WHY DO TEXT-TO-VIDEO MODELS STRUGGLE WITH SPATIAL RELATIONS?

Anonymous authors
Paper under double-blind review

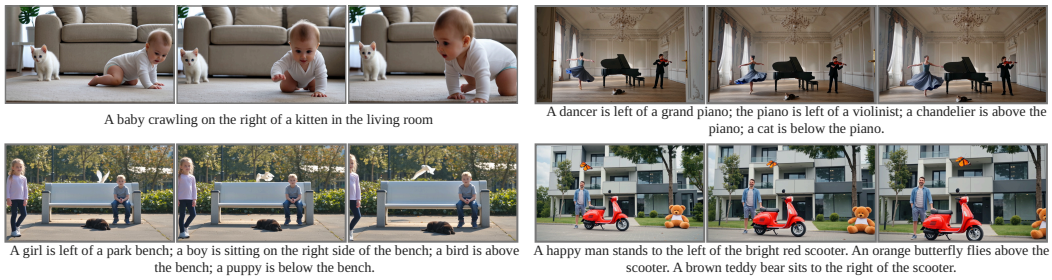

Figure 1: **Teaser of SPARC.** Training-free spatial control from text alone for generating videos with specified spatial relations.

ABSTRACT

Text-to-video (T2V) models have achieved impressive generation quality, yet they remain unreliable in establishing spatial relations among multiple objects. In this work, we investigate the sources of spatial-relation failures in transformer-based T2V models. Our analysis identifies two key limitations. First, the frozen text encoder weakly represents the implicit complementary spatial role of reference objects. Second, joint conditioning produces ambiguous object grounding during early denoising. Based on these observations, we propose Spatial Patching and Anchored Relational Conditioning (SPARC), a training-free approach for improving spatial controllability directly from text prompts. Our method first applies Complementary Spatial Token Patching to strengthen the spatial representation of reference objects. It then uses Anchor-to-Graph Denoising to progressively establish relations around a shared spatial anchor. Both components operate at inference time with frozen model parameters and do not require additional spatial inputs. Experiments across two T2V backbones and multiple spatial benchmarks demonstrate consistent improvements in spatial controllability. As a training-free method, SPARC achieves a 12.5% relative improvement in the T2V-CompBench spatial score over the second-best method. Our code will be released upon acceptance.

1 INTRODUCTION

In recent years, video generation models have become sufficiently advanced to generate realistic videos. These models have also become commercially available for video generation with synchronized audio (Collins, 2025). Yet, these advances do not guarantee that text-to-video (T2V) models can correctly interpret and realize the compositional structure specified in a text prompt (Sun et al., 2025). This limitation is particularly evident in prompts that specify spatial relations among multiple objects. State-of-the-art T2V models can generate the requested objects, but often fail to establish the intended spatial relationships (Huang et al., 2024). Even when the correct relation is established, it may not remain consistent throughout the video (Wu et al., 2025). Reliable spatial control is important not only for compositional accuracy, but also for the efficient and practical use of T2V models. In this regard, it is necessary to infer the correct placement of objects directly from text

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

without requiring external spatial information. Despite its significance, the underlying reasons why current T2V models fail to establish spatial relationships remain unclear.

Research on this subject has generally focused on a few approaches to improve spatial control. Existing methods use external spatial conditions such as layouts and bounding boxes (Feng et al., 2025), masks (Yariv et al., 2025), and object trajectories (He et al., 2025). Although these methods improve accuracy, they also come with some limitations. Using these methods in real-world scenarios requires additional effort to provide external inputs (Rahman et al., 2026). Moreover, fine-tuning or training is required for these spatial control methods when applying them to future models, reducing their feasibility and adaptability (Zhang et al., 2025a). An ideal solution would infer the correct spatial configuration directly from the given prompt without relying on predefined spatial constraints. This motivates a training-free approach that uses only the text prompt for spatial control. Achieving this goal first involves identifying why existing T2V models remain unreliable when processing spatial-relation prompts.

To understand this failure, we analyze how spatial roles are represented and grounded in transformer-based T2V models during early generation. First, we find a strong asymmetry between explicitly stated and implicitly inferred relational roles in the text representation. Second, we observe ambiguous object grounding when multiple objects and relations are processed simultaneously during early denoising. These analyses reveal limitations in both the representation of relational roles and the grounding of multiple object relations.

Based on these observations, we propose **SPARC** (Spatial Patching and Anchored Relational Conditioning), a training-free method to improve spatial controllability in transformer-based T2V models. Our method addresses the identified limitations through two complementary mechanisms. First, we modify text embeddings with the Complementary Spatial Token Patching algorithm to strengthen the representation of spatial roles for objects. Second, we propose Anchor-to-Graph Denoising to reduce interference during the grounding of multiple object relations in the early generation steps. SPARC improves spatial alignment without requiring additional training, bounding boxes, masks, or other external spatial inputs.

Our main contributions are listed as follows: (1) We analyze transformer-based T2V models under spatial relation prompts and identify two sources of failure: weak representation of spatial roles in object embeddings and ambiguous grounding of multiple object relations. (2) We propose SPARC, a training-free approach that improves spatial controllability using only text prompts. (3) We conduct extensive experiments across multiple T2V backbones and spatial benchmarks, showing consistent improvements in spatial controllability, with a 12.5% relative improvement in the T2V-CompBench spatial score.

2 Related Work

2.1 Spatially Controllable Video Generation

Spatial controllability in video generation encompasses where objects appear in a frame and how they move over time. Several approaches leverage explicit spatial conditions either from users or from external algorithms. Examples of these spatial conditions include layouts (Feng et al., 2025), bounding boxes (He et al., 2025), segmentation masks (Yariv et al., 2025), object trajectories (Rahman et al., 2026), depth maps (Chen et al., 2023), and pose sequences (Zhang et al., 2024). These conditions provide the model with explicit cues about where the object should appear and how it is moving over time. Although spatial conditions substantially improve object positioning in generated videos, they depend on external planning or guidance, introducing additional complexity into the video generation process (Zhang et al., 2026).

Another direction of research is training or adapting a model for spatial control. These methods propose control modules or spatial adapters that link conditions to the video generator (Zhang et al., 2025b; Lin et al., 2025). Their aim is to improve the accuracy of the model with minimal training (Guo et al., 2024). However, such components are often tied to particular model architectures or require model-specific adaptation, which limits their direct transfer to new T2V backbones (Lin et al., 2025). In contrast, a preferable approach would improve spatial control without modifying or retraining the underlying generation model.

To reduce architecture-specific adaptation, researchers have explored training-free methods that manipulate the inference process. Common training-free strategies include manipulating attention maps (Xu et al., 2026), modifying latent variables (Qiu et al., 2024), and injecting spatial guidance during denoising (Sun et al., 2026). Nevertheless, these methods often require user-provided guidance or external assistance. Despite these developments, the reasons why T2V models fail to establish spatial relations when relying solely on text prompts are not yet fully understood.

2.2 Spatial Relation Understanding in Generative Models

Understanding spatial relations in generative models involves the correct positioning of objects relative to each other. Existing research has focused mainly on single-instance generation (Zhou et al., 2024; Ian Chan et al., 2026). However, when two objects are introduced into a frame with relative spatial positions, generative models still struggle to satisfy the intended spatial relationship (Zhang et al., 2025a). Problems such as inconsistent object placement and the failure to generate one of the objects can occur in the output, as illustrated in Fig. 5. This challenge becomes more difficult as the number of objects and relations increases (Yiflach et al., 2026).

A further challenge is maintaining the spatial relationship throughout the video (Sun et al., 2025). Object movement and camera motion may cause the spatial arrangement to become inconsistent over time. Understanding this behavior is important for both evaluating models and improving spatially consistent generation. Existing benchmarks recognize this gap and develop evaluation methods based on relational prompts (Lin et al., 2024). They generally use predefined prompts and apply object detection algorithms to identify the objects in generated videos (Huang et al., 2024; Sun et al., 2025). The detected bounding boxes are compared based on their center positions to determine whether the spatial relation is satisfied. However, these evaluations primarily reveal whether a model satisfies the specified relation and provide limited insight into why the model fails.

Previous studies have mainly focused on improving or evaluating spatially controllable generation (Chen et al., 2023; Sun et al., 2025). The sources of spatial-relation failures have received comparatively less attention (Liu et al., 2025). In particular, the behaviors of T2V models that lead to incorrect object placement and inconsistent spatial relationships remain insufficiently investigated. This makes it difficult to develop a general and training-free solution. Our work addresses this gap by investigating why transformer-based T2V models fail to establish spatial relations directly from text prompts and by proposing SPARC to address these failures.

3 Diagnosing Spatial-Relation Failures

3.1 Relational Video Generation

We consider a relational text prompt $p = (o_t, r, o_r)$, where o_t and o_r denote the target and reference objects, respectively, and r denotes their desired spatial relation, e.g., *left of*, *right of*, *above*, *behind*, etc. For each relation r, we denote its complementary relation by $\bar{r}$. Then, we define the semantically equivalent inverse prompt as $p^{\text{inv}} = (o_r, \bar{r}, o_t)$, such that $p^{\text{inv}} \equiv p$. For example, $p =$ "a cat is left of a dog" is equivalent to $p^{\text{inv}} =$ "a dog is right of a cat."

Let E denote a frozen text encoder that produces the contextualized token representations $H = E(p) = \{h_1, \ldots, h_L\}$, where L denotes the number of text tokens. We denote the contextualized representations of the target and reference objects by h_{tar} and h_{ref}, respectively. Let v_θ denote a pretrained video diffusion transformer (DiT) (Wan et al., 2025) with frozen parameters θ, which generates a video V conditioned on H through an iterative generation process.

The goal of relational video generation is to generate a video V such that $\mathcal{R}(V, o_t, o_r) = r$. Here, $\mathcal{R}(\cdot)$ represents the spatial relation between the corresponding objects in the video. Despite explicitly specifying r in the input prompt, pretrained video DiTs frequently fail to establish or maintain the desired spatial relation, as further demonstrated by the prompt-level reasoning experiments in Appendix D.7. We next investigate this failure from two perspectives: relational-role representation in the textual condition and relational grounding during video generation.

3.2 Relational-Role Asymmetry

To establish the intended spatial relationship, the textual representation must contain sufficient information about the relational roles of the objects. We therefore investigate whether the frozen text encoder represents spatial information for both objects. For this purpose, we train a six-class single-layer linear probe to predict the spatial role encoded in each object embedding. The six classes are *left*, *right*, *above*, *below*, *front*, and *behind*. In the two-object setting, we apply the probe separately to h_{tar} and h_{ref}. For $p = (o_t, r, o_r)$, r is explicit for o_t, while $\bar{r}$ must be inferred for o_r. For example, for $p = (\text{cat}, \textit{left}, \text{dog})$, $r = \textit{left}$ is explicit for the cat, whereas $\bar{r} = \textit{right}$ must be inferred for the dog.

We use the probe to examine h_{tar} and h_{ref} on the spatial prompts from T2V-CompBench (Sun et al., 2025). A prediction is considered correct if it matches r for h_{tar} or $\bar{r}$ for h_{ref}. Probe accuracy is computed as the percentage of correctly predicted roles across the evaluation prompts. As shown in Tab. 1, r is predicted from h_{tar} with 88.33% accuracy, whereas $\bar{r}$ is predicted from h_{ref} with only 32.78%. This reveals a strong asymmetry between explicitly stated and implicitly inferred spatial roles.

Table 1: **Relational-role asymmetry.** Linear-probe and reference-role conditioning results.

Linear Probe	**Acc. (%)**
o_t in p	88.33
o_r in p	32.78
o_r in p^{inv}	**89.17**
Role Conditioning Spatial Score (%)	
h_{ref} conditioned on r	43.87
Unmodified h_{ref}	47.83
h_{ref} conditioned on $\bar{r}$	**58.58**

The weak encoding of $\bar{r}$ in h_{ref} has two possible explanations. The encoder may lack the corresponding spatial knowledge, or it may fail to infer the role when it is implicit. To investigate these, we also evaluate the linear probe on the semantically equivalent inverse prompt $p^{\text{inv}} = (o_r, \bar{r}, o_t)$, which makes $\bar{r}$ explicit for o_r. Under p^{inv}, the probe accuracy for o_r increases from 32.78% to 89.17%. This indicates that the encoder contains the relevant spatial knowledge, but fails to reliably infer the implicit role of o_r.

We next examine whether the role encoded in h_{ref} affects spatial generation. The unmodified h_{ref} obtained from p serves as the baseline. We then construct two auxiliary prompts, $(o_r, \bar{r})$ and (o_r, r). To assess the effect of each role, we separately replace the baseline h_{ref} with the o_r embedding obtained from each auxiliary prompt. We evaluate the generated videos using the T2V-CompBench spatial score (Sun et al., 2025), which measures whether the generated objects satisfy the requested spatial relation. As shown in Tab. 1, conditioning h_{ref} on the correct role $\bar{r}$ increases the spatial score from 47.83% to 58.58%, whereas conditioning it on r reduces the score to 43.87%. These results show that the spatial role encoded in h_{ref} directly affects the generated relation, motivating our *Complementary Spatial Token Patching*. Additional training details, probe statistics, and robustness analyses are provided in Appendix A.1.

3.3 Delayed Relational Grounding

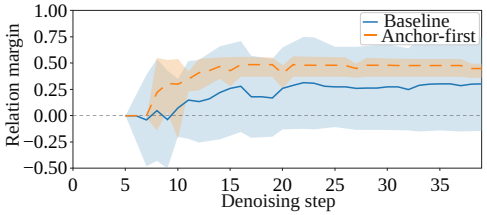

Figure 2: **Relational grounding trajectory.** Signed relation margin across denoising steps under baseline and anchor-first conditioning.

Even when spatial information is encoded in the textual condition, the model must still translate this information into the output video. Conventional transformer-based T2V models use a single encoded sequence throughout generation (Kong et al., 2025). This sequence contains all objects and their spatial relations. Therefore, the model must place the objects and establish their relations simultaneously. We hypothesize that simultaneously grounding multiple objects and relations produces overlapping attention patterns. This can lead to ambiguous object-to-region assignments, preventing the model from establishing a clear spatial ordering between the objects during the early denoising steps. To test this hypothesis, we compare two conditioning settings. In the baseline

setting, the full relational prompt conditions all denoising steps. In the anchor-first setting, we condition the early denoising steps on the reference object before introducing the full relational prompt. We analyze both denoising trajectories on multi-object prompts.

We track the emergence of the requested relation along the denoising trajectory. At denoising step t, let z_t denote the noisy video latent, σ_t the noise level associated with step t, and $v_\theta(z_t, t, H)$ the prediction of the frozen video DiT with parameters θ, conditioned on the textual representation H. We estimate the corresponding clean latent as

$$\hat{z}_0^{(t)} = z_t - \sigma_t v_\theta(z_t, t, H), \tag{1}$$

where $\hat{z}_0^{(t)}$ denotes the estimated clean latent at denoising step t. We then decode $\hat{z}_0^{(t)}$ into video frames and localize the corresponding objects. From their relative positions, we compute the signed relation margin. A positive margin indicates the correct spatial ordering, whereas a negative margin indicates an incorrect ordering. A margin near zero indicates weak relational grounding.

We evaluate 30 two-object prompts spanning six spatial relations: *left*, *right*, *above*, *below*, *front*, and *behind*. We use five prompts per relation and generate each prompt with five random seeds. At each denoising step, we first compute the relation margin for each prompt-seed pair and then average across all evaluated prompts and seeds. The solid curve in Fig. 2 shows the mean relation margin, while the shaded region denotes one standard deviation across prompt-seed pairs. The relational margin remains near zero during early denoising and becomes only weakly positive during later steps. Across the late generation steps, the baseline achieves a mean margin of only 0.062, indicating that the requested relation remains weakly established. In contrast, anchor-first conditioning produces an earlier and stronger positive relation margin, suggesting that early object anchoring facilitates relational grounding. For the baseline setting, attention visualizations show substantial overlap among object-token attention maps during early denoising, suggesting ambiguous object-to-region assignments. Additional attention visualizations are provided in Appendix A.2. Together, these observations suggest that ambiguous early object grounding is associated with the delayed and weak formation of the requested spatial relation.

4 SPARC

Our analysis identifies two sources of spatial-relation failure: asymmetric encoding of complementary spatial roles and ambiguous relational grounding during early denoising. We address these failures with two training-free components: Complementary Spatial Token Patching and Anchor-to-Graph Denoising. Complementary Spatial Token Patching addresses relational-role asymmetry by strengthening the spatial information associated with reference objects. Anchor-to-Graph Denoising addresses ambiguous relational grounding by progressively introducing object relations around a spatial anchor during early denoising. Both components operate during inference with a frozen text encoder and video DiT, as illustrated in Fig. 3.

We first use an off-the-shelf scene-graph parser (Schuster et al., 2015) to extract the objects and spatial relations from the input prompt. The resulting scene graph is defined as

$$\mathcal{G} = (\mathcal{N}, \mathcal{E}), \qquad \mathcal{E} = (e_m)_{m=1}^{M}, \qquad e_m = \left(o_t^{(m)}, r_m, o_r^{(m)}\right), \tag{2}$$

where $\mathcal{N}$ denotes the set of object nodes and $\mathcal{E}$ denotes the set of directed relation edges. Object nodes retain their associated modifiers and actions, while $\mathcal{E}$ represents only the spatial relations between them. M denotes the number of extracted relations. The objects $o_t^{(m)}$ and $o_r^{(m)}$ are the target and reference objects of the m-th relation, respectively.

4.1 Complementary Spatial Token Patching

Consider an edge $e = (o_t, r, o_r)$ associated with the scene graph. The spatial role r of the target object is explicitly stated, whereas the complementary role $\bar{r}$ of the reference object remains implicit. Our analysis in Sec. 3.2 shows that the frozen text encoder represents such implicit roles weakly. We therefore construct the semantically equivalent inverse prompt

$$p^{\text{inv}} = (o_r, \bar{r}, o_t), \qquad H^{\text{inv}} = E(p^{\text{inv}}). \tag{3}$$

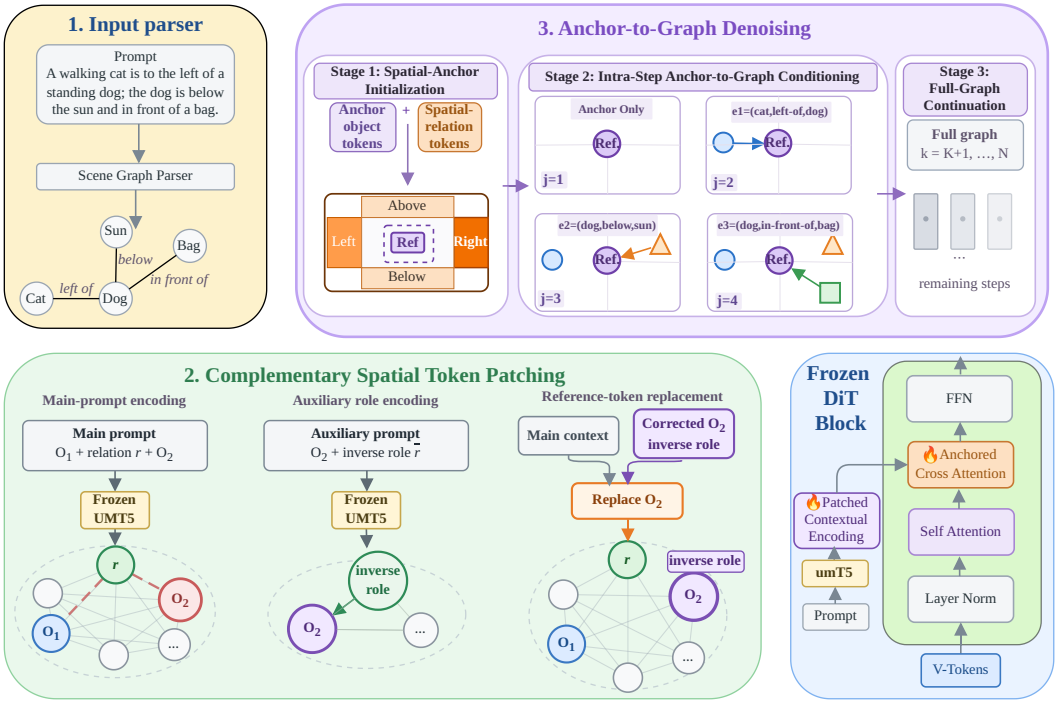

Figure 3: **Overview of SPARC.** The input prompt is parsed into a scene graph, followed by Complementary Spatial Token Patching and Anchor-to-Graph Denoising for training-free spatial control.

In p^{inv}, the original reference object o_r occupies the target position, making its complementary role $\bar{r}$ explicit. Let $h^{\text{inv}}_{\text{tar}}$ denote the contextualized representation of this object in H^{inv}. We replace its original reference-object representation in the main context with

$$\widetilde{h}_{\text{ref}} = h^{\text{inv}}_{\text{tar}}. \tag{4}$$

For a reference object spanning multiple text tokens, this replacement is applied to all token representations corresponding to o_r. All other token representations remain unchanged. We apply this operation independently to every relational edge in $\mathcal{E}$, producing the patched contextual representation $\widetilde{H}$.

4.2 Anchor-to-Graph Denoising

Complementary Spatial Token Patching improves the textual spatial representation, but does not address the ambiguous relational grounding identified in Sec. 3.3. We therefore introduce Anchor-to-Graph Denoising, which progressively introduces object relations around a shared spatial anchor during the first K denoising steps. Given the scene graph $\mathcal{G}$, we select the object with the highest node degree as the spatial anchor o_a. Starting from o_a, the conditioning context is progressively expanded to the complete scene graph.

As illustrated in Fig. 3, Anchor-to-Graph Denoising consists of three stages. In **Stage 1**, Spatial-Anchor Initialization adds a short denoising phase equivalent to 2.5% of the standard denoising steps, using only the anchor-object tokens and the spatial-relation terms (e.g., *left* or *right*) before the main denoising process begins. This establishes an initial spatial structure for subsequent relational grounding. In **Stage 2**, Intra-Step Anchor-to-Graph Conditioning is applied during the first K denoising steps. At each denoising step, the procedure starts from o_a traverses the relation edges in the order returned by the scene-graph parser. Each edge updates the hidden representation through the modified cross-attention, and the resulting representation is passed to the next edge. After all edges are processed, we complete the DiT forward pass and perform the latent update. In **Stage 3**, the remaining denoising steps use the complete patched prompt context $\widetilde{H}$. We use $K = 4$ in all experiments and provide the full algorithm and sensitivity analysis in Appendix B.

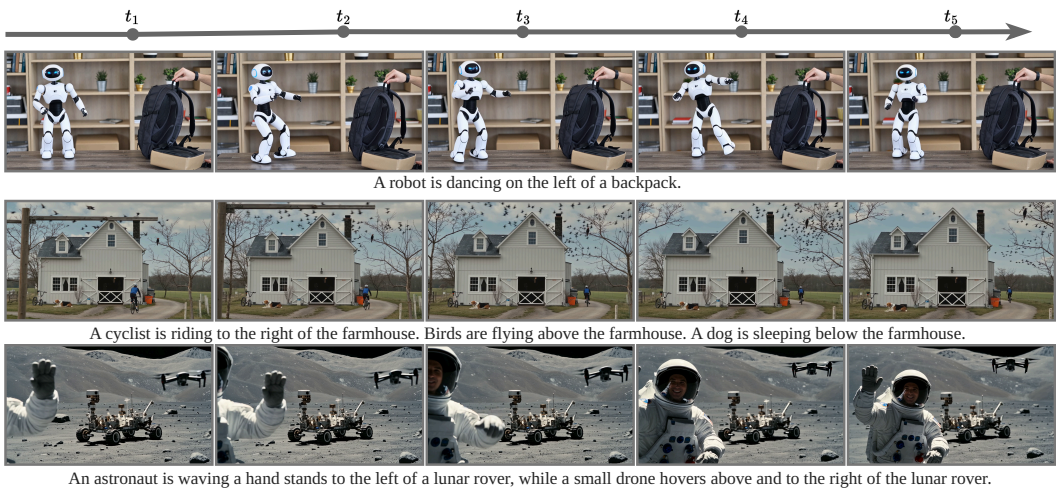

Figure 4: **Qualitative results of SPARC.** Each row shows five uniformly sampled frames from a generated video, illustrating spatial consistency over time.

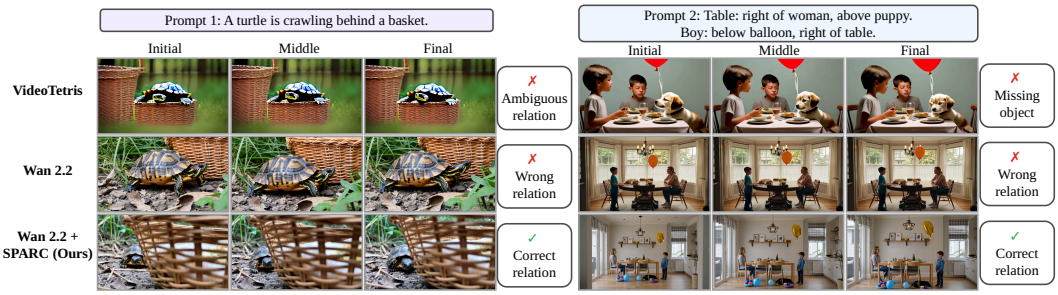

Figure 5: **Qualitative comparison with representative baselines.** Initial, middle, and final frames are shown for each generated video. SPARC more consistently satisfies and maintains the requested spatial relations over time.

5 Experiments

Baselines. We compare SPARC with a broad set of representative T2V systems, including CogVideoX (Yang et al., 2025), HunyuanVideo (Kong et al., 2025), Kandinsky (Arkhipkin et al., 2026), LTX-Video (HaCohen et al., 2024), LVD (Lian et al., 2024), Mochi (Team, 2024), NOVA (Deng et al., 2025), Pyramid Flow (Jin et al., 2025), Tora (Zhang et al., 2025b), VideoCrafter2 (Chen et al., 2024), VideoTetris (Tian et al., 2024), and Wan 2.2 (Wan et al., 2025). We evaluate all baselines under the same two-object and multi-object spatial-relation settings.

Experimental Setup. We evaluate spatial-relation generation using VBench (Huang et al., 2024) and T2V-CompBench (Sun et al., 2025), which provide established two-object spatial-relation prompts. To further evaluate more complex relational scenes, we construct a multi-relation setting from these benchmarks. We combine relational prompts to form scenes containing 3–5 objects and multiple spatial relations. For each T2V system, we use the official model checkpoint and the inference configuration recommended by the authors. The same model-specific generation settings are used for both the two-object and multi-relation evaluations.

Evaluation Metrics. We evaluate spatial-relation accuracy using four complementary metrics. For VBench (Huang et al., 2024), we follow its spatial-relationship evaluation, which detects the specified objects and determines their relative positions from the predicted bounding boxes. For T2V-CompBench (T2V-C) (Sun et al., 2025), we use its detection-based spatial metric based on GroundingDINO (Liu et al., 2024), with segmentation and depth estimation additionally used for depth-based relations such as *in front of* and *behind*. We further report VISOR (Wang et al., 2026), which evaluates whether both objects are detected and whether their bounding-box centers satisfy

Table 2: **Spatial controllability comparison.** Results on the T2V-CompBench and VBench spatial prompts using four spatial evaluators. Avg. denotes their unweighted mean. † denotes methods specifically designed for spatial/layout control. Rows highlighted in gray indicate our method, SPARC. Higher is better for all metrics. Best and second-best results are shown in **bold** and underlined, respectively.

Method	T2V-CompBench Prompts					VBench Spatial Prompts				
	T2V-C ↑	VBench ↑	VQA ↑	VISOR ↑	Avg. ↑	T2V-C ↑	VBench ↑	VQA ↑	VISOR ↑	Avg. ↑
CogVideoX	51.25	25.78	8.26	37.82	30.78	56.16	51.80	34.73	46.35	47.26
HunyuanVideo	50.31	26.61	9.44	39.97	31.58	46.23	59.77	4.96	30.73	35.42
LTX-Video	55.96	25.96	8.76	50.93	35.40	49.70	61.68	11.33	34.30	39.25
LVD†	5.99	13.78	12.42	50.46	20.66	47.91	37.02	29.82	38.24	38.25
Mochi	58.18	27.46	14.38	51.72	37.94	65.44	63.02	45.30	55.87	57.41
NOVA	55.64	21.97	11.54	53.49	35.66	61.94	57.10	31.87	49.18	50.02
Pyramid Flow	52.80	15.39	7.38	43.28	29.71	47.05	34.43	8.74	34.82	31.26
Tora†	46.28	19.67	4.37	28.96	24.82	41.15	43.95	8.41	27.60	30.28
VideoCrafter2	46.66	9.09	4.69	43.33	25.94	35.52	34.84	3.21	26.04	24.90
VideoTetris†	52.26	9.19	7.08	53.59	30.53	44.30	40.03	10.09	35.49	32.48
Kandinsky	57.28	31.68	13.73	58.39	40.27	55.69	72.22	32.84	56.85	54.40
Kandinsky+SPARC	61.53	31.55	15.49	60.63	42.30	64.73	**75.80**	**54.12**	**62.64**	**64.32**
Wan 2.2	63.90	49.85	16.31	65.63	48.92	63.08	66.94	29.00	48.29	51.83
Wan 2.2+SPARC	**71.87**	**54.13**	**23.71**	**77.37**	**56.77**	**69.43**	68.75	45.89	59.52	60.90

the requested spatial relation, and VQAScore (Lin et al., 2024), which measures prompt–visual alignment through visual question answering. For video-level evaluation, frame-level scores are averaged across the sampled frames of each generated video. All metrics are originally bounded in $[0, 1]$ and are multiplied by 100 for readability. We additionally report Avg., the unweighted mean of the four metrics.

5.1 Spatial Controllability Results

Two-object spatial controllability. Table 2 compares SPARC with representative T2V models on the spatial prompts of T2V-CompBench (Sun et al., 2025) and VBench (Huang et al., 2024). SPARC yields substantial improvements on several individual spatial evaluators. For example, on VBench spatial prompts, Kandinsky+SPARC improves VQAScore by 21.28 points, while Wan2.2+SPARC improves VQAScore by 16.89 points. On T2V-CompBench set, Wan2.2+SPARC further improves VISOR by 11.74 points. On the T2V-C metric, Wan2.2+SPARC improves the score from 63.90 to 71.87, corresponding to a 12.5% relative improvement over the second-best method. Consistent gains in both Wan2.2 and Kandinsky indicate that the effectiveness of SPARC is not restricted to a single T2V backbone. Overall, these results suggest that strengthening complementary spatial roles and relational grounding improves spatial alignment across different backbones, benchmarks, and evaluators.

Multi-relation spatial controllability. Table 3 evaluates the more challenging multi-object setting, where each prompt contains multiple objects and spatial relations. SPARC yields consistent gains on Wan2.2 across all four metrics. In particular, Wan2.2+SPARC improves VQAScore by 12.22 points, T2V-C by 5.82 points, and VISOR by 5.06 points, achieving the best overall performance across all four evaluators. Kandinsky+SPARC also shows a substantial improvement in VQAScore, increasing it by 17.86 points. These gains are reflected in the aggregate performance: Wan2.2+SPARC increases the average score from 42.72 to 49.65, corresponding to a 16.2% relative improvement, while Kandinsky+SPARC increases the average score from 33.01 to 38.24, corresponding to a 15.8% relative improvement. Overall, these results suggest that SPARC remains beneficial as relational complexity increases. Additional analyses, including per-relation results for each spatial relation and performance across different numbers of objects, are provided in Appendix D.

Qualitative results. Figures 4 and 5 further illustrate the spatial improvements. As shown in Fig. 4, SPARC preserves the specified object arrangements across uniformly sampled video frames, including scenes with multiple simultaneous relations. In Fig. 5, representative baselines exhibit ambiguous or reversed relations and may omit requested objects, whereas SPARC establishes the specified relations and maintains them from early to late frames. These examples qualitatively sup-

Table 3: **Multi-relation spatial controllability.** Results on multi-relation prompts using four spatial evaluators.

Method	T2V-C	VBench	VQA	VISOR	Avg.
CogVideoX	51.49	9.97	21.04	28.52	27.76
HunyuanVideo	55.10	7.59	19.61	37.73	30.01
LTX-Video	57.59	9.17	9.00	33.37	27.28
LVD	46.61	4.86	12.91	30.56	23.74
Mochi	60.87	4.52	43.93	49.21	39.63
NOVA	52.18	8.54	15.57	23.75	25.01
Pyramid Flow	60.30	6.86	13.30	40.69	30.29
Tora	51.66	11.34	11.99	34.95	27.49
VideoCrafter2	54.07	4.32	3.59	43.24	26.31
VideoTetris	55.21	2.09	9.28	40.93	26.88
Kandinsky	60.13	11.76	12.15	48.01	33.01
Kandinsky+SPARC	60.40	11.32	30.01	51.21	38.24
Wan 2.2	67.45	10.34	32.99	60.09	42.72
Wan 2.2+SPARC	**73.27**	**14.96**	**45.21**	**65.15**	**49.65**

Table 4: **Component ablation of SPARC.** Results on multi-relation prompts using Wan2.2 as the base model. The 41-step baseline controls for the additional inference cost of SPARC.

Variant	T2V-C ↑	VBench ↑
Base (40 steps)	67.45	10.34
Base (41 steps)	67.48	11.19
CSP only	71.95	14.50
A2G Stage 1	72.11	13.95
A2G Stage 2	72.22	8.24
A2G	72.68	11.81
Base+SPARC	**73.27**	**14.96**

port the improvements observed in the quantitative evaluation. For a more comprehensive qualitative comparison, please refer to the supplementary video.

5.2 Ablation Studies

Table 4 analyzes the contribution of each component of SPARC in the multi-relation setting using Wan2.2. We separately evaluate CSP, the individual stages of A2G, and their combination. CSP alone provides substantial gains over the baseline, improving T2V-C from 67.45 to 71.95 and VBench from 10.34 to 14.50. The A2G variants also improve T2V-C, with the complete A2G formulation reaching 72.68. Combining CSP and A2G in SPARC achieves the strongest overall performance, reaching 73.27 T2V-C and 14.96 VBench. These results indicate that the two components provide complementary benefits, with CSP strengthening relational-role representation and A2G improving the grounding of spatial relations during denoising.

SPARC adds only approximately 2.5% inference time compared with the base Wan2.2 model. To determine whether the observed gains arise simply from this additional computation, we additionally evaluate a runtime-matched baseline by increasing Wan2.2 from 40 to 41 denoising steps. The additional denoising step produces only a small change in performance, from 67.45 to 67.48 on T2V-C and from 10.34 to 11.19 on VBench, remaining substantially below SPARC at 73.27 and 14.96, respectively. This comparison indicates that the improvements are attributable to the proposed spatial-conditioning mechanisms rather than merely to additional inference computation.

6 Conclusion

We investigated why transformer-based text-to-video models remain unreliable in generating spatial relations directly from text. Our analysis revealed two complementary failure sources: asymmetric representation of implicit spatial roles and ambiguous relational grounding during early denoising. Based on these observations, we introduced SPARC, a training-free framework that improves spatial controllability directly from text prompts without requiring external spatial guidance, such as bounding boxes, masks, or trajectories. Complementary Spatial Token Patching and Anchor-to-Graph Denoising directly address the two failure modes identified in our analysis. Experiments across multiple T2V backbones and increasingly complex relational prompts demonstrate improved spatial controllability without additional training or explicit spatial inputs. A current limitation is that SPARC constructs a fixed scene graph from the input prompt and therefore does not explicitly model spatial relations that change over time. Extending the framework to dynamic relational graphs is an important direction for future work.

AI Assistance Disclosure

We used an LLM to assist execution of the research, including software environment setup, dependency resolution, debugging, and experiment-script preparation. We used generative AI to generate a multi-object prompts set to obtain consistent multi-relation prompts from the existing dataset. We used generative AI tools to improve the clarity and readability of the manuscript. All AI-assisted outputs were reviewed and validated by the authors, who take responsibility for the final content of this work.

References

Vladimir Arkhipkin, Vladimir Korviakov, Nikolai Gerasimenko, Denis Parkhomenko, Viacheslav Vasilev, Alexey Letunovskiy, Nikolai Vaulin, Maria Kovaleva, Ivan Kirillov, Lev Novitskiy, Denis Koposov, Nikita Kiselev, Alexander Varlamov, Dmitrii Mikhailov, Vladimir Polovnikov, Andrey Shutkin, Julia Agafonova, Ilya Vasiliev, Anastasiia Kargapoltseva, Anna Dmitrienko, Anastasia Maltseva, Anna Averchenkova, Olga Kim, Tatiana Nikulina, and Denis Dimitrov. Kandinsky 5.0: A family of foundation models for image and video generation, 2026. URL https://arxiv.org/abs/2511.14993.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320. IEEE, 2024.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023.

Eli Collins. Fuel your creativity with new generative media models and tools. Google Blog, May 2025. URL https://blog.google/innovation-and-ai/products/generative-media-models-io-2025/. Published May 20, 2025.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, volume 2025, pp. 44730–44745, 2025.

Weixi Feng, Chao Liu, Sifei Liu, William Yang Wang, Arash Vahdat, and Weili Nie. Blobgen-vid: Compositional text-to-video generation with blob video representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12989–12998, June 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

Weijie He, Mushui Liu, Yunlong Yu, Zhao Wang, and Chao Wu. Dyst-xl: Dynamic layout planning and content control for compositional text-to-video generation. *arXiv preprint arXiv:2504.15032*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024.

Weng Ian Chan, Hiroaki Santo, Yasuyuki Matsushita, and Fumio Okura. Instance-wise distribution control of text-to-image diffusion models. *Pattern Recognition*, 172:112614, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.112614. URL https://www.sciencedirect.com/science/article/pii/S0031320325012774.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Long Lian, Baifeng Shi, Adam Yala, Boyi Li, et al. Llm-grounded video diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 50207–50227, 2024.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In *International Conference on Learning Representations*, volume 2025, pp. 101103–101147, 2025.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, pp. 366–384, 2024.

Bingyan Liu, Chengyu Wang, Tongtong Su, Huan Ten, Jun Huang, Kailing Guo, and Kui Jia. Understanding attention mechanism in video diffusion models. *arXiv preprint arXiv:2504.12027*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models, 2024.

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*, pp. 8748–8763. PmLR, 2021.

Aimon Rahman, Jiang Liu, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Yusheng Su, Vishal M. Patel, Zicheng Liu, and Emad Barsoum. Movi: Training-free text-conditioned multi-object video generation. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1036–1040, 2026. doi: 10.1109/ICASSP55912.2026.11461763.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pp. 70–80, 2015.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416, 2025.

Zhengyang Sun, Yu Chen, Xin Zhou, Xiaofan Li, Xiwu Chen, Dingkang Liang, and Xiang Bai. When numbers speak: Aligning textual numerals and visual instances in text-to-video diffusion models. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2026.

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

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ddFN3lWpIr.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6602–6611. IEEE, 2019.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=HbTxc6U1fO.

Qianxun Xu, Chenxi Song, Yujun Cai, and Chi Zhang. Switchcraft: Training-free multi-event video generation with attention controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Guy Yariv, Yuval Kirstain, Amit Zohar, Shelly Sheynin, Yaniv Taigman, Yossi Adi, Sagie Benaim, and Adam Polyak. Through-the-mask: Mask-based motion trajectories for image-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18198–18208. IEEE, 2025.

Sapir Esther Yiflach, Yuval Atzmon, and Gal Chechik. Data-driven loss functions for inference-time optimization in text-to-image. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3525–3535. IEEE, 2026.

Gaoyang Zhang, Bingtao Fu, Qingnan Fan, Qi Zhang, Runxing Liu, Hong Gu, Huaqi Zhang, and Xinguo Liu. Compass: Enhancing spatial understanding in text-to-image diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15253–15265. IEEE, 2025a.

Wanyue Zhang, Lin Geng Foo, Thabo Beeler, Rishabh Dabral, and Christian Theobalt. Vhoi: Controllable video generation of human-object interactions from sparse trajectories via motion densification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4009–4021, 2026.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, XIAOPENG ZHANG, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5a79AqFr0c.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2063–2073. IEEE, 2025b.

Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6818–6828, June 2024.