VideoBooth Review

Abstract

The paper proposes a fine-tuning framework for personalized text-to-video generation conditioned on a single reference image ("image prompt"). The core is a coarse-to-fine mechanism: (i) a CLIP-image-driven encoder that replaces the text tokens corresponding to the subject (coarse identity) and (ii) a multi-scale *attention injection* that appends the image latent as additional keys/values in cross-frame attention (fine details, temporal consistency).

1 Core Concepts and Method

Problem. Purely text-driven video diffusion models often fail to match a specific subject's appearance. *VideoBooth* augments text prompts with a single *image prompt* to better preserve subject identity during video generation [1].

Backbone. The method is built on a latent text-to-video diffusion model (a Stable-Diffusion-style U-Net extended with temporal/cross-frame attention) [7, 5, 6].

Coarse (token-level) image injection. A CLIP image encoder extracts a global image feature which an MLP maps into the text-embedding space. The mapped vector replaces the token(s) in the text prompt that refer to the subject. This provides a *coarse, high-level* identity prior shared across layers.

Fine (feature-level) image injection. To restore spatial details, the reference image is passed through the Stable Diffusion's VAE to obtain a latent at matching scales; noise is added according to the current diffusion time-step. In each cross-frame attention layer, the method *concatenates* the image latent's projected keys/values to the keys/values of the frame(s), then:

1. updates the *first frame's* values using attention over (image + first-frame) keys/values
2. reuses these *updated first-frame values* as a persistent reference for subsequent frames to promote temporal consistency

Multi-scale injection (low to high resolution) yields a *coarse-to-fine* image-conditioning pipeline.

Training schedule. Two-stage training is essential. First learn the token-level image encoder (with K/V adapters) so the model reliably grounds identity, then train attention injection so the model uses image latents to refine spatial details. Joint training leads to over-reliance on the fine path and temporal artifacts (as shown in their ablations) [1].

2 Strengths

- **Identity injection at the token level.** Replacing the subject span in the text embedding with mapped CLIP-image features is a clean way to *directly* bind the image prompt to the rest of the text.
- **Coarse-to-fine refinement** The multi-scale attention injection complements the token-level identity by restoring spatial detail and keeping it consistent across frames.
- **Temporal propagation trick.** Refining the first frame and reusing its values for subsequent frames is an effective way toward identity stability with little engineering overhead.
- **Tuning-free inference.** Unlike DreamBooth/Textual Inversion [3, 2], a single trained model generalizes to many subjects with a single image prompt.
- **Empirical quality.** On image-alignment metrics (CLIP-Image, DINO) and in qualitative/user studies, VideoBooth preserves the prompt image appearance better than adapted baselines [1].

3 Weaknesses

- **Unexplained frozen query projection, with no ablation.** The method learns separate K/V projections for the image latents but (effectively) leaves queries unchanged. Without an ablation on learning/finetuning query projections for image-conditioned attention, it's unclear if capacity is limited here and may lead to suboptimal results.
- **Compute overhead from K/V concatenation.** Passing the latens of the frames through two different projections and then concatenate them *doubles* the cost of performing attention, which raises both memory and latency.
- **First-frame anchoring limits long sequences.** Always propagating from the first frame can cause drift or rigidity in long videos (identity tied to outdated pose/lighting), and error propagation if frame 1 is suboptimal. A rolling window or periodic re-anchoring would likely improve long-horizon generation.
- **Metric bias toward CLIP space.** Using CLIP both as a building block (coarse encoder) and as a primary evaluator (CLIP-Text/CLIP-Image) risks favorable bias. DINO helps, and the user study is positive, but additional human studies or downstream task metrics would strengthen claims.
- **Backbone selection.** The U-Net backbone is effective but lags behind *Diffusion Transformers (DiT)* in scalability/quality on image generation [8]. A transformer video diffuser could unify spatial/temporal attention and simplify injection via tokenization.

4 Suggestions and Future Improvements

- Replace global first-frame anchoring with:
  - a *sliding window* over recent k frames
  - *periodic keyframes* (every m frames) where image injection reoccurs to refresh identity under pose/lighting shifts
- Backbone modernization by adopting a diffusion transformer (DiT) [8]:
  - Remove cross attention and pass image and text conditioning through the adaptive layer normalization mechanism [8]
  - keep cross-frame and temporal attention only
  - DiTs' capacity can be scaled more easily
  - DiTs improve global coherence and detail at higher resolution
- Ablations and diagnostics.

  Add studies on: (i) learning query projections, (ii) cost/quality trade-offs of compact image-token banks vs. full K/V concat, (iii) long-horizon robustness with different temporal strategies, (iv) CLIP-free human evaluations.

5 Experiments, Dataset, and Results

Dataset. The VideoBooth dataset assembles (caption, image-prompt, video) triplets by segmenting first-frame subjects from WebVid and filtering them by size/motion. This is a practical approach to creating a labeled dataset for the task, although it inherits WebVid's biases and favors salient foreground objects.

Baselines. Adapting Textual Inversion [2] and DreamBooth [3] to video provides important reference points (optimization-based personalization), while ELITE [4] offers an encoder-based comparator. It would be informative to also include *fine-tuned* video baselines (e.g., one-shot Tune-A-Video [5]) under a fixed compute budget.

Metrics and ablations. The quantitative gains in DINO/CLIP-Image and the user study support the identity-preservation claim. The ablation set convincingly shows: (i) coarse-only misses detail, (ii) fine-only overfits the first frame (temporal artifacts), and (iii) two-stage training is necessary. Adding the query-projection ablation and long-horizon tests would round this out.

6 Conclusion

VideoBooth is a well-motivated and thoughtfully engineered step toward personalized text+image-to-video generation. Its token-level image injection (coarse) plus multi-scale attention injection (fine) is simple, modular, and empirically effective. The paper would be stronger with ablations on query projections, efficiency analyses, and long-horizon conditioning.

References

[1] 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. *arXiv preprint arXiv:2312.00777*, 2023.

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *arXiv preprint arXiv:2208.12242*, 2022.

[4] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. *arXiv preprint arXiv:2302.13848*, 2023.

[5] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Ying Shan, and Mike Zheng Shou. Tune-A-Video: One-shot Tuning of Image Diffusion Models for Text-to-Video Generation. *arXiv preprint arXiv:2212.11565*, 2022.

[6] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. *arXiv preprint arXiv:2303.13439*, 2023.

[7] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent Video Diffusion Models for High-Fidelity Video Generation with Arbitrary Lengths. *arXiv preprint arXiv:2211.13221*, 2022.

[8] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. (*arXiv:2212.09748*)

[9] Zheng Chong, Liqian Ma, Zhifei Zhang, Zhe Lin, and Dong Xu. CatVTON: Concatenation Is All You Need for Virtual Try-On with Diffusion Models. *arXiv preprint arXiv:2407.15886*, 2024.