-4o to identify the latest opening-credit frame and override the detected boundary.

(iii) Fixed-length chunking. Within the remaining scene spans, we cut non-overlapping 10-second clips at 16 fps (161 frames per clip, matching the Wan2.1 input convention). The pipeline yields 22,478 clips (*Tom and Jerry*: 2,809; *We Bare Bears*: 8,187; *Mr. Bean*: 1,372; *Young Sheldon*: 10,110), of which 21,226 receive valid captions in the prompt-generation stage above.

For each domain, we define the set of key characters: *Mr. Bean* (`Mr. Bean`), *Tom and Jerry* (`Tom`, `Jerry`, `Spike`), *We Bare Bears* (`Ice Bear`, `Grizzly`, `Panda`), and *Young Sheldon* (`Sheldon Cooper`, `Missy`, `Mary Cooper`, `George Sr.`).

Prompt: ***Tom*** *and* ***Ice Bear*** *get jobs at a bakery.* ***Tom*** *keeps chasing* ***Jerry*** *through cake trays, and* ***Ice Bear*** *calmly decorates a three-tier wedding cake. The bride ends up choosing* ***Ice Bear****'s version over the original.*



Prompt: ***Mary Cooper*** *and* ***Panda*** *host a tea party.* ***Mary*** *brings her finest china and proper etiquette.* ***Panda*** *makes cute cupcakes with bear faces.*



Figure 5: **Comparison on multi-subject interaction.** Results from SkyReels-A2 [Fei et al., 2025] (top row) and ours (bottom row).

Prompt: ***Panda*** *is at a karaoke bar, singing loudly with his brothers, closing his eyes as if he were a superstar on TV.*



Figure 6: **Comparison on single-subject generation.** From top to bottom: Wan2.1-I2V [Team Wan, 2025], SkyReels-A2 [Fei et al., 2025], and ours. Comparisons against VideoBooth [Jiang et al., 2024] and DreamVideo [Wei et al., 2024] are deferred to the supplementary.

4 Experiments

4.1 Benchmarks

We propose the **Cross-Character Generalization Benchmark**, a social stress test that places each character in novel scenes paired with previously uncoexistent partners (e.g., *Mr. Bean* meeting *Tom and Jerry*). We evaluate our method using a comprehensive set of metrics along two dimensions. For overall video quality and temporal coherence, we adopt Consistency, Motion, Dynamic, Quality, and Aesthetic from VBench [Huang et al., 2024]. To assess character-level consistency and interaction, we employ a vision–language model (VLM) and introduce four specialized metrics: Identity Fidelity (ID-FID), Motion Authenticity (MOT-AUTH), Style Coherence (STY-COH), and Interaction Plausibility (INT-PLAUS). Specifically, we leverage Gemini-1.5-Flash [Team, 2024a] as the VLM backbone for these evaluations.

Video Quality and Temporal Consistency. (1) **Consistency** evaluates the overall video-text consistency across frames computed by ViCLIP [Wang et al., 2023]. (2) **Motion** measures the level of smoothness of generated motions. (3) **Dynamic** quantifies the degree of motion dynamics using RAFT [Teed and Deng, 2020]. (4) **Quality** measures the imaging quality, referring to the distortion

(e.g., over-exposure, noise, blur) by the image quality predictor MUSIQ [Ke et al., 2021]. (5) **Aesthetic** evaluates the artistic and beauty value perceived by humans towards each video frame using the LAION aesthetic predictor.

Character Consistency and Interaction. (1) **Identity Fidelity (ID-FID)** evaluates how well the generated video preserves each character's visual identity and distinctive features. The VLM assesses facial feature consistency, body proportions, characteristic attributes (e.g., Jerry's mouse ears, Tom's whiskers), and overall color scheme. A score of 10 indicates perfect identity preservation, where the character is immediately recognizable; a score of 1 indicates the character is completely unrecognizable. (2) **Motion Authenticity (MOT-AUTH)** measures the authenticity of character-specific movements and behaviors relative to their canonical personality traits. The evaluation considers motion patterns (e.g., Jerry's quick scurrying, Tom's exaggerated sneaking), behavioral consistency, and expression of personality through movement. The VLM analyzes temporal sequences to assess alignment with the character's known behavior patterns. (3) **Style Coherence (STY-COH)** assesses the consistency of each character's original artistic and visual style. This includes animation style (e.g., cartoon vs. realistic), aesthetic coherence with the source material, art direction fidelity, and rendering consistency. The VLM compares the generated video to its learned representation of the character's canonical appearance and stylistic conventions. (4) **Interaction Plausibility (INT-PLAUS)** evaluates the naturalness and plausibility of multi-character interactions. The assessment considers spatial relationships, timing and coordination, believability of reactions and responses, and physical dynamics between characters. For single-character videos, this metric evaluates interactions with the environment and scene elements.

4.2 Comparison

We benchmark our method against state-of-the-art video generation baselines, including two single-subject customization approaches—VideoBooth [Jiang et al., 2024] and DreamVideo [Wei et al., 2024]—as well as foundation image-to-video models Wan2.1-I2V [Team Wan, 2025] and SkyReels-A2 [Fei et al., 2025], both of which support single- and multi-character generation. For multi-subject customization specifically, we compare directly against SkyReels-A2. Note that Wan2.1-T2V cannot directly generate videos from a character image. Thus, we first employ OmniGen [Xiao et al., 2025] to synthesize an image using the prompt and reference, which is then used as input to Wan2.1-I2V.

For single-subject evaluation, we generate 50 videos featuring 10 characters—five from cartoons (`Tom`, `Jerry`, `Grizzly`, `Ice Bear`, `Panda`) and five from live-action series (`Mr. Bean`, `Sheldon`, `George`, `Mary`, `Missy`). For multi-subject evaluation, we generate 50 videos, each featuring 2–3 characters interacting within the same scene (noting that SkyReels-A2 supports fewer than three characters). These interactions span a wide range of scenarios, including inter-style (cartoon with real-life), intra-style (within cartoons or within real-life), inter-series (across different shows), and intra-series (within the same show). All reference images are included in the Appendix.

Figure 6 shows the qualitative comparison on single-subject generation against the two strongest baselines, Wan2.1-I2V [Team Wan, 2025] and SkyReels-A2 [Fei et al., 2025]; the comparisons against VideoBooth [Jiang et al., 2024] and DreamVideo [Wei et al., 2024], which fail to preserve the visual identity of the reference character at all, are deferred to the supplementary (Figure 18). Wan2.1-I2V and SkyReels-A2 retain identity to some extent—though they struggle with facial details—but fail to synthesize character-faithful motions. In contrast, our method consistently preserves visual fidelity and generates character-faithful motion.

Figure 5 presents qualitative comparisons on multi-character interaction, both within the same style and across different styles. While SkyReels-A2 [Fei et al., 2025] can synthesize acceptable single-subject videos, it struggles with complex interactions across multiple characters, especially in multi-style settings. Although it can place several characters into a shared scene, the results often exhibit visual inconsistencies and unnatural interactions. In contrast, our method enables contextually coherent interactions without compromising character identity or native style.

Quantitative results across nine metrics are reported in Table 1. Our method consistently outperforms prior approaches in both single- and multi-subject settings, demonstrating stronger identity preservation, faithful motion synthesis, and coherent style maintenance across diverse interaction scenarios. Additional comparison results are provided in the Appendix. Figure 3 surveys seven further interactions our model produces (one per row, five frames each), pairing characters across



Figure 7: **Effect of training scheme.** Top: Stage 1 alone (full-parameter fine-tuning on the four-show data) collapses both characters to a single cartoon style. Middle: Stage 2 alone (rank-32 LoRA trained directly on the base model, two-show data, 2,000 steps) already separates the styles—cartoon Spike next to live-action Mr. Bean—but grounds identities less precisely (a less faithful Spike and a duplicate Bean-like person). Bottom: our full two-stage pipeline preserves both identities and their native styles.

Table 1: **Comparison of recent video generation models across evaluation dimensions.** The first group of columns reports VBench metrics; the last four report VLM-based character-level scores (1–10). **Bold** indicates the best performance per column.

		VBench Metrics [Huang et al., 2024]					**VLM Metrics**			
Methods	**Subject**	Consistency	Motion	Dynamic	Quality	Aesthetic	ID-Fid	Mot-Auth	Sty-Coh	Int-Plaus
VideoBooth [Jiang et al., 2024]	Single	0.129	0.978	0.509	0.641	0.490	4.45	3.72	5.43	4.44
DreamVideo [Wei et al., 2024]	Single	0.185	0.956	–	0.627	0.500	4.51	4.16	6.82	5.37
Wan2.1-I2V [Team Wan, 2025]	Single	0.068	0.983	0.653	0.719	0.586	5.27	5.10	7.94	6.41
SkyReels-A2 [Fei et al., 2025]	Single	0.147	0.978	0.784	**0.723**	0.585	**6.17**	4.55	7.82	6.78
Ours	Single	**0.189**	**0.984**	**1.000**	0.576	**0.597**	6.12	**5.41**	**8.06**	**7.24**
SkyReels-A2 [Fei et al., 2025]	Multiple	0.131	0.965	0.979	**0.714**	0.537	6.17	4.55	6.28	4.94
Ours	Multiple	**0.183**	**0.984**	**0.986**	0.686	**0.581**	**6.48**	**5.50**	**7.26**	**5.22**

Table 2: **Comparison with multi-reference video generators on 40 cross-character prompts.** *Left:* VLM scores (1–10) as *Gemini / Qwen3-VL*; **bold**: best per judge. *Right:* blind user study on 10 of the prompts (11 participants, 110 judgements); vote share with 95% Wilson intervals, chance = 25%.

	VLM metrics (Gemini / Qwen3-VL)				**User study** (vote share, %)			
Methods	ID-Fid	Mot-Auth	Sty-Coh	Int-Plaus	Identity	Style	Interaction	Quality
VACE-14B [Jiang et al., 2025]	8.22 / 3.89	6.33 / **6.00**	8.00 / **7.33**	6.00 / 5.56	6.4 [3–13]	6.4 [3–13]	3.6 [1–9]	3.6 [1–9]
Phantom-Wan-14B [Liu et al., 2025]	8.89 / 4.56	5.44 / 5.22	7.44 / 6.89	**6.11 / 5.78**	19.1 [13–27]	18.2 [12–26]	21.8 [15–30]	18.2 [12–26]
SkyReels-A2 [Fei et al., 2025]	8.56 / 4.89	5.56 / 5.44	7.56 / 6.44	5.00 / 5.33	17.3 [11–25]	16.4 [11–24]	11.8 [7–19]	13.6 [8–21]
Ours	**9.00 / 6.11**	**7.22** / 5.11	**8.11** / 7.22	5.44 / 5.56	**57.3** [48–66]	**59.1** [50–68]	**62.7** [53–71]	**64.5** [55–73]

shows—both realistic *Young Sheldon* characters with cartoon characters from *Tom and Jerry* and *We Bare Bears*, and realistic *Young Sheldon* characters with *Mr. Bean*—all of which never co-occur in the original training videos; full per-row composition is described in the supplementary.

Multi-reference baselines and user study. We further compare against three recent multi-reference subject-to-video generators, VACE-14B [Jiang et al., 2025], Phantom-Wan-14B [Liu et al., 2025], and SkyReels-A2 [Fei et al., 2025] (official checkpoints and settings), on 40 cross-character prompts spanning all four shows and all style pairings. Baselines receive one reference image per character; ours receives only the prompt. We additionally re-score all videos with the open-weight Qwen3-VL-30B-A3B [Bai et al., 2025] and run a blind four-way user study on 10 of the prompts (11 participants, randomized order, method names hidden). As shown in Table 2, our method attains the highest Identity Fidelity under both judges and is preferred by users on all four criteria with non-overlapping confidence intervals; the preference is strongest on live-action and mixed pairs, while for cartoon pairs from famous franchises a reference image is often sufficient (details in supplementary Section D.1).

Table 3: **Effect of Synthetic Data Augmentation.** We vary the proportion of synthetic videos relative to the original dataset and evaluate across the same dimensions used in Table 1.

	VBench Metrics [Huang et al., 2024]					VLM Metrics			
Ratio	Subject-C	Background-C	Motion-S	Dynamic	Quality	ID-FID	MOT-AUTH	STY-COH	INT-PLAUS
5%	0.874	0.908	0.983	0.900	0.684	7.67	6.03	7.15	**4.83**
10%	**0.881**	**0.915**	**0.985**	0.950	**0.696**	**8.33**	6.72	**7.33**	4.78
20%	0.844	0.891	0.978	**1.000**	0.673	8.30	**7.10**	7.08	3.90

4.3 Ablation study

Training scheme. Figure 7 contrasts Stage 1 alone (full-parameter fine-tuning on the four-show data) with our full pipeline on a cross-style prompt (live-action *Mr. Bean* meeting cartoon Spike). Stage 1 alone learns both characters but collapses them to a single cartoon style—the style-delusion failure motivating Section 3.2—whereas our full pipeline keeps Mr. Bean live-action alongside cartoon Spike. We also train the Stage-2 LoRA directly on the base model (*Stage-2 only*; two shows, 2,000 steps; see supplementary). It already separates the styles (middle row) but grounds identities less precisely, supporting our decomposition: Stage 1 grounds identities, while Stage 2 provides cross-style mixing and can serve as a lightweight recipe on its own.

Augmentation ratio. We analyze how varying the ratio of synthetic augmentation data affects the mitigation of style delusion. We vary the proportion of synthetic composites relative to the original four-show data (Section 3.2) at 5%, 10%, and 20%. As summarized in Table 3, a moderate ratio (10%) maximizes the character-level metrics (best ID-FID, STY-COH, and overall video quality), while excess augmentation (20%) materially hurts interaction plausibility (INT-PLAUS drops from 4.78 to 3.90), confirming that synthetic cross-style data should remain a small fraction of the training mix.

Per-character data requirement. Over 15 characters with 55–5,509 training clips, solo ID-FID declines only from 7.00 (head, $\geq$ 1,265 clips) to 6.83 (464–956) and 6.25 (55–84); see supplementary Sections D.4–D.7 for details, unseen characters, and failure cases.

5 Discussion

Limitations. Introducing a new character requires curating footage and re-training; characters outside the cast can be composed through the base model's prior, but likeness-defined identities may drift. Specialization also costs general quality (VBench aesthetic -9.2% on 20 general prompts), pulls outputs toward a broadcast-TV look, and lets look-alike identities interfere (Mickey Mouse gets Jerry's face). Finally, scenes with five or more characters may omit some of them, and rare interaction verbs or untrained rendering styles fall back to the patterns and the two style registers of our data, although identities are preserved. All stem from the training distribution (Sections D.5–D.7).

Broader applications and future work. Cross-character content is a long-standing production practice (e.g., *Who Framed Roger Rabbit*, *Space Jam*, *Kingdom Hearts*) whose artist and rights-clearance costs scale poorly with the number of IPs; MiMix targets this pain point. Promising directions include running Stage 1 as a low-rank adaptation to reduce interference with the base model, cross-show multi-character and style-swap augmentation, a learned character-conditioning adapter [He et al., 2024] for open-world characters, and validating the full pipeline on other backbones beyond our LTX-2.3 feasibility study.

Ethics and copyright statement. Our copyrighted training content (*Tom and Jerry*, *We Bare Bears*, *Mr. Bean*, *Young Sheldon*) is used for non-commercial academic research only, under fair-use / text-and-data-mining exceptions. We do not redistribute the raw episodes; only derivative captions, metadata, and the processing pipeline are released under a non-commercial research-use license. Characters remain the property of their rights holders, and production deployment requires separate licensing. As these exceptions are jurisdiction-dependent, researchers reproducing this work should verify compliance with their local copyright law.

Acknowledgments and Disclosure of Funding

This work was supported by internal funding from the Institute of Foundation Models (IFM) and the Metaverse Center at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). The authors declare no competing interests.

References

Kling ai. Technical report, Kuaishou, 2024. URL https://klingai.kuaishou.com/.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

Yuanhao Cai, He Zhang, Xi Chen, Jinbo Xing, Yiwei Hu, Yuqian Zhou, Kai Zhang, Zhifei Zhang, Soo Ye Kim, Tianyu Wang, et al. Omnivcus: Feedforward subject-driven video customization with multimodal control conditions. *arXiv preprint arXiv:2506.23361*, 2025.

Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23516–23527, 2025a.

Tsai-Shien Chen et al. Multi-subject open-set personalization in video generation. 2025b.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. UniMax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *ICLR*, 2023.

Google DeepMind. Veo: Google's next-gen video generation model. https://deepmind.google/technologies/veo, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

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 *CVPR*, 2024.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *ICCV*, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

Weiming Kong, Yanghan Tuo, Jiangfeng Jia, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

Hengjia Li et al. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. *arXiv preprint arXiv:2411.17048*, 2024.

Feng Liang et al. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *ICCV*, pages 14951–14961, 2025.

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

OpenAI. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/, August 2024a. Accessed: 2025-09-10.

OpenAI. Video generation models as world simulators. Technical report, OpenAI, 2024b. URL https://openai.com/index/video-generation-models-as-world-simulators/.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Gemini Team. Gemini: A family of highly capable multimodal models, 2024a.

Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024b.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Zhao Wang, Aoxue Li, Enze Xie, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *CVPR*, 2024.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025.

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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions claimed in the abstract and Section 1 (the Cross-Character Embedding (CCE) and Cross-Character Augmentation (CCA) framework, the Cross-Character Generalization Benchmark, and the curated multi-show dataset) are described in Section 3 and validated by the experiments and ablations in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses the main limitations: introducing a new character requires curating footage and re-training; specializing on four shows measurably reduces general generation quality and causes identity interference between visually similar characters (quantified with VBench in the supplementary); and scenes with five or more characters, interaction verbs far from the training data, and rendering styles absent from training remain failure modes. The supplementary additionally reports per-character data requirements and generalization to unseen characters.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical framework for multi-character video generation; it does not include formal theorems or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 describes the full two-stage training recipe (Stage 1: full-parameter fine-tuning of Wan2.1-T2V-14B for $5{,}000$ steps with Adam at learning rate 10^{-5}, batch size 64; Stage 2: rank-32 LoRA fine-tuning for $1{,}000$ steps with Adam at learning rate 2×10^{-5}, batch size 64; FP16 and gradient clipping in both stages) and the three-stage data curation pipeline (PySceneDetect with threshold 40 and minimum scene length 10 s; perceptual-hash intro/outro filtering; 10 s clips at 16 fps yielding 161-frame inputs that match the Wan2.1 convention; GPT-4o captioning with 8 uniformly sampled frames per clip and a per-show character identification list). The shared GPT-4o instruction template and the per-show character rosters are reproduced in the supplementary.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the data-processing pipeline (scene detection, intro/outro filtering, clip chunking, GPT-4o captioning, and Wan2.1 VAE/T5 feature encoding) along with the generated metadata and captions; the project page (https://tingtingliao.github.io/mimix) links to qualitative results and code. Raw episodes of *Tom and Jerry*, *We Bare Bears*, *Mr. Bean*, and *Young Sheldon*are not redistributed because they are copyrighted television content; the supplementary describes how to obtain the source videos for non-commercial research collaboration. Note that we do not collect private information, infer sensitive attributes, perform biometric identification, or target non-public individuals.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training hyperparameters (two-stage recipe: Stage 1 full-parameter fine-tuning of Wan2.1-T2V-14B at learning rate 10^{-5} for $5{,}000$ steps; Stage 2 rank-32 LoRA at learning rate $2{\times}10^{-5}$ for $1{,}000$ steps; batch size 64; Adam; FP16; gradient clipping) and dataset

details (374 episodes / 22,478 clips, 10 s clips at 16 fps with 161 frames per clip) are reported in Section 3. The evaluation set (50 single-subject and 50 multi-subject prompts spanning intra-/inter-style and intra-/inter-series scenarios) and metric definitions are reported in Section 4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The main benchmark (Table 1, Table 3) reports point estimates on a fixed set of 50 single-subject and 50 multi-subject prompts without error bars, as re-running each baseline with multiple seeds was prohibitively expensive for 14B-scale video diffusion models. For the user study (Table 2) we report 95% Wilson score confidence intervals on vote shares, and we verify the VLM-based ranking with a second, open-weight judge (Qwen3-VL).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 3 reports both the algorithmic configuration (two-stage fine-tuning of Wan2.1-T2V-14B: Stage 1 full-parameter fine-tuning with Adam at lr 10^{-5} for 5,000 steps, batch size 64; Stage 2 rank-32 LoRA with Adam at lr 2×10^{-5} for 1,000 steps, batch size 64; FP16) and the hardware budget: Stage 1 on 72 NVIDIA A100 GPUs for $\approx$ 57 hours, Stage 2 on 64 A100 GPUs for $\approx$ 12 hours.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The work uses publicly broadcast television content for non-commercial research only and does not redistribute it (see the Ethics and copyright statement in Section 5). The user study collected only anonymous preference judgements on generated videos and no personal data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: *Positive impact:* the framework enables creative applications such as fan-style storytelling, animation prototyping, and education by lowering the barrier to character-faithful video generation. *Negative impact:* like all generative video models, the technology could be misused for unauthorized impersonation or disinformation involving fictional or real personas; we mitigate this by training only on copyrighted fictional characters and not redistributing the raw video corpus. The Discussion (Section 5) further notes the requirement of LoRA fine-tuning per character, which itself acts as a soft barrier to indiscriminate impersonation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We do not redistribute the raw television episodes (which are copyrighted) and we release the dataset metadata, captions, and processing code only for non-commercial research. The character-specific LoRA weights are tied to fictional characters from licensed shows, so the released artefacts cannot be used to impersonate identifiable real individuals out-of-the-box.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All borrowed components are cited with their source publications: Wan2.1 [Team Wan, 2025], SkyReels-A2 [Fei et al., 2025], VACE [Jiang et al., 2025], Phantom [Liu et al., 2025], VideoBooth [Jiang et al., 2024], DreamVideo [Wei et al., 2024], OmniGen [Xiao et al., 2025], LTX-2 [HaCohen et al., 2026], SAM2 [Ravi et al., 2024], InsightFace [Deng et al., 2019], GPT-4o [OpenAI, 2024a], Gemini [Team, 2024a], Qwen3-VL [Bai et al., 2025], LoRA [Hu et al., 2022], VBench [Huang et al., 2024], ViCLIP [Wang et al., 2023], RAFT [Teed and Deng, 2020], and MUSIQ [Ke et al., 2021]. Each is used in accordance with its respective license. The TV-show source material is copyrighted; we use it only for non-commercial research and do not redistribute the raw videos.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets (the curated multi-show dataset of 374 episodes and 22,478 clips, the four-stage data-processing pipeline, the LoRA-finetuned MiMix models, and the Cross-Character Generalization Benchmark) are documented in Section 3, Section 4, and the supplementary. The dataset README, the per-show character roster, the GPT-4o captioning template, and the evaluation prompt set are released alongside the code under a non-commercial research-use license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We conducted a blind user study (Section 4 and supplementary Section D.1). The supplementary reproduces the full instructions shown to participants (pick one of four videos per question, given the character reference images and the prompt) and the four questions (identity, style, interaction, overall quality). Participants were recruited outside the project and participation was anonymous.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study involved only anonymous preference judgements on generated, non-offensive video clips, collected no personal or identifying information, and posed no foreseeable risk to participants.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: GPT-4o is used as a core component of the data pipeline to generate the structured `[character]/[action]/[scene-style]` captions; the full instruction template and per-show character rosters are reproduced in the supplementary. Gemini-1.5-Flash is used as the VLM evaluator for the four character-level metrics (ID-FID, MOT-AUTH, STY-COH, and INT-PLAUS) defined in Section 4. Qwen3-VL-30B-A3B-Instruct is used as a second, open-weight VLM judge to verify the robustness of these metrics. GPT-4 is also used for grammar and language polishing of the manuscript, as declared in the supplementary "Use of LLM" section.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Supplementary Material

A Use of LLM

We used a large language model (GPT-4) to assist with grammar correction and language refinement throughout the writing process. Specifically, the model was employed to polish sentence structure, improve clarity, and ensure consistency in technical terminology. No content or experimental results were generated by the model; all scientific contributions and analysis were conducted by the authors.

B Dataset

Character Distribution. Figure 8 visualizes the per-character occurrence frequencies across all curated clips. Protagonist-centered shows (*Mr. Bean*, *Young Sheldon*) maintain high main-character presence, while ensemble shows (*We Bare Bears*) and classic animation (*Tom and Jerry*) display more balanced distributions across their main casts.

Captioning Prompt to GPT-4o. We provide GPT-4o with 8 frames uniformly sampled across each clip, together with show metadata and a per-show character identification list, along with carefully designed instructions to ensure precise and consistent character references. The shared instruction template and the per-show metadata (`movie_name`, `show_kind`, `scene_style`, and the canonical character roster) are shown below.

C Experiments

VLM evaluation details. The four character-level metrics defined in Section 4 of the main paper (ID-FID, MOT-AUTH, STY-COH, INT-PLAUS) are produced by Gemini-1.5-Flash [Team, 2024a]. Each generated clip is uploaded to the Gemini Files API in full—the model receives the whole video, not pre-sampled frames—and the four metrics are queried in four separate API calls with metric-specific prompts; we use `temperature` $= 0.3$, `top_p` $= 0.95$, `top_k` $= 40$, `max_output_tokens` $= 1024$, with a 2 s rate-limiting delay between calls.

Each prompt asks Gemini to return an integer score on a 1–10 scale together with a short free-text justification. Score anchors at the extremes are stated explicitly in the prompt (e.g. "10 = perfect identity preservation; 1 = character unrecognizable"). The cues each metric prompt asks the model to consider are:

- ID-FID (*identity preservation*): facial features, body proportions, distinctive characteristic attributes (e.g. Jerry's mouse ears, Tom's whiskers), and overall colour scheme.
- MOT-AUTH (*character-faithful motion*): motion patterns typical of the character, behavioural traits (e.g. Jerry's quick scurrying, Tom's exaggerated sneaking), personality expression through movement, and consistency with known mannerisms.

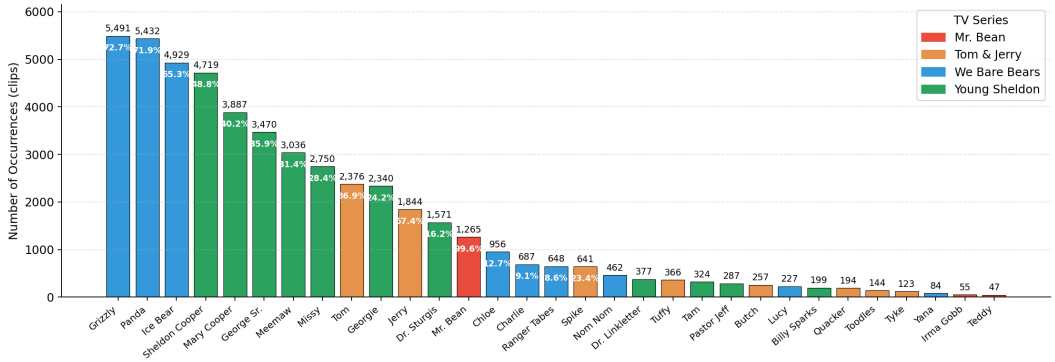


Figure 8: **Character distribution in our dataset.** The bar chart shows the number of appearances of each character in the video-caption pairs.

- STY-COH (*style preservation*): animation style (cartoon vs. realistic), aesthetic match with the source material, art-direction consistency, and rendering style appropriate to the character.
- INT-PLAUS (*interaction naturalness*): for multi-character clips, spatial relationships between characters, timing and coordination, believability of reactions, and physical plausibility; for single-character clips the prompt is rephrased to evaluate interaction with the environment.

The per-metric numbers reported in Tables 1 and 2 of the main paper are arithmetic means of these per-clip 1–10 scores across the 50 single-subject and 50 multi-subject evaluation videos. Internally we also compute a single weighted composite score with weights $0.30/0.30/0.25/0.15$ for ID-FID/MOT-AUTH/STY-COH/INT-PLAUS; only the per-metric averages appear in the main paper.

Reference images. We provide 10 reference images together with their corresponding masks for evaluating VideoBooth [Jiang et al., 2024], DreamVideo [Wei et al., 2024], Wan2.1 [Team Wan, 2025], and SkyReels-A2 [Fei et al., 2025], as illustrated in Figure 16.

Single-Subject Comparison. We provide additional qualitative comparisons with VideoBooth [Jiang et al., 2024], DreamVideo [Wei et al., 2024], Wan2.1 [Team Wan, 2025], and SkyReels-A2 [Fei et al., 2025] in Figure 18, 19, 20, and 17. These examples further demonstrate our method's ability to preserve subject identity and produce coherent motion across diverse scenarios.

Multi-Subject Comparison. We present additional qualitative comparisons with SkyReels-A2 [Fei et al., 2025] in Figure 15. The results demonstrate our method's capability to synthesize natural and coherent interactions among multiple characters while faithfully preserving their identities, characteristic behaviors, and distinctive visual styles. In comparison to SkyReels-A2, our approach yields fewer style inconsistencies and artifacts, producing more realistic and engaging multi-subject generations.

D Additional Experiments

Because Gemini-1.5-Flash, the VLM judge used for Tables 1–2 of the main paper, has been retired by its provider, the new VLM scores in this section are produced by its successor in the Gemini Flash family with otherwise identical metric prompts and score anchors (Section C); all methods within each study are scored by the same judge, so within-study comparisons are unaffected.

D.1 Multi-reference baselines, open-source judge, and user study

Protocol. We compare against three open-source multi-reference subject-to-video methods, each with its officially released 14B-scale checkpoint and default inference settings (Table 4). All baselines receive identical inputs per prompt: one solo reference frame per character extracted from the source shows, and the same natural-language prompt (scene action, per-character appearance description, and target style). Our model receives the same content in its structured caption format and no reference image. The 40 prompts pair two characters each, cover all four source shows, and span real×real, cartoon×cartoon, and real×cartoon combinations in both realistic and cartoon target styles; all methods use a fixed seed.

Table 4: **Inference settings of the multi-reference comparison.**

Method	Reference input	Resolution / length	Sampling
VACE-14B [Jiang et al., 2025]	multi-image	832×480, 81 frames @16 fps	50 steps, CFG 5.0
Phantom-Wan-14B [Liu et al., 2025]	up to 4 images	832×480, 121 frames @24 fps	50 steps, CFG 5.0 / 7.5
SkyReels-A2 [Fei et al., 2025]	1–4 images	832×480, 81 frames @15 fps	50 steps, CFG 5.0
Ours	none (names in prompt)	896×512, 41 frames @16 fps	50 steps, CFG 3.0

Open-source VLM judge. We re-score every video with Qwen3-VL-30B-A3B-Instruct [Bai et al., 2025], run locally on 32 uniformly sampled frames per video, with exactly the same metric prompts, score anchors, reference images, and character names as the Gemini evaluation. Qwen scores are stricter and more compressed, but both judges rank our method first on Identity Fidelity (Table 2). Motion Authenticity and Interaction Plausibility are nearly flat under Qwen (spread below 0.9 points across methods) for two reasons: Gemini ingests the full video while Qwen sees 32 sparse frames,

which makes temporal qualities harder to judge, and Qwen3-VL-30B-A3B is a mixture-of-experts model with only $\sim$3B active parameters per token.

User study. Ten of the 40 prompts were used for a blind four-way study. Each participant saw all ten scenes; for each scene, the four videos were shown side by side in a random left-to-right order together with the character reference images and the prompt, and the participant chose one video for each of four questions (identity, style, interaction, overall quality). Method names were never shown, and videos were served under opaque identifiers. The instructions shown to participants were: *"Each scene shows two characters. Use the reference images (how the characters should look) and the prompt to judge four videos, shown in random order. Pick one video per question. The videos loop automatically and have no sound."* Eleven participants recruited outside the project took part; participation was anonymous and only answers and timings were stored. Intervals in Table 2 are 95% Wilson score intervals [Wilson, 1927]. Table 5 breaks down the quality vote by pairing type: the reference-based baselines are competitive on cartoon$\times$cartoon pairs of globally famous characters, whereas the gap opens on live-action and mixed pairs, where a reference image does not convey how a character moves or should be re-styled. Figure 9 shows qualitative examples.

Table 5: **User-study quality vote by pairing type** (share of votes, %).

Method	real$\times$real	cartoon$\times$cartoon	mixed, realistic scene	mixed, cartoon scene
VACE-14B	0	9	3	0
Phantom-Wan-14B	27	**41**	12	0
SkyReels-A2	0	18	15	9
Ours	**73**	32	**70**	**91**

D.2 Stage-2-only training details

The Stage-2-only model in Figure 7 skips Stage 1 entirely and trains a rank-32 LoRA from the untouched Wan2.1-T2V-14B on the query/key/value/output projections and both feed-forward projections of every transformer block, with a constant learning rate of 2×10^{-5}, bf16 precision, and FSDP hybrid sharding on 16 A100 GPUs. The data are all 4,005 real clips of *Mr. Bean* and *Tom and Jerry* plus their four cross-style composite sets. Checkpoints are saved every 1,000 steps; we report the 2,000-step checkpoint. Because this LoRA must acquire the two-show character distribution and the style-mixing behavior jointly through a rank-32 bottleneck, identity errors are expected; the cross-style separation it already exhibits indicates that this capability does not depend on Stage 1.

D.3 Cross-backbone feasibility on LTX-2.3

To probe whether the character-tagged captioning recipe transfers beyond Wan2.1, we apply it to LTX-2.3 [HaCohen et al., 2026], a 22B audio–video model whose VAE, text encoder, and transformer differ from those of Wan2.1. The study uses an independently prepared corpus of about 28k clips from four animated series, structured `[character]` and `[scene]` captions following the Stage-1 conditioning principle, and a rank-32 LoRA on the attention projections for 2,000 steps. On 16 held-out prompts that combine characters from different series (768$\times$448, 121 frames, 24 fps), the adapted model places named characters from different series in a shared scene while preserving their visual identities (Figure 10). This is component-level evidence rather than a full replication: it uses a different, animation-only dataset, LoRA instead of full Stage-1 fine-tuning, and no Stage-2 CCA. We therefore describe the recipe as architecture-preserving and potentially portable, not as backbone-agnostic.

D.4 Long-tail characters

We evaluate 15 characters of the final model, spanning 55 to 5,509 training clips, under four conditions each with shared scene templates: two solo scenes in novel settings and two cross-style mixing scenes with a fixed high-frequency partner of the opposite style (60 videos). Table 6 and Figure 11 report the VLM scores by frequency tier. Degradation is gradual rather than a cliff: across a 100$\times$ reduction of per-character data, solo ID-FID decreases by only 0.75 points, mid-tail characters below 1,000 clips match the head tier, and a clear identity slip appears only for the rarest character (Irma Gobb, 55

Prompt: *Mr. Bean and Spike stand together in a quiet sunlit park, talking softly under tall trees.*



Prompt: *Mary Cooper and Tom pause to talk in the middle of a quiet public library with rows of bookshelves.*



Figure 9: **Qualitative comparison with multi-reference baselines.** Rows within each block, top to bottom: VACE-14B, Phantom-Wan-14B, SkyReels-A2, and ours. The reference-based baselines tend to re-style the live-action character as a cartoon or let one identity dominate the clip, whereas our model keeps both identities in their native styles.

clips). Each tier mean averages few videos, so differences of a few tenths of a point should not be over-interpreted.

D.5 Generalization to unseen characters

We write eight well-known characters that never appear in our training data (Superman, Snow White, Donald Trump, Iron Man, Pikachu, SpongeBob, Elsa, Harry Potter) in the same caption format as the training cast and generate them with the final model, alone and mixed with head-tier cast members under realistic and cartoon scene styles (Figure 12). The unseen identities are recognizably rendered—costume, hair, and signature props carry most of the identity—while the trained partner



Figure 10: **Cross-series character mixing on LTX-2.3.** Three held-out prompts, each combining characters from two different animated series (four frames per row), generated after adapting LTX-2.3 with our character-tagged captions. Audio is generated jointly but not shown.

Table 6: **Quality versus per-character training frequency** (VLM scores, 1–10; solo / cross-style mix).

Tier (training clips)	ID-FID	STY-COH	MOT-AUTH	INT-PLAUS
Head (1,265–5,509): Grizzly, Sheldon, Tom, Mr. Bean	7.00 / 6.62	7.00 / 6.62	7.00 / 6.62	6.62 / 5.75
Mid tail (464–956): Chloe, Spike, Nom Nom	6.83 / 6.83	6.83 / 6.50	6.50 / 6.33	6.50 / 5.50
Deep tail (123–377): 6 characters	6.58 / 6.67	6.92 / 6.75	6.83 / 6.25	6.67 / 5.75
Extreme tail (55–84): Yana, Irma Gobb	6.25 / 6.75	5.75 / 6.75	6.25 / 6.50	6.00 / 5.50

stays on-model. Cross-style control extends beyond the cast: Superman becomes a coherent 2D cartoon next to Grizzly, and Snow White is re-rendered photorealistically beside Mr. Bean. Identities defined mainly by facial likeness, such as Harry Potter, drift.

D.6 Effect of specialization on general generation

We compare the base Wan2.1-T2V-14B with the final MiMix model on 20 plain-text prompts with identical seeds: 12 general IPs outside our cast (6 requesting a realistic and 6 a cartoon style) and 8 character-free scenes probing landscapes, indoor scenes, fluid physics, crowds, close-ups, and camera motion. Each model runs at its recommended settings (base: 81 frames, 832×480; ours: 161 frames, 896×512). Table 7 reports VBench scores and Figure 13 shows one frame per prompt.

Every prompt still yields a coherent video, and most identities remain recognizable, but the cost is measurable and concentrated in three forms: a *style pull* toward the broadcast-TV aesthetic of the training series, *broadcast artifacts* such as subtitle-like overlays and mid-clip cuts, and *identity interference* on cartoon IPs that resemble trained characters. Aesthetic quality drops twice as much on cartoon IPs as on realistic ones. Part of the subject-consistency drop reflects protocol rather than quality: this frame-similarity score penalizes motion, and within our own model moving clips average 80.4 versus 94.3 for near-static ones, while our clips are twice as long and far more dynamic (dynamic degree $35 \rightarrow 70$). Because our model is tuned on a structured caption format, plain sentences are out-of-distribution; rewriting the 12 IP prompts in the caption format restores several identities (e.g., SpongeBob goes from absent to on-model), but not Mickey Mouse, who is still drawn with Jerry's face. Note that these caption-format prompts also include an explicit appearance string and were generated with the Stage-1 checkpoint. We regard identity interference between visually similar characters as a genuine limitation of specializing on a small cast; running Stage 1 as a low-rank adaptation is a natural way to reduce it, which we leave to future work.

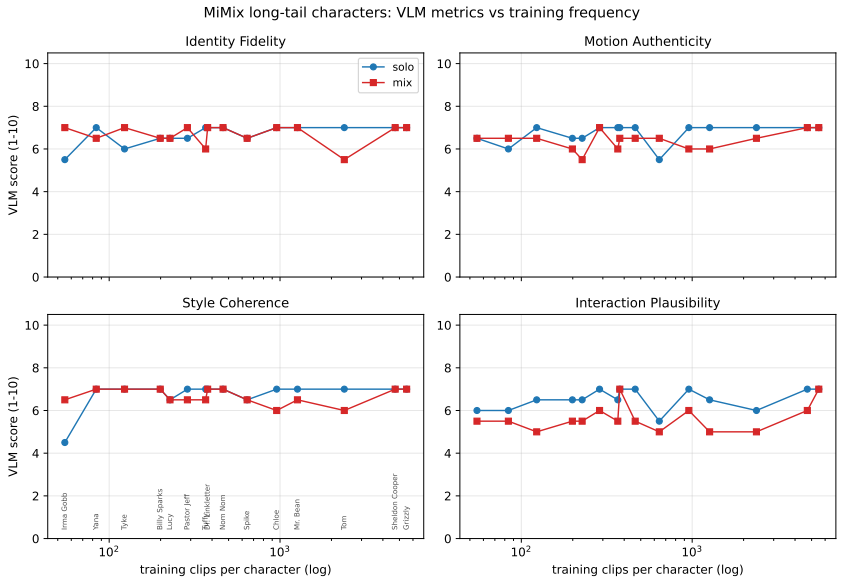


Figure 11: **Per-character VLM score versus training-clip count** (log scale) for solo and cross-style mix conditions.

Table 7: **VBench before (Base) and after (Ours) specialization** on 20 general prompts; relative change in parentheses.

	Realistic IPs (6)		Cartoon IPs (6)		General scenes (8)		Average (20)	
Dimension	Base	Ours	Base	Ours	Base	Ours	Base	Ours
Subject Consistency	95.78	84.78 (−11.5%)	95.02	83.07 (−12.6%)	96.43	84.74 (−12.1%)	95.81	84.25 (−12.1%)
Background Consistency	96.84	93.23 (−3.7%)	95.42	90.80 (−4.8%)	96.44	91.02 (−5.6%)	96.25	91.62 (−4.8%)
Motion Smoothness	99.46	98.56 (−0.9%)	98.66	97.32 (−1.4%)	98.92	98.45 (−0.5%)	99.01	98.14 (−0.9%)
Dynamic Degree	33.33	50.00 (+50.0%)	50.00	100.00 (+100.0%)	25.00	62.50 (+150.0%)	35.00	70.00 (+100.0%)
Aesthetic Quality	62.03	57.69 (−7.0%)	64.72	54.69 (−15.5%)	63.83	59.95 (−6.1%)	63.56	57.69 (−9.2%)
Imaging Quality	62.16	60.62 (−2.5%)	74.56	70.91 (−4.9%)	67.74	61.66 (−9.0%)	68.11	64.12 (−5.9%)

D.7 Failure-case analysis

We stress-test the final model along three axes—interaction complexity, number of characters, and style diversity—using only well-trained characters ($\geq$ 1,265 clips) and a fixed seed, so that failures reflect the stress factor rather than data scarcity (Figure 14).

Interaction complexity. Two-character interaction holds across every style pairing, including sustained physical contact (Mr. Bean bandaging Sheldon's elbow), coordinated whole-body motion (Jerry and Panda dancing), object hand-over (Mr. Bean passing a cup of tea to Panda), and shared activity (Sheldon and Tom playing chess). Verbs far from the training distribution, such as a formal waltz, fall back to the nearest familiar interaction pattern.

Number of characters. Scenes with three and four characters are reliable: every identity remains distinct and on screen for the whole clip, with cartoon and live-action characters sharing one room. With five or six characters, the scene and styles still hold, but one or two characters never enter the shot. Multi-character clips in our corpus are plentiful but always same-show ensembles, while the cross-show augmentation set contains only about ten clips with three or four characters and none with more. Phrasing the cast as in the training captions (a pose inside each character description and one action clause each) is important for four-character scenes.

Style diversity. With the pair and action fixed, we request eight rendering styles absent from training (anime, 3D CGI, watercolor, pixel art, manga ink, claymation, oil painting, film noir). Each request resolves onto the nearer of the two registers in our data—anime and watercolor to cel cartoon, 3D CGI and claymation to live action—yet both identities and the cross-style pairing survive every shift.



Figure 12: **Unseen characters composed with the trained cast.** Each row shows one never-trained character: solo (left), mixed with a realistic cast member in a realistic scene (middle), and mixed with a cartoon cast member in a cartoon scene (right).

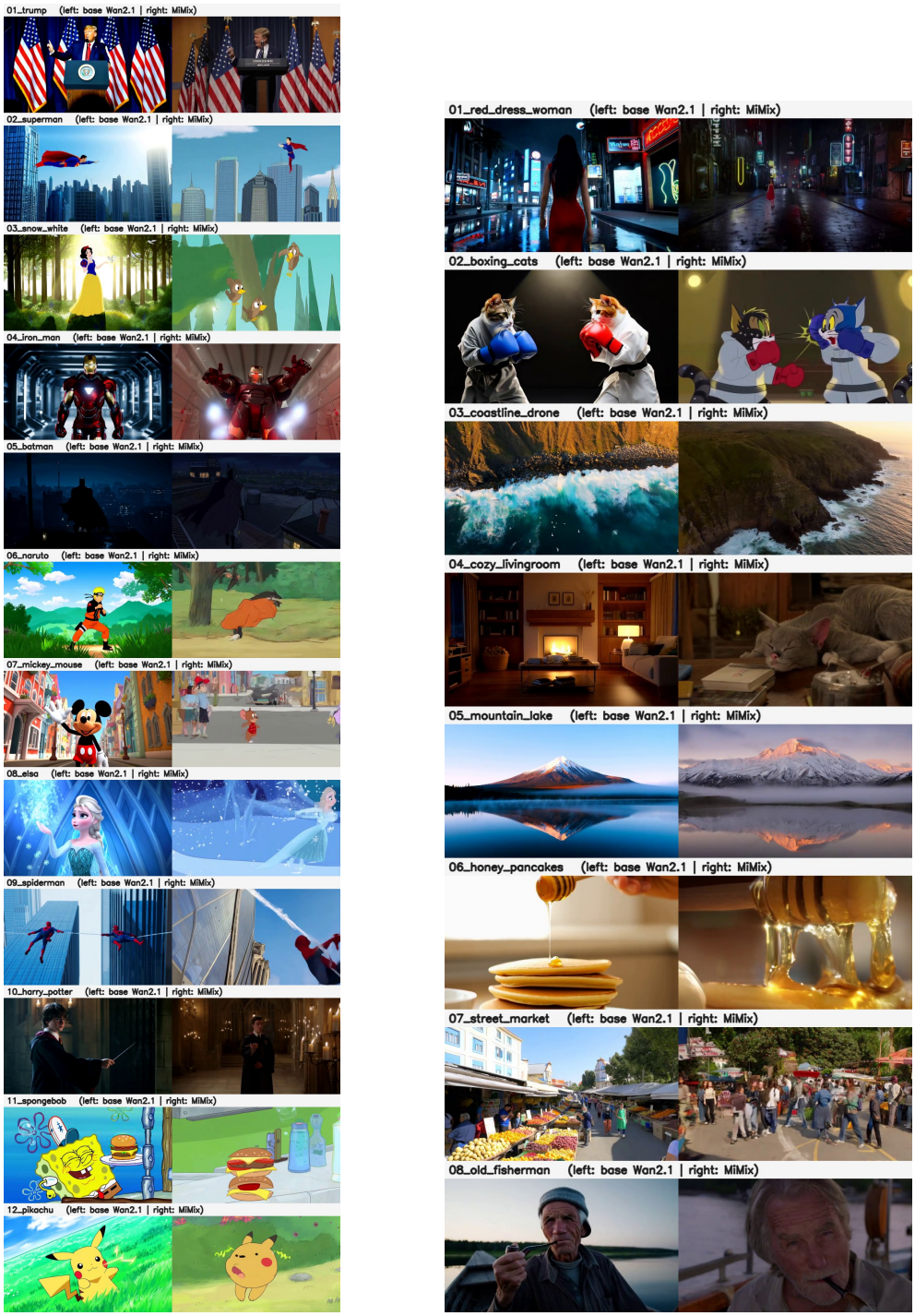


Figure 13: **Base Wan2.1 (left of each pair) versus our model (right of each pair)** on 12 general IPs (left column) and 8 character-free scenes (right column), identical prompts and seeds.



Figure 14: **Stress tests.** Each row shows two clips (two frames each). Top: complex interactions (Mr. Bean bandaging Sheldon's elbow; Sheldon and Tom playing chess). Middle: four characters, all of whom stay in frame (left), and six requested characters, of whom five appear (right). Bottom: untrained styles resolve onto our two registers—anime to cel cartoon (left) and 3D CGI to live action (right)—while both identities are preserved.

All three limits share one cause, the edge of the training distribution, and suggest data-side remedies: cross-show multi-character augmentation, interaction-diverse captioning, and style-swap augmentation.

E Extended Figure Captions

Extended caption for Figure 3 of the main paper. The diverse-interaction grid (Figure 3 of the main paper) surveys seven further interactions generated by our method (one example per row, five frames per example):

- **Row 1.** *Sheldon Cooper* (*Young Sheldon*) meets *Mr. Bean* at a restaurant booth.
- **Row 2.** *Mary Cooper* (*Young Sheldon*) encounters cartoon *Tom* (*Tom and Jerry*) in a library.
- **Row 3.** *Dr. Sturgis* (*Young Sheldon*) sits with cartoon *Tuffy* (*Tom and Jerry*) at a dining table.
- **Row 4.** *Mary Cooper* (*Young Sheldon*) shares a meal with cartoon *Grizzly* (*We Bare Bears*) in her home.
- **Row 5.** *Meemaw* (*Young Sheldon*) and cartoon *Ice Bear* (*We Bare Bears*) stand together in a yard.
- **Row 6.** *Georgie* (*Young Sheldon*) meets cartoon *Panda* (*We Bare Bears*) in a parking lot at night.
- **Row 7.** *George Sr.* (*Young Sheldon*) and *Mr. Bean* share a moment together.

The seven examples include cross-show pairings of realistic *Young Sheldon* characters with cartoon characters from *Tom and Jerry* and *We Bare Bears*, as well as cross-show realistic pairings between *Young Sheldon* and *Mr. Bean*. None of these character pairings co-occur in the original training videos, yet the model places them coherently together while preserving each character's canonical visual appearance and native style.

GPT-4o captioning instruction (shared across shows)

The same template is sent to GPT-4o for every show; only the four placeholders `<movie_name>`, `<show_kind>`, `<scene_style>`, and `<character_list>` are filled in per show.

Instruction body (abridged): *You are captioning a clip from "`<movie_name>`", a `<show_kind>`. The characters listed below are fictional roles in this show; describe what these named fictional characters are doing in this scripted scene.*

Output format — exactly one line, in this structure:

[character]*NAME*:*appearance***[/character]**
[character]*NAME*:*appearance***[/character]**...
[action]*description***[/action]**
[scene-style]<scene_style>**[/scene-style]**

Rules.

- One [character] block per character actually visible in the clip; use only canonical names from <character_list>—no aliases, no "a man".
- appearance: 5–15 words on clothing, posture, and items *as visible in this clip*; avoid "often" or "usually".
- action: 25–50 words, concrete and visual, mentioning the setting.
- Avoid hedging ("seems", "appears", "may", "suggests"); do not mention the show title, the frames, the video, or this prompt.

Per-show metadata and canonical character rosters

Mr. Bean `movie_name`: (scripted live-action British comedy series) | `scene_style`: `real`

`Mr. Bean`	main character, a man typically in a brown tweed jacket and red tie
`Irma Gobb`	Mr. Bean's girlfriend, a woman with curly brown hair
`Teddy`	Mr. Bean's small brown stuffed teddy bear (he treats it as alive)

Tom and Jerry `movie_name`: (animated cartoon) | `scene_style`: `cartoon`

`Tom`	gray cat (main character)
`Jerry`	small brown mouse
`Tuffy`	small grey mouse, often wears a white diaper (alias: `Nibbles`)
`Spike`	gray bulldog
`Tyke`	Spike's small brown puppy
`Quacker`	yellow duckling
`Butch`	black alley cat
`Toodles`	white female cat with a bow

We Bare Bears `movie_name`: (animated cartoon) | `scene_style`: `cartoon`

`Grizzly`	main brown bear, outgoing and energetic (aliases: `Grizz`, `Grizzly Bear`)
`Panda`	main black-and-white panda bear, shy (alias: `Panda Bear`)
`Ice Bear`	main white polar bear, calm (alias: `Polar Bear`)
`Chloe`	young Korean-American girl with glasses, the bears' kid friend
`Nom Nom`	mean small koala internet celebrity
`Charlie`	tall, brown, scruffy sasquatch
`Ranger Tabes`	enthusiastic female park ranger
`Lucy`	young woman with curly brown hair, Chloe's friend
`Yana`	Russian woman, Ice Bear's love interest

Young Sheldon `movie_name`: (scripted live-action sitcom) | `scene_style`: `real`

`Sheldon Cooper`	main character, a young boy with short brown hair (alias: `Sheldon`)
`Mary Cooper`	Sheldon's mother, a woman with curly auburn hair (alias: `Mary`)
`George Sr.`	Sheldon's father, a middle-aged man, high-school football coach (aliases: `George`, `George Cooper`)
`Georgie`	Sheldon's older brother, a teenage boy with light brown hair (alias: `George Jr.`)
`Missy`	Sheldon's twin sister, a young girl with light brown hair
`Meemaw`	Sheldon's grandmother, an older woman with short white hair and glasses (alias: `Connie Tucker`)
`Dr. Sturgis`	Sheldon's physics mentor, an older man with grey hair and glasses (aliases: `John`, `John Sturgis`)
`Dr. Linkletter`	Sheldon's professor, an older man with grey hair (alias: `Linkletter`)
`Pastor Jeff`	the family's pastor, a middle-aged man
`Tam`	Sheldon's friend, a boy with black hair
`Billy Sparks`	the Coopers' neighbor, a boy with curly red hair

Prompt: Panda, Mr. Bean, and Sheldon attend a tech startup pitch event. Panda pitches an app for bear selfies, Bean presents a banana phone, and Sheldon critiques everyone's latency stats. Nobody wins.



Prompt: Mr. Bean and Tom accidentally get locked inside a fancy hotel kitchen. While Tom chases a mouse under silver trays, Bean tries to cook pasta but ends up flooding the entire room with foam from a dishwasher he mistook for an oven.



Prompt: Tom and Grizzly get jobs as mall security. Grizzly tries to be friendly with everyone, while Tom gets distracted chasing Jerry through every store. By the end, they both get promoted for "enthusiasm."



Figure 15: **Comparison on multi-subject interaction.** Results from SkyReels-A2 [Fei et al., 2025] (top row) and ours (bottom row).

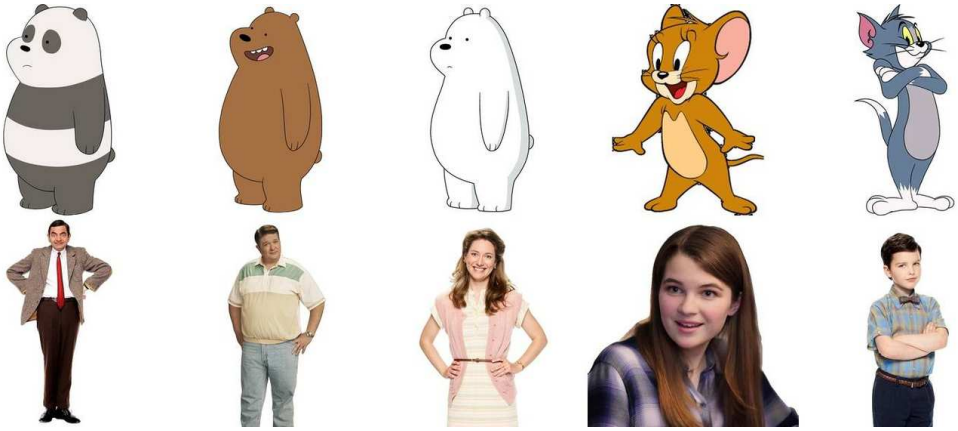
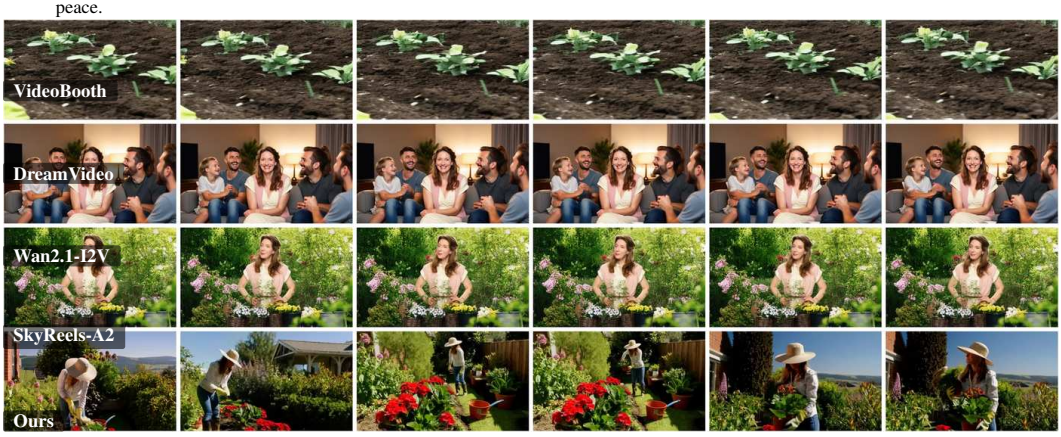


Figure 16: **Reference images** of 5 cartoon characters including Panda, Grizzly, Ice Bear, Tom and Jerry, and 5 realistic characters including Mr. Bean, George Sr., Mary Cooper, Missy and Sheldon.

Prompt: Mary Cooper is planting flowers in the garden, wearing gloves and a hat, enjoying the sunshine and the quiet moment of peace.



Prompt: George Sr. is barbecuing outside, flipping burgers with a big smile, while the family gathers around waiting for food.



Prompt: Missy is standing on a stage, nervously holding a microphone, trying to sing while imagining herself as a famous star.

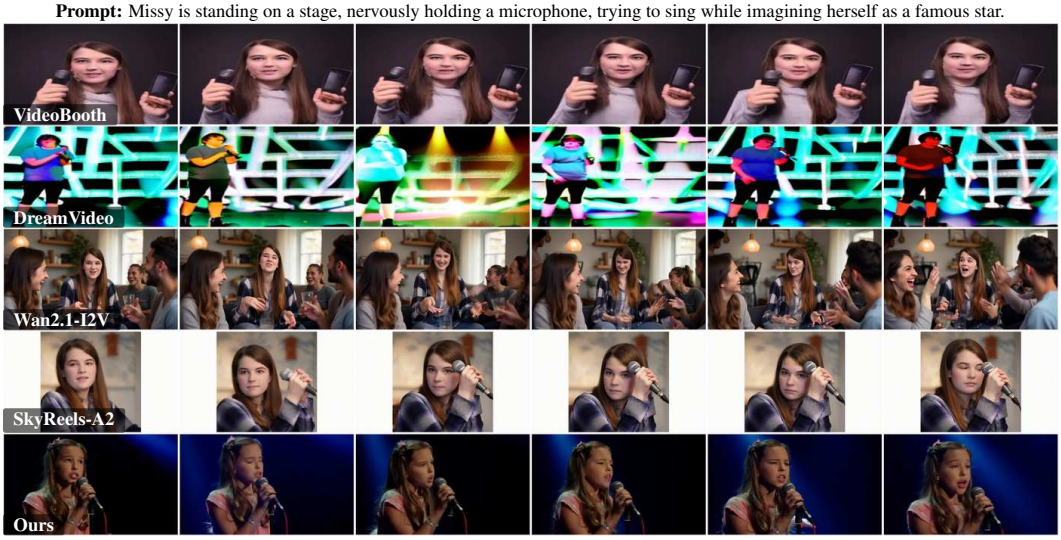
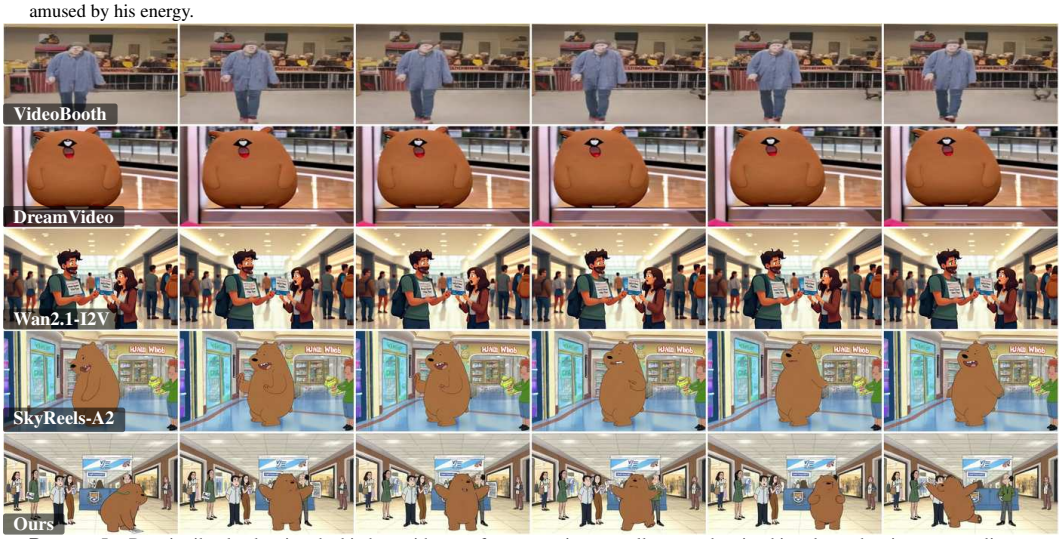


Figure 17: **Comparison on single-subject generation.** From top to bottom: results from VideoBooth [Jiang et al., 2024], DreamVideo, Wan2.1 [Team Wan, 2025], SkyReels-A2 [Fei et al., 2025] and ours.

Prompt: Tom is sitting on the floor with a bowl of milk, drinking it happily while closing his eyes in comfort.



Prompt: Grizzly is standing in a shopping mall, handing out flyers with too much enthusiasm, while people look confused but also amused by his energy.



Prompt: Ice Bear is silently cleaning the kitchen with great focus, wearing a small apron, showing his calm and serious personality compared to his brothers.

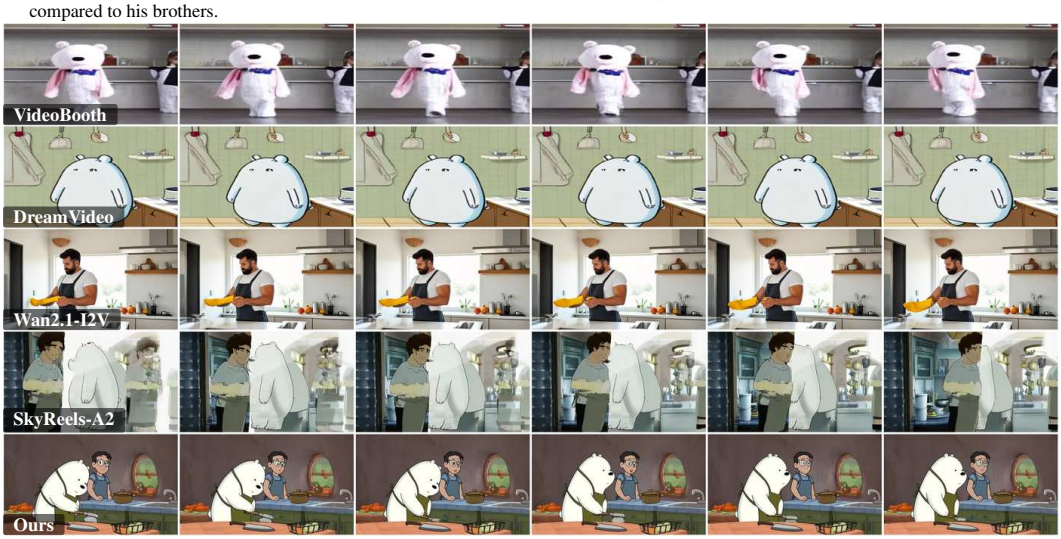
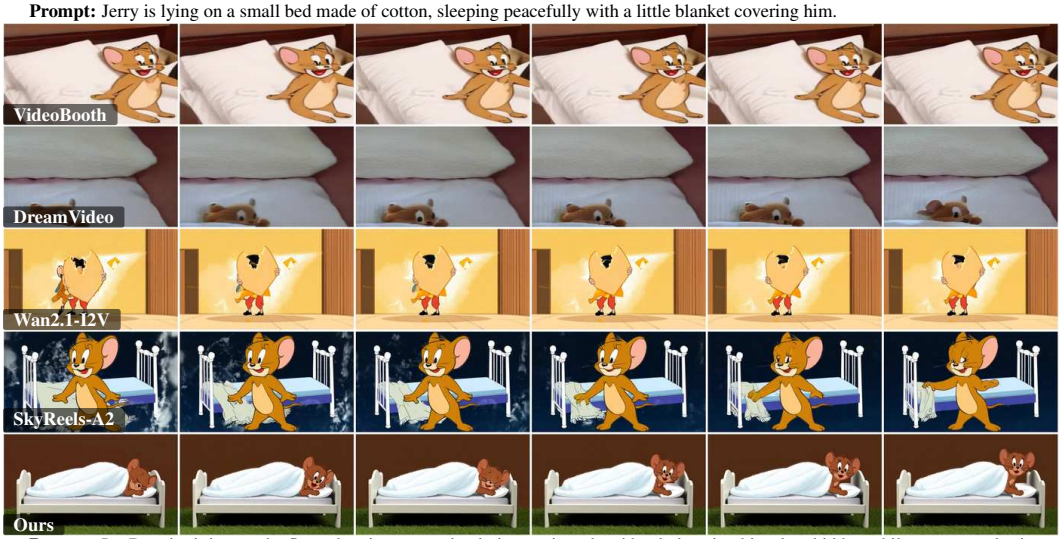
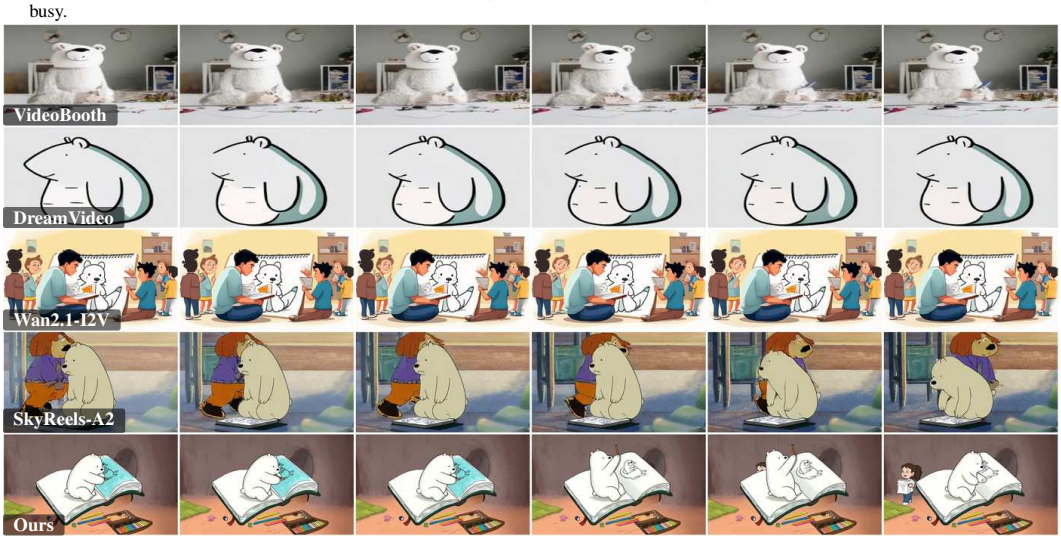


Figure 18: **Comparison on single-subject generation.** From top to bottom: results from VideoBooth [Jiang et al., 2024], DreamVideo, Wan2.1 [Team Wan, 2025], SkyReels-A2 [Fei et al., 2025] and ours.

Prompt: Jerry is lying on a small bed made of cotton, sleeping peacefully with a little blanket covering him.



Prompt: Ice Bear is sitting on the floor, drawing cute animal pictures in a sketchbook, keeping his talent hidden while everyone else is busy.



Prompt: Panda is holding a big milk tea, sipping happily while scrolling through his favorite social media app, completely ignoring the world around him.

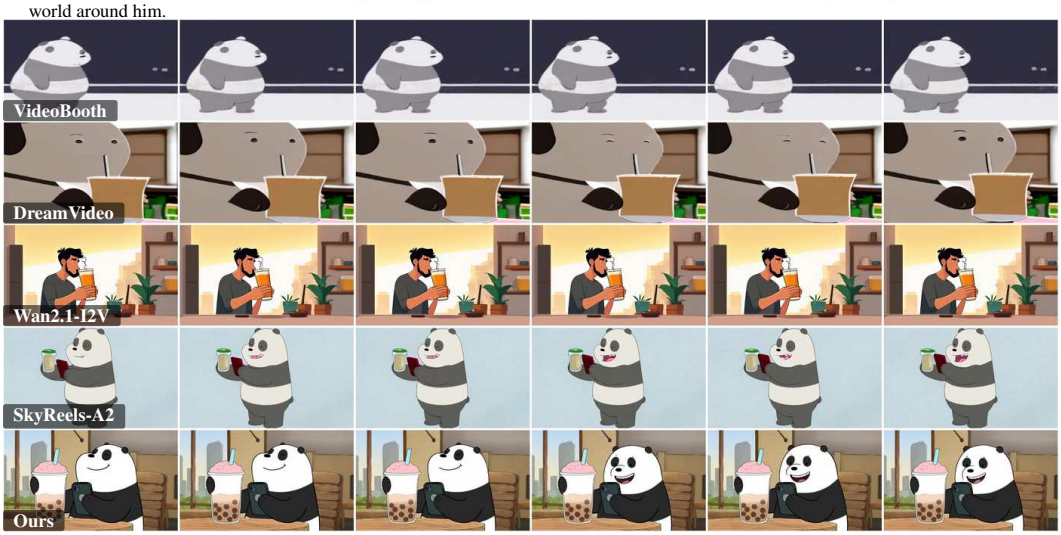
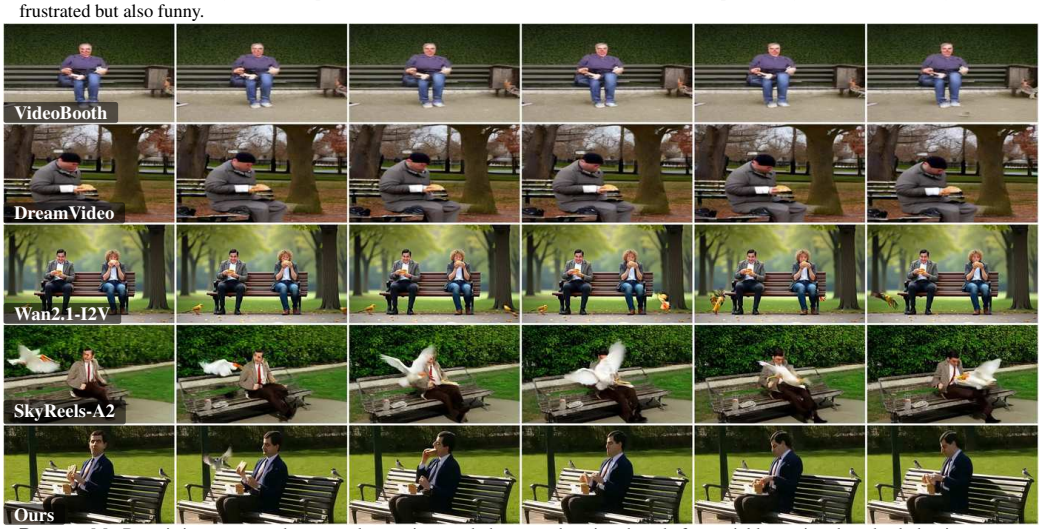
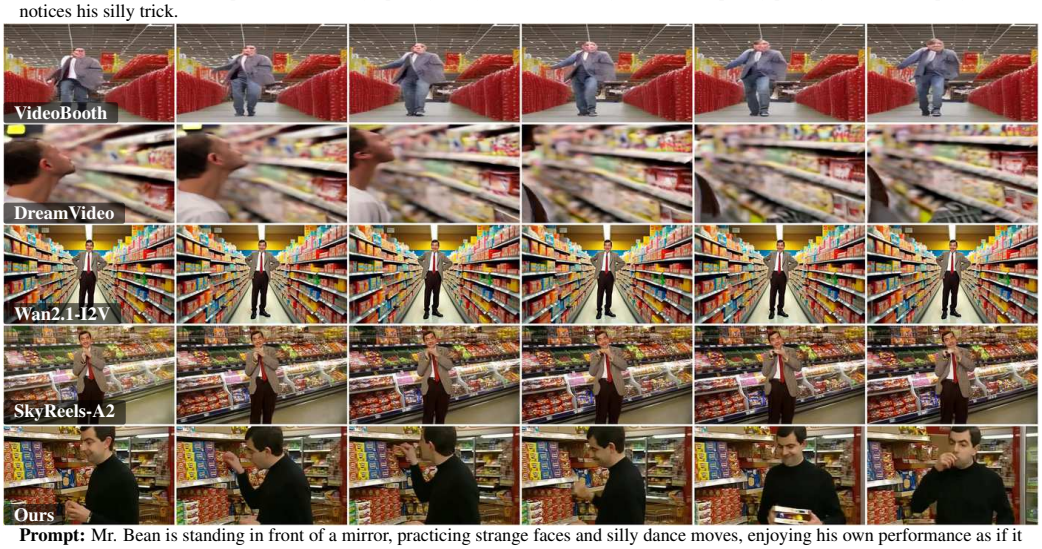


Figure 19: **Comparison on single-subject generation.** From top to bottom: results from VideoBooth [Jiang et al., 2024], DreamVideo, Wan2.1 [Team Wan, 2025], SkyReels-A2 [Fei et al., 2025] and ours.

Prompt: Mr. Bean is sitting alone on a park bench, trying to eat his sandwich while a bird keeps stealing the crumbs, making him look frustrated but also funny.



Prompt: Mr. Bean is in a supermarket, secretly opening candy boxes and tasting them before quickly putting them back, hoping no one notices his silly trick.



Prompt: Mr. Bean is standing in front of a mirror, practicing strange faces and silly dance moves, enjoying his own performance as if it were a comedy show.



Figure 20: **Comparison on single-subject generation.** From top to bottom: results from VideoBooth [Jiang et al., 2024], DreamVideo, Wan2.1 [Team Wan, 2025], SkyReels-A2 [Fei et al., 2025] and ours.