Synthesizing Compositional Videos from Text Description

Prajwal Singh* Kuldeep Kulkarni† Shanmuganathan Raman* Harsh Rangwani†
CVIG Lab, IIT Gandhinagar, India*
Adobe Research, India†
{singh_prajwal, shanmuga}@iitgn.ac.in, {kulkulka, hrangwani}@adobe.com

Abstract

Existing pre-trained text-to-video diffusion models can generate high-quality videos, but often struggle with misalignment between the generated content and the input text, particularly while composing scenes with multiple objects. To tackle this issue, we propose a straightforward, training-free approach for compositional video generation from text. We introduce Video-ASTAR for test-time aggregation and segregation of attention with a novel centroid loss to enhance alignment, which enables the generation of multiple objects in the scene, modeling the actions and interactions. Additionally, we extend our approach to the Multi-Action video generation setting, where only the specified action should vary across a sequence of prompts. To ensure coherent action transitions, we introduce a novel token-swapping and latent interpolation strategy. Extensive experiments and ablation studies show that our method significantly outperforms baseline methods, generating videos with improved semantic and compositional consistency alongside improved temporal coherence [1].

1. Introduction

Recent advancements in large-scale Text-to-Video (T2V) generation have significantly improved the ability to produce photorealistic videos from natural language descriptions [8, 10, 17, 20]. These models can generate both short and long videos with high fidelity, making them increasingly popular for creative applications, virtual environments, and assistive technologies. However, a persistent challenge arises when the input text involves multiple objects, actions, or their combinations. In such scenarios, generated videos often exhibit missing entities, misplaced interactions, or incorrect temporal ordering of actions. This issue is referred to as compositional generation, where models are expected to faithfully generate scenes involving multiple distinct components described in the input.

While composition has been extensively studied in the Text-to-Image (T2I) literature, with several methods proposed to improve fine-grained alignment between entities and their spatial relationships [1, 2, 6, 15], relatively few works have addressed the challenge in the T2V domain [21, 25]. The difficulty in extending compositional generation to video lies in the increased complexity of maintaining both spatial and temporal coherence across frames. Unlike images, videos require models to capture temporal relationships between multiple entities and actions while preserving their individual attributes across time.

To address these limitations, we propose an optimization-based framework, Video-ASTAR, for generating temporally coherent videos rich in composition from textual prompts. Our method builds upon the pre-trained VideoCrafter2 diffusion model [4]. Inspired by ASTAR [1], we introduce attention segregation and retention mechanisms for video generation to gain finer control over attention maps, allowing distinct textual tokens to modulate different object regions in space and time. This design encourages disentangled control over multiple objects and their interactions, directly improving the composition. We also introduce a novel Centroid Loss to enhance entity-action interactions by aligning the spatial attention maps of relevant entities. We further identify specific cross-attention layers within the model that are most effective for controlling entity, enabling better grounding of text in video generation.

In addition to spatial composition, we consider the challenge of Multi-Action Video Generation (MAVG), which involves generating a video that captures a coherent sequence of actions conditioned on a textual description [5, 13, 16, 21, 24]. Most current T2V models struggle to maintain spatial-temporal consistency over extended durations, especially when the text contains multiple actions. This leads to abrupt transitions in the generated video, which affects the overall video quality and semantic alignment with the input. For handling MAVG, we extend our compositional generation framework by introducing Token

[1] Project page: https://prajwalsingh.github.io/Video-ASTAR/

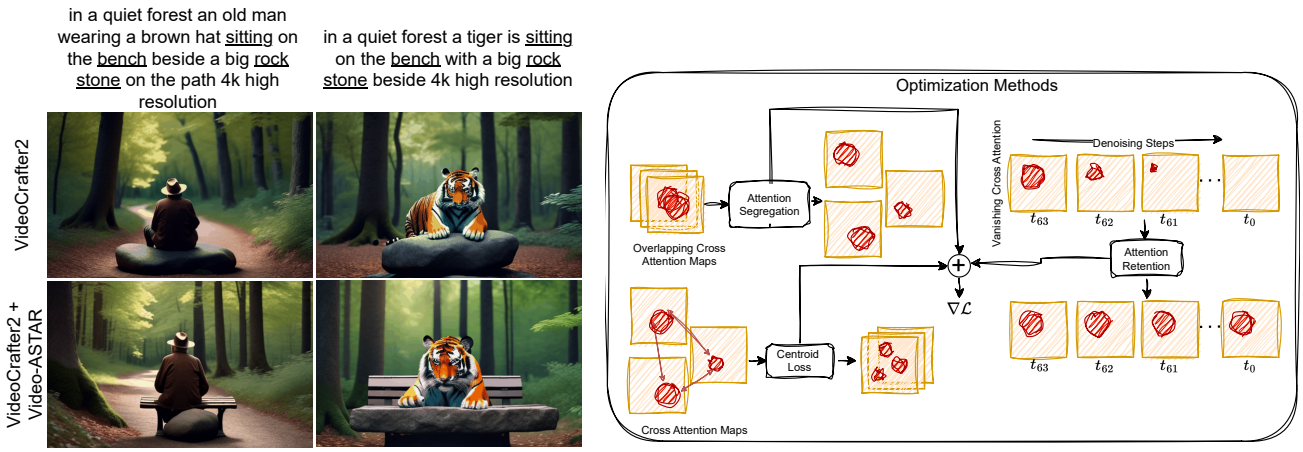

Figure 1. **Video-ASTAR Framework.** We introduce Video-ASTAR, which enforces the consistent generation of concepts in the prompt (*left*), via a latent optimization framework based on losses on corresponding concept token attention maps (*right*) in VideoCrafter2.

Swapping and Latent Interpolation strategies. These techniques aim to preserve entity identity across multiple action phases and ensure smoother transitions between distinct actions. As a result, our method enables the generation of videos that are not only visually coherent but also accurately aligned with complex multi-action textual prompts.

The following are the contributions of our work improving text-based video synthesis:

1. *Optimization-Based Framework Video-ASTAR for Compositional Video Generation:* We present a training-free approach that leverages the pre-trained VideoCrafter2 [4] diffusion model to generate videos with rich and coherent composition, following content from textual descriptions accurately.
2. *Centroid Loss for Attention Interaction:* We introduce a novel loss function, Centroid Loss, which promotes stronger interaction between entities in the attention maps. This helps maintain consistent and accurate representations of objects and actions throughout the denoising steps in the diffusion process.
3. *Method for Attention Control:* To improve attention segregation and retention across video frames, we re-model ASTAR [1] loss across frames and extend it with a mean-threshold-based mask creation strategy. Specifically, we generate a mask from the first frame using mean attention values and apply this mask consistently across all frames. This enables stable tracking of tokens throughout the video. Additionally, we identify the cross-attention layers within the diffusion model that are most critical to controlling the generation, enabling more targeted manipulation of visual content. (Figure 1)

2. Related Works

We find that the object/action being missed is one of the main issues in compositional video generation. To address this problem, there are different strategies that are being used for images and videos.

Compositional Image Generation. Attend-and-Excite [2] addresses semantic consistency issues in text-to-image generation by leveraging cross-attention maps to guide the generative process. This method ensures that key objects and relationships described in the text are accurately represented in the generated images by iteratively optimizing a loss on attention maps during the denoising process to align them with semantic concepts in the input text. However, a limitation of Attend-and-Excite is the vanishing of certain concepts from the attention map.

One way this issue is addressed by ASTAR [1], by introducing an attention retention loss to ensure that all tokens maintain some association with pixels throughout the denoising process. Building on this, we introduce an ASTAR-based loss with our proposed Centroid Loss to enhance attention retention further. Additionally, we present an efficient approach for calculating the mask used in ASTAR's attention retention loss, improving its overall effectiveness.

Compositional Video Generation. In the Video Composition (VICO) work [25], the authors proposed a method to reformulate video generation as a flow equalization problem, where temporal dynamics and spatial relationships are represented through structured flow fields. The approach decouples content generation and motion synthesis, ensuring accurate representation of both static and dynamic components. VideoTetris [21] frames video generation as a block-wise process, where "content blocks" representing objects and "motion blocks" capturing temporal dynamics are independently generated and combined based on the input text. In contrast, MagicComp [26] and VideoRepair [14] are training-free approaches, but they rely on LLM for planning and refinement. Our proposed method operates on spatial cross-attention maps and uses only the loss functions to ensure correct temporal and spatial relationships for compositional video generation. Some recent methods tackle com-

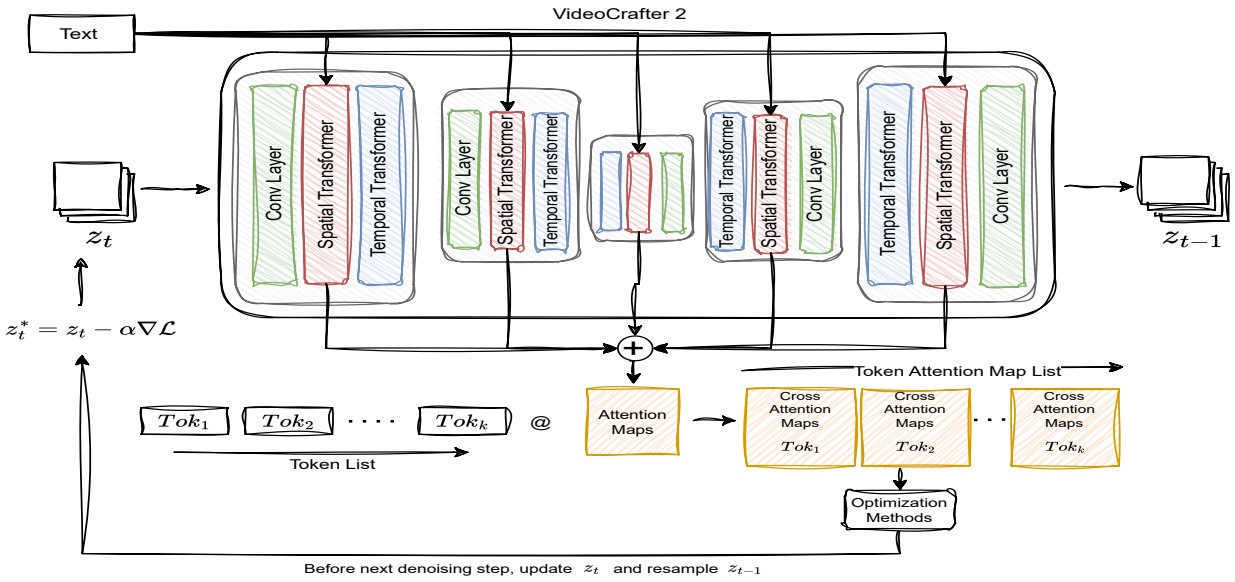

Figure 2. **Video-ASTAR Framework.** The figure illustrate VideoCrafter2 (VC2) [4] framework with proposed latent optimization step. The VC2 video latent diffusion model conditions the noisy latent input on the given text prompt, and during the denoising step, the proposed method ensures that the attention map attends to each token according to the text. The final denoised latent is then given as input to the pre-trained latent-to-frame decoder network of VC2 for video generation.

positional video generation by redesigning models. GenMAC [11] uses multi-agent collaboration, while DreamRunner [23] employs retrieval-augmented motion priors. In contrast, Video-ASTAR requires no new architecture, modules, or retraining; it operates purely at test time on existing diffusion models, offering a lightweight and complementary alternative to training-heavy approaches.

Multi-Action Videos. FIFO [13] introduced a training-free diagonal denoising framework for multi-action video generation, with the goal of producing infinitely long videos. While effective for generating extended video sequences, FIFO struggles to capture multiple actions accurately and often results in inconsistent backgrounds and incoherent video frames. To address these shortcomings, we extended the Video-ASTAR optimization to FIFO, achieving improved background consistency and frame coherence.

3. Method

Latent Video Diffusion Model (LVDM). In this work we have used pre-trained VideoCrafter2 [4] model for text-to-video generation. The VideoCrafter2 (VC2) pipeline consists of two components: 1) a video VAE and 2) a video latent diffusion. A pretrained VAE from the Stable Diffusion [19] model is used as a video autoencoder, where each frame is encoded independently into the latent z_0 and the VC2 process 16 frames at a time. The video latent diffusion model is then trained on encoded latent z_0 with the following forward (Eq. 1 and 2) and reverse denoising (Eq. 3) :

$$q(z_{1:T}|z_0) := \prod_{t=1}^{T} q(z_t|z_{t-1}), \tag{1}$$

$$q(z_t|z_{t-1}) := \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbb{I}) \tag{2}$$

$$p_\theta(z_{t-1}|z_t) := \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, y), \Sigma_\theta(z_t, t, y)) \tag{3}$$

where each video latent is of shape $(C \times F \times H \times W)$, T is the number of total diffusion timesteps, and β_t is the noise level at timestep t. This will give us a set of noisy video latent z_t at any timestep t as discussed in [3]. The μ_θ and Σ_θ are learned U-Net parameters conditioned on timestep t and text prompt y. The video latent shape of pre-trained VC2 model is $z_0 \in \mathbb{R}^{4\times16\times40\times64}$.

Cross Attention Maps. To enforce the semantic conditioning from the text prompt in the denoising step of U-Net diffusion, cross-attention is used. As shown in Figure 2, the text prompt is given as input to the cross-attention layer in the spatial transformer layer of VideoCrafter2 across the different stages. The text prompt y is encoded using the pre-trained CLIP [18] text encoder ϕ followed by computing the cross attention:

$$CrossAttn(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right).V, \tag{4}$$

$$Q = W_Q^i \,.\, \psi(z_t), K = W_K^i \,.\, \phi(y), V = W_V^i \,.\, \phi(y) \tag{5}$$

here, $\psi(z_t) \in \mathbb{R}^{F\times N\times d^i}$ represents N spatially flattened video latent tokens each with d dimension across all the

Figure 3. **Token Visualization.** The figure shows the cross-attention map between text and video frame pooled from a downsampling layer in a spatial transformer block of VideoCrafter2 [4].

frames F, $W^{(i)}$ is learnable weight matrix on each attention head i, and, $\phi(y) \in \mathbb{R}^{1\times 77\times 1024}$ represents CLIP text encoding with 77 tokens and each with dimension 1024. After the computation, we get the cross attention map (A) of shape $\mathbb{R}^{F\times N\times 77}$, where each cell represents the attention score between pixel and text token across all the frames. In the Figure 3, we have shown the cross-attention map between a text token and a video frame pixel.

Text-to-Video (T2V). In the inference step, starting with a user-provided text description encoded with CLIP text embedding $\phi(y)$ and randomly sampled latent noise (z_t), denoising steps are performed using the DDIM framework to produce conditioned latent representations ($\hat{z_0}$). These denoised latent vectors are then passed through a pre-trained LVDM [9] video decoder to synthesize the final video.

Compositional Generation. For compositional video generation, we propose Video-ASTAR with Centroid Loss directly to the accumulated attention maps from the Spatial Transformer. Although ASTAR [1] was initially proposed for images, we extend it to video frames by adding the temporal dimension to the loss function. We also modified the mask generation method, replacing the graph component computation with mean thresholding *i.e.*, attention values below the mean are masked to zero and above the mean to one for each frame. The mask from the first video frame is used for the attention retention loss, resulting in stable video generation with rich details.

1) Attention Segregation Loss. At each denoising step t we have a pair of cross-attention maps (A^m, A^n) for each concept (nouns) and reduce the intersection-over-union value in all locations (i, j) for the pair of attention maps [1]. This encourages the attention map of each entity to have the least possible overlap. Here, $\hat{C}$ is the total number of token pairs.

$$\mathcal{L}_{seg} = \sum_{\substack{m,n\in\hat{\mathcal{C}}\\ \forall m>n}} \left[\frac{\sum_{ij} min([A_t^m]_{ij}, [A_t^n]_{ij})}{\sum_{ij}([A_t^m]_{ij} + [A_t^n]_{ij})}\right] \tag{6}$$

2) Attention Retention Loss. Given attention A^m at each time step $t-1$ and masked attention map M from time step t. The loss encourages maximizing the intersection-over-union value across all the pixels (i, j) of the attention map where the mask is set to one. Here, C is the total number of tokens.

$$\mathcal{L}_{ret} = \sum_{m\in\mathcal{C}} \left[1 - \frac{\sum_{ij} min([A_{t-1}^m]_{ij}, [M_t^m]_{ij})}{\sum_{ij}([A_{t-1}^m]_{ij} + [M_t^m]_{ij})}\right] \tag{7}$$

3) Centroid Loss. This loss encourages attention maps to close to each other. To achieve this, we first created P number of token pairs. The loss function is computed based on an attention map tensor $\hat{A}[i, j, t, h, w]$ with shape $(P, 2, T, H, W)$, where, $j \in \{0, 1\}$ is index for tokens a ($j = 0$) and b ($j = 1$) in each pair.

The row centroid (y-coordinate) for token j in pair i at frame t is computed:

$$RowCent_{[i,j,t]} = \left(\sum_{h=0}^{H-1}\left(\sum_{w=0}^{W-1}\hat{A}[i,j,t,h,w] \cdot (h+1)\right) / \sum_{h'=0}^{H-1} y_{\text{loc}}[h']\right) \tag{8}$$

The column centroid (x-coordinate) is computed similarly:

$$ColCent_{[i,j,t]} = \left(\sum_{w=0}^{W-1}\left(\sum_{h=0}^{H-1}\hat{A}[i,j,t,h,w] \cdot (w+1)\right) / \sum_{w'=0}^{W-1} x_{\text{loc}}[w']\right) \tag{9}$$

The 2D centroid for each token is:

$$CentPerFrame_{[i,j,t]} = \begin{bmatrix} RowCent_{[i,j,t]} \\ ColCent_{[i,j,t]} \end{bmatrix} \tag{10}$$

The loss is Euclidean distance between the centroids of tokens a ($j = 0$) and b ($j = 1$) for each pair i and frame t:

$$\mathcal{L}_{cent} = \sqrt{\left(CentPerFrame_{[i,0,t]} - CentPerFrame_{[i,1,t]}\right)^2} \tag{11}$$

Here, $y_{\text{loc}}[h] = h + 1$ for $h = 0, \ldots, (h_{attn} - 1)$, and $x_{\text{loc}}[w] = w+1$ for $w = 0, \ldots, (w_{attn}-1)$. The loss function that is used for optimization is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \mathcal{L}_{ret} + \mathcal{L}_{cent} \tag{12}$$

We use gradient descent to optimize latent z_t based on the

(a) (b)

Figure 4. **Qualitative Comparison.** (a) The figure illustrates the effect of different optimization functions with VideoCrafter2 [4], (b) shows the effect of different parameters on the Video-ASTAR loss.

loss above. The Figure 1 shows optimization flow (right) along with its effect (left). In the supplementary material, we have discussed the *concept selection strategy* for selecting nouns/verbs for the optimization process.

Multi-Action Video Generation (MAVG). In addition to compositional generation, text-to-video models also struggle to follow sequences of actions described in the text while maintaining semantic consistency across frames. FIFO [13] addresses this by proposing an optimization-based method for long video generation and extends it to support multi-prompt inputs. However, as shown in the top row of Figure 9 and Figure 10, FIFO often produces inconsistencies across frames. To address this, we propose two techniques:

1) Token Swapping. For multi-action sequence prompts that differ only in their actions while the rest of the text remains similar, we preserve the initial prompt-generated tokens and modify only the action-specific tokens. Specifically, we replace the action and end tokens (<eot>) from the first prompt with those from the second prompt and incorporate them into the second prompt. This approach ensures consistent background and context across the frames of the multi-action video, preventing abrupt changes.

2) Latent Interpolation. In FIFO [13] for generating the infinite frames, the authors have proposed a diagonal denoising method where new noise is input at the end (starting point of the denoising step) of the queue, and by the time it reaches the front (denoised latent variable), it can be used to synthesize the new video frames. In our work, we proposed a modification to this step where, instead of directly pushing a new noisy frame at the end of the queue, we push the weighted average of the last N frames of the queue with the new noisy frame.

Figure 5. **Qualitative Comparison Across Methods.** In this figure, we have shown the effect of different methods when integrated with the VideoCrafter2 [4] text-to-video generation network.

$$F_{noisy} = w_1 \times F_N + w_2 \times \epsilon_{\mathcal{N}(0,I)} \tag{13}$$

where epsilon is a new sampled noise from the standard normal distribution and w_1 and w_2 are weights for interpolation, which are kept as 0.4 and 0.6 for the proposed method. With these modifications in the FIFO pipeline, we also used the proposed losses for compositional video.

4. Experiment and Results

In the first part of this section, we have discussed different baseline methods used for the comparison. The second part

Methods	Spatial Relationship [84] (↑)	Multiple Objects [82] (↑)	Overall Consistency [93] (↑)	Temporal Consistency [259] (↓)
VideoCrafter2 [4]	43.99%	42.53%	**28.24%**	0.2899
VideoCrafter2 + Attend and Excite [2]	42.47%	33.31%	27.98%	0.5446
VideoCrafter2 + VICO [25]	42.35%	41.06%	27.90%	0.3798
VideoCrafter2 + Video-ASTAR (Ours)	**44.13%**	**50.66%**	27.33%	**0.2378**

Table 1. **Quantitative Comparison** with baseline methods using VBench [12] across three different prompt categories (total prompts in each category are shown in brackets []) on three random seed values. We report the average across all runs.

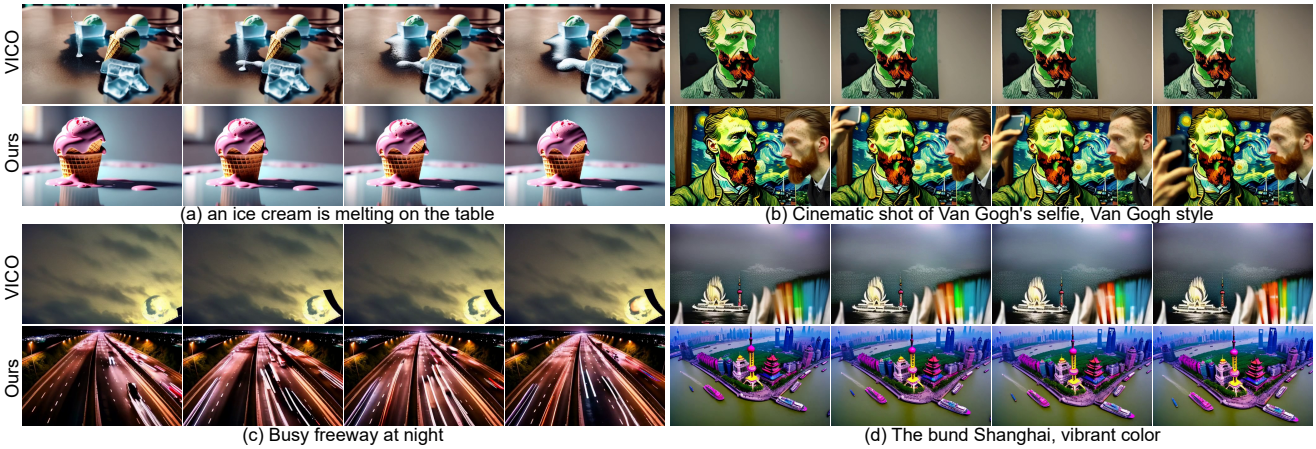

Figure 6. **Overall Consistency.** Qualitative comparison between the results generated using VICO [25] and our proposed optimization method for compositional video generation.

Configuration	Spatial (↑)	Multiple Object (↑)	Overall (↑)	Temporal (↓)
Video-ASTAR	**49.20%**	**61.51%**	**27.85%**	0.2061
w/o Segregation loss($\mathcal{L}_{seg}$)	47.38%	38.95%	27.12%	0.2156
w/o Retention loss ($\mathcal{L}_{ret}$)	43.18%	50.38%	26.95%	0.2125
w/o Centroid loss ($\mathcal{L}_{cent}$)	47.28%	56.33%	**27.85%**	**0.1994**

Table 2. **Loss Ablation.** The table shows quantitative analysis of the proposed framework with different loss configurations.

of the section discusses the evaluation criteria and ablations.

Baselines. We compare our method with the baseline VideoCrafter2 [4], a text-to-video generation model. Additionally, we evaluate against VICO [25], which enhances VideoCrafter2 by introducing a max-flow-based [7] optimization to refine the spatial-temporal layout of latent noise for better compositional control. We also consider VideoCrafter2 combined with the Attend-Excite mechanism [2], originally proposed for compositional image generation. In VICO, the Attend-Excite loss is adapted to guide the alignment between the generated video content and input text prompts, improving compositional fidelity [25].

Evaluation Metrics. To evaluate the compositional generation, we used VBench [12], consisting of different dimensions to measure the quality of generated videos. In our work, we have used three such dimensions, which are a) Spatial Relation: it measures if the generated videos follow the spatial relationship given in the text description between the objects, b) Multiple Object Composition: to measure if frames consist of all the objects mentioned in the text, and c) Overall Consistency: to evaluate how well the generated video align with given text prompt using ViCLIP [22]. Apart from these three dimensions, we also measure the generated video quality on the temporal dimension. To measure it, we extract the clip [18] feature ($\psi \in \mathbb{R}^d$) for each frame (F) and compute the sum of the differences between consecutive frames to get the perceptual difference.

$$\mathcal{D}_{percep} = \sum_{i=1}^{|F|} 1 - \cos(\psi_{F[:-1]}, \psi_{F[1:]}) \qquad (14)$$

For fair comparison, we rerun all the methods on three different random seeds and report the mean of them.

Frame Mask	Spatial (↑)	Multiple Object (↑)	Overall (↑)	Temporal (↓)
All Frames	38.34%	40.17%	**27.89%**	0.2213
Only First Frame	**39.92%**	**48.86%**	27.34%	**0.2143**

Table 3. **Frame Mask Ablation.** Shows the effect of using the first frame vs all frames mask during the optimization.

Implementation Details. Our method is optimization-based and requires no training. We use the FIFO implementation [13] with VideoCrafter2 as the backbone, which employs the DDIM sampler with 64 steps. For compositional video generation, we optimize latents for 25 steps using Video-ASTAR (with centroid loss) and vanilla gradient descent, with learning rate defined by a scale factor s_f=20

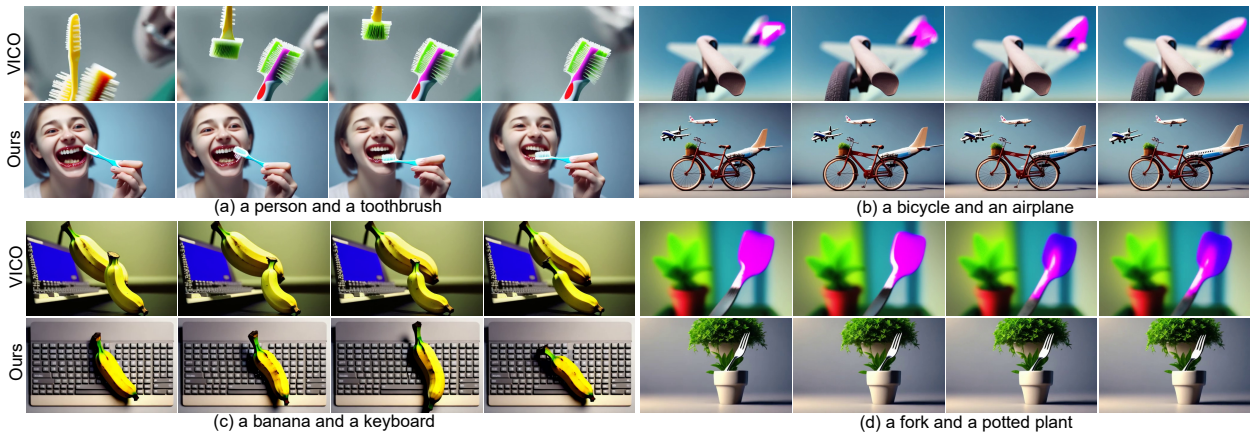

Figure 7. **Multiple Objects.** Qualitative comparison between the results generated using VICO [25] and our proposed optimization method for compositional video generation.

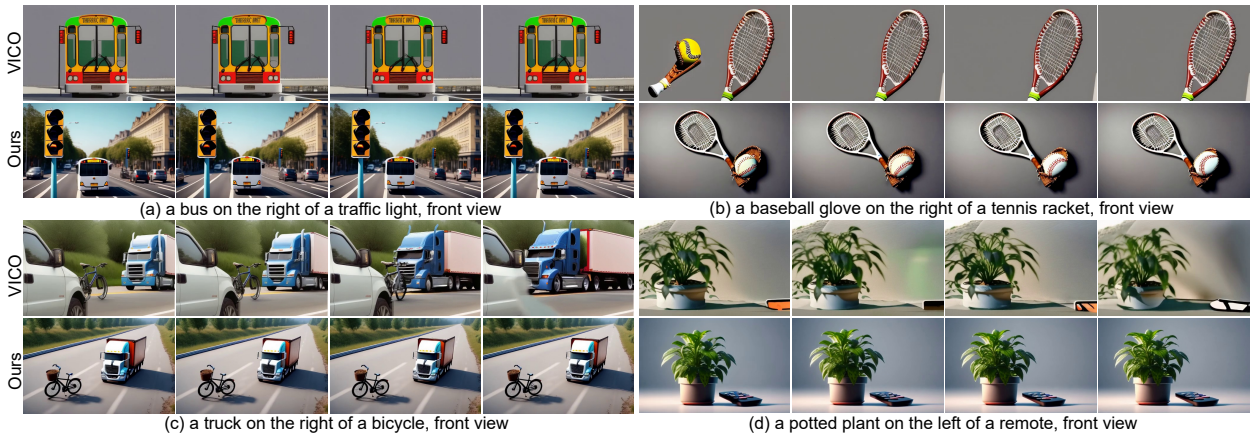

Figure 8. **Objects Spatial Relationship.** Qualitative comparison between the results generated using VICO [25] and our proposed optimization method for compositional video generation.

and scale range $s_r \in [1, 0.5]$. For multi-action prompts, we further apply token swapping and queue latent interpolation.

Quantitative Comparison. Table 1 shows that Video-ASTAR achieves competitive results and outperforms the baseline on most VBench dimensions. In particular, the Multiple Objects metric improves significantly, attributed to attention retention and centroid loss maintaining active and spatially consistent attention maps.

Qualitative Comparison. We also compare with VICO [25] under different seeds. As shown in Figs. 6, 7, and 8, our results are more consistent and coherent. Unlike VICO, we identified transformer blocks $\{9, 17, 18, 19\}$ as most effective and computed losses only from their cross-attention maps to produce stable results.

Ablation. We evaluate the effect of different optimization methods on VideoCrafter2 (VC2) [4]. As shown in Fig. 4(a), VC2 fails to generate the object "bench." Applying A&E loss [2] recovers it in the first prompt but fails in the second and shows alignment issues. In contrast, Video-ASTAR recovers missing entities in both prompts, with better contextual understanding and frame consistency. Table 2 reports the effect of different loss combinations, showing that $\mathcal{L}_{seg}$, $\mathcal{L}_{ret}$, and $\mathcal{L}_{cent}$ all contribute meaningfully (evaluated with seed 42). Here, $\mathcal{L}_{seg}$ reduces overlap between token attentions, $\mathcal{L}_{ret}$ maintains token focus across denoising steps, and $\mathcal{L}_{cent}$ improves compositional generation by enhancing entity interaction. Fig. 4(b–c) shows that fixed thresholding misses entities, while mean-based thresholding captures them more reliably. Figs. 4(d–e) further show that reusing a binary mask from the first frame preserves context (e.g., consistently generating a "big rock"). Table 3 confirms that using the first-frame mask outperforms recomputing masks at each frame. Finally, Fig. 5 compares optimization methods on VC2, illustrating that our approach better captures prompt entities. More results and details are provided in the supplement.

Multi-Action. We conducted a study on multi-action sequence generation, where the input prompts contain multiple actions while keeping the surrounding context unchanged. As shown in Figure 9 and Figure 10, FIFO [13]

(a) in a quiet park an old man wearing a brown hat walking on the path 4k high resolution, in a quiet park an old man wearing a brown hat sitting on the path 4k high resolution

(b) in a quiet forest a tiger is walking on the path with a big rock stone aside 4k high resolution, in a quiet forest a tiger is sitting on a big rock stone aside the path 4k high resolution

Figure 9. **Token Swapping.** Qualitative comparison between the results generated using the FIFO [13] baseline and with token swapping.

(a) in a quiet park an old man wearing a brown hat walking on the path with a bench aside 4k high resolution, in a quiet park an old man wearing a brown hat sitting on a bench aside the path 4k high resolution

Figure 10. **Latent Interpolation.** Qualitative comparison between the results generated using the FIFO [13] baseline and with our proposed optimization method for MAVG. In our case, the bench remains consistent and avoids hallucination.

exhibits inconsistency in long video sequences. In contrast, our proposed methods, token swapping and latent interpolation, enable FIFO to generate coherent action sequences with consistent object presence. For example, in Figure 10, instead of hallucinating a new bench, the ”old man” continues to interact with the existing bench in frames.

Limitations. We find that with Video-ASTAR, we do get some minor loss in overall consistency as seen in Table 1 as a tradeoff for having multiple objects in the prompt.

5. Conclusion

This work tackles the challenge of compositional video generation from a text prompt by employing a straightforward, training-free method using a pre-trained video diffusion model. We have introduced Video-ASTAR with centroid loss for the optimization process over the cross-attention map. With comprehensive empirical results and ablation studies, we show its effectiveness over baselines. In addition to compositional generation, we study the multi-action video generation problem and have proposed a token swapping and latent interpolation method for generating consistent, action-specific videos, ensuring that the transitions between actions are smooth and logical. Our approach emphasizes the potential of optimization-based techniques for achieving controllable and grounded video generation grounded in composition, all without requiring additional training or fine-tuning of large-scale diffusion models, thereby streamlining the process of video generation in a way that is both efficient and effective. This work will pave the way for further work on control of T2V models.

6. Acknowledgment

The work is done as part of the internship at Adobe by Prajwal Singh. Additionally, the work was supported by the Prime Minister Research Fellowship, awarded to Prajwal Singh (PMRF2122-2557), and by the Jibaben Patel Chair in Artificial Intelligence, held by Shanmuganathan Raman.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023. 1, 2, 4

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1, 2, 6, 7

[3] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3

[4] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 1, 2, 3, 4, 5, 6, 7

[5] Hyungjin Chung, Dohun Lee, and Jong Chul Ye. Acdc: Autoregressive coherent multimodal generation using diffusion correction. *arXiv preprint arXiv:2410.04721*, 2024. 1

[6] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 1

[7] Lester R Ford Jr and Delbert R Fulkerson. Maximal flow through a network. *Canadian journal of Mathematics*, 8: 399–404, 1956. 6

[8] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pages 393–411. Springer, 2024. 1

[9] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 4

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1

[11] Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. Genmac: compositional text-to-video generation with multi-agent collaboration. *arXiv preprint arXiv:2412.04440*, 2024. 3

[12] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6

[13] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *arXiv preprint arXiv:2405.11473*, 2024. 1, 3, 5, 6, 7, 8

[14] Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. Videorepair: Improving text-to-video generation via misalignment evaluation and localized refinement. *arXiv preprint arXiv:2411.15115*, 2024. 2

[15] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 1

[16] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024. 1

[17] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 1

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 6

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[20] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[21] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, Di ZHANG, et al. Videotetris: Towards compositional text-to-video generation. *Advances in Neural Information Processing Systems*, 37:29489–29513, 2024. 1, 2

[22] 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. 6

[23] Zun Wang, Jialu Li, Han Lin, Jaehong Yoon, and Mohit Bansal. Dreamrunner: Fine-grained compositional story-

to-video generation with retrieval-augmented motion adaptation. *arXiv preprint arXiv:2411.16657*, 2024. 3

[24] Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. *arXiv preprint arXiv:2410.10816*, 2024. 1

[25] Xingyi Yang and Xinchao Wang. Compositional video generation as flow equalization. *arXiv preprint arXiv:2407.06182*, 2024. 1, 2, 6, 7

[26] Hongyu Zhang, Yufan Deng, Shenghai Yuan, Peng Jin, Zesen Cheng, Yian Zhao, Chang Liu, and Jie Chen. Magiccomp: Training-free dual-phase refinement for compositional video generation. *arXiv preprint arXiv:2503.14428*, 2025. 2