STRUCTURE-ANCHORED CONDITIONAL RESPONSE STEERING FOR HUMAN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Structurally conditioned human video generators already follow trajectories strongly constrained by pose, identity, and reference appearance, making unrestricted classifier-free guidance (CFG) prone to perturbing otherwise useful structural information. We revisit guidance from a conditional-response perspective and distinguish target-specific semantic and structural responses through selective changes of the conditioning configuration while keeping the remaining context fixed. Based on these axis-wise responses, we propose a structure-anchored steering formulation in which the model's own local response geometry determines where intervention is applied, while a bounded intervention budget regulates semantic steering in direction and diffusion time before the structural response is recomposed with controlled output magnitude. On Wan-Animate, our method achieves the best numerical results among the evaluated guidance methods on CLIP, T2V, MPJPE, aggregate VBench2 Human Fidelity Score, and both VideoScore2 visual-quality and physical-consistency criteria. Cross-backbone evaluation on other DiT-based algorithms further examines the transfer of the proposed guidance principle. These results suggest that guidance for structured human video generation is better formulated as controlled conditional-response steering than as unrestricted transformation of a single conditional discrepancy.

1 INTRODUCTION

Controllable human video generation increasingly combines high-capacity video generators with structural conditions such as human pose, facial identity, and reference appearance to produce temporally coherent and controllable motion (Ho et al., 2022; Blattmann et al., 2023; Brooks et al., 2024; Zhang et al., 2025a; Wang et al., 2025; Cheng et al., 2025). These structural signals are useful precisely because they strongly constrain the sampling trajectory Xue et al. (2025); Seedance et al. (2026); Sun et al. (2026). Text, by contrast, supplies complementary semantics—for example, the interacted object, action context, or fine-grained attributes required by the target prompt. Guidance in this setting is therefore different from unconstrained text-to-video generation: it follows the pattern of additional semantic intervention acts on a trajectory that is already substantially specified by auxiliary structure and appearance.

Classifier-free guidance (CFG) strengthens conditional generation by extrapolating a conditional discrepancy (Ho & Salimans, 2022). Recent training-free methods improve this mechanism from different perspectives. CFG++ modifies the guided sampling update (Chung et al., 2024); APG decomposes and rescales the guidance update (Sadat et al., 2024); ADG regulates guidance in the angular domain (Jin et al., 2025b); and CFG-Ctrl formulates guidance through feedback control (Wang et al., 2026a). These methods substantially improve generic guidance, but mainly address how a conditional discrepancy should be regulated or integrated into sampling. For structurally conditioned human video generation, an additional question becomes critical: *which conditional response should be guided when the trajectory is already strongly anchored by human structure and appearance?*

We address this question through *axis-wise conditional responses*. Let T denote the target prompt, T_0 a generic positive semantic anchor, $S = (P, F)$ the structural conditions, and R the shared reference conditions. At the same sampling state, we compare the fully conditioned prediction $v(T, S, R)$ with two selectively modified configurations, $v(T_0, S, R)$ and $v(T, S_0, R)$. Their finite

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.

differences define target-specific semantic and structural responses while keeping the remaining context fixed. This construction changes the object on which guidance operates: rather than treating all conditioning information as a single discrepancy to amplify, it exposes how distinct conditioning axes influence an already structured trajectory.

Our central principle is to *regulate these responses according to the geometry they induce*. The local angular change caused by each condition contrast defines a response-induced allocation field, so intervention is concentrated at spatio-temporal locations where the corresponding conditioning axis produces a stronger directional change in the model prediction. Semantic intervention is then bounded in direction and diffusion time, after which the structural response is recomposed and the output magnitude is stabilized. These operations are not introduced as independent modules; together they instantiate a single bounded conditional-intervention principle, summarized in Fig. 1.

We evaluate the method primarily on Wan-Animate (Cheng et al., 2025), with cross-backbone evaluation on PoseAnything (Wang et al., 2026b) reported in the appendix. On Wan-Animate, Ours achieves the best numerical results among the evaluated guidance methods on CLIP, T2V, MPJPE, aggregate VBench2 Human Fidelity, and both VideoScore2 Visual Quality and Physical Common Sense Consistency. It also obtains the highest Human Clothes and Human Identity scores, while Human Anatomy remains the only reported Human Fidelity component on which it does not lead. Qualitative comparisons further show more stable preservation of activity-specific human–object configurations under challenging interactions.

Our contributions are threefold:

- We formulate guidance for structurally conditioned human video generation through **axis-wise conditional responses**, isolating target-specific semantic and structural effects by selectively changing one conditioning source while keeping the remaining context fixed.
- We introduce a **bounded response-steering formulation** in which the model's own conditional geometry induces token-wise intervention allocation, while semantic steering is controlled in direction and diffusion time before structural recomposition and magnitude stabilization.
- We provide **comprehensive empirical evaluation** on Wan-Animate, where the proposed method improves semantic alignment, pose adherence, aggregate human fidelity, visual quality, and physical consistency over the evaluated guidance baselines, together with cross-backbone and sensitivity analyses in the appendix.

2 A Conditional-Response View of Guidance

Classifier-free guidance can be viewed as modifying the model prediction through the difference between two conditioning configurations. This perspective is particularly useful for structurally conditioned human video generation, where text, pose, face, and reference appearance contribute different types of information to the same sampling trajectory. We first introduce a common conditional-response formulation for existing guidance methods, and then instantiate it along the semantic and structural conditioning axes used by our method.

2.1 Generalized Conditional-Response Guidance

Let x_t denote the current state of a diffusion or flow-based sampler, and let $v_\theta(x_t, t; c)$ denote the model prediction under conditioning configuration c. Given a target configuration c^+ and a reference configuration c^-, we define their conditional response as

$$r_t(c^+, c^-) = v_\theta(x_t, t; c^+) - v_\theta(x_t, t; c^-). \tag{1}$$

A broad class of training-free guidance methods can then be described through two operations. First, a response-regulation operator transforms the predictions and their conditional response:

$$\tilde{v}_t = \Phi_t\left(v_t^+, v_t^-, r_{0:t}; \eta\right), \tag{2}$$

where $v_t^+ = v_\theta(x_t, t; c^+)$, $v_t^- = v_\theta(x_t, t; c^-)$, and η denotes method-specific guidance parameters. The resulting prediction is then incorporated into the sampler through

$$x_{t'} = \Psi_t\left(x_t, \tilde{v}_t; t, t'\right), \qquad t' < t, \tag{3}$$

where Ψ_t denotes the sampling update.

Equations 1–3 are not intended as a new guidance algorithm. Rather, they provide a common notation for separating three design choices: *which conditional contrast is constructed*, *how the resulting response is regulated*, and *how the guided prediction is integrated into sampling*. Figure 1(b) illustrates representative methods under this view.

Vanilla CFG (Ho & Salimans, 2022) linearly extrapolates a single conditional response. APG (Sadat et al., 2024) further regulates this response through projection, rescaling, and momentum, while ADG (Jin et al., 2025b) operates on its angular geometry. CFG-Ctrl (Wang et al., 2026a) introduces feedback-based regulation over the guidance process. These methods differ substantially in their regulation mechanisms, but primarily operate on a single conditional contrast. CFG++ (Chung et al., 2024) additionally modifies how the guided prediction is incorporated into the sampling dynamics. This distinction motivates us to examine not only *how* a response is regulated, but also *which conditioning response should be guided* in a structurally controlled human-video generator.

2.2 Structure-Anchored Axis-Wise Responses

We consider a human-video generator conditioned on a target prompt T, structural conditions $S = (P, F)$ consisting of pose and face guidance, and shared reference conditions R. In addition to the target prompt, we introduce a generic positive semantic anchor

$$T_0 = \text{“a person is performing an action”}, \tag{4}$$

which preserves generic human-action semantics without specifying the target-specific action content.

At the same state x_t, we evaluate three conditioning configurations:

$$v_J = v_\theta(x_t, t; T, S, R), \tag{5}$$

$$v_B = v_\theta(x_t, t; T_0, S, R), \tag{6}$$

$$v_{\bar{S}} = v_\theta(x_t, t; T, S_0, R), \tag{7}$$

where S_0 denotes the structure-ablated condition used by the implementation. The first prediction contains the complete target conditioning, the second replaces only the target text with the generic positive anchor while retaining the same structural and reference context, and the third removes the structural condition while preserving the target text and reference context.

These contrasts define two axis-wise conditional responses:

$$r_t^T = v_J - v_B, \tag{8}$$

$$r_t^S = v_J - v_{\bar{S}}. \tag{9}$$

The semantic response r_t^T isolates the target-specific change induced by replacing the generic human-action semantics T_0 with the target prompt T, while keeping S and R fixed. The structural response r_t^S measures the change induced by the pose and face conditions under the same target prompt and reference context. We refer to (r_t^T, r_t^S) as *axis-wise conditional responses*. Here, “axis-wise” refers only to selective intervention on different conditioning sources; it does not assume that the two responses are orthogonal, statistically independent, or form a complete linear basis.

The semantic construction in Eq. 8 also provides a direct comparison with conventional CFG extrapolation. If the target-specific semantic response is scaled linearly,

$$v_\lambda = v_B + \lambda r_t^T = v_J + (\lambda - 1) r_t^T, \tag{10}$$

then for $\lambda \geq 1$,

$$\|v_\lambda - v_J\|_2 = (\lambda - 1)\|r_t^T\|_2. \tag{11}$$

Thus, increasing λ guarantees a larger intervention relative to the native target-conditioned prediction v_J, but does not specify where that intervention should act, how its direction should be regulated, or for how long strong guidance should remain active.

This motivates the formulation developed in Sec. 3. Rather than treating r_t^T as a globally scalable Euclidean residual, we use the local response geometry to determine where semantic and structural intervention is relevant, regulate semantic steering through a bounded spatio-temporal budget, and finally recompose the structural response.

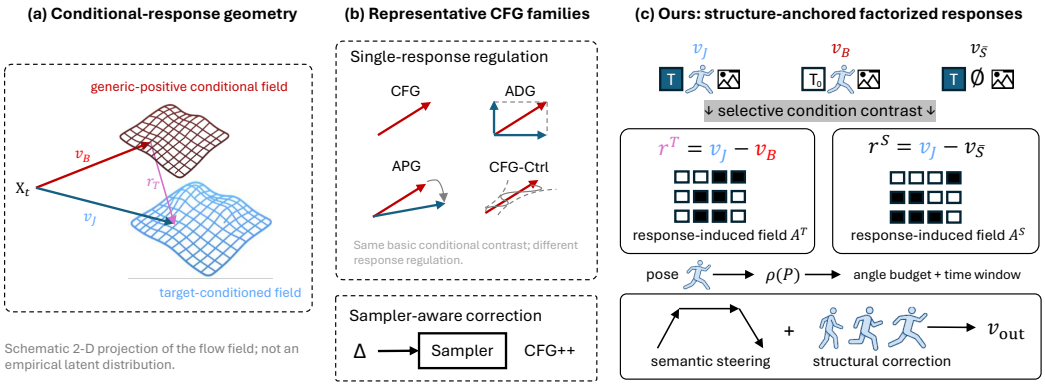


Figure 1: **Conditional-response view of guidance and our structure-anchored realization. (a)** At the same flow state x_t, replacing the generic positive semantic anchor T_0 with the target prompt T changes the predicted field from v_B to v_J, defining the target-specific semantic response $r_t^T = v_J - v_B$. **(b)** Representative CFG methods can be organized according to how a conditional response is regulated and incorporated into sampling. **(c)** Our formulation constructs axis-wise semantic and structural responses through selective condition contrasts.

3 Structure-Anchored Conditional Response Steering

Rather than applying a single global gain to the semantic response r_t^T, we regulate the intervention according to its local conditional geometry and the motion characteristics of the requested sequence. Starting from the axis-wise responses defined in Sec. 2.2, our method follows three steps. First, the response geometry itself determines where each conditioning axis is locally influential. Second, semantic intervention is performed in the angular domain under a shared sequence-level budget that regulates both directional deviation and temporal exposure. Finally, the structural response is recomposed with the steered semantic prediction and the magnitude of the generic structure-anchored prediction is restored. The complete procedure is illustrated in Fig. 1(c).

3.1 Response-Induced Local Allocation

Let q index a spatio-temporal latent location, and define the normalized prediction

$$\hat{v}(q) = \frac{v(q)}{\|v(q)\|_2 + \epsilon}. \tag{12}$$

For the semantic axis, the generic structure-anchored prediction v_B serves as the reference prediction; for the structural axis, the reference prediction is $v_{\bar{S}}$. We measure the local directional response of conditioning axis $k \in \{T, S\}$ as

$$\theta_t^k(q) = \arccos\left(\langle \hat{v}_J(q), \hat{v}_{\bar{k}}(q) \rangle\right), \tag{13}$$

where

$$v_{\bar{T}} = v_B, \qquad v_{\bar{S}} = v_\theta(x_t, t; T, S_0, R). \tag{14}$$

The angle $\theta_t^k(q)$ measures how strongly the local prediction direction changes when conditioning axis k is replaced or removed. We therefore derive an allocation field directly from this conditional response:

$$A_t^k(q) = \left[\frac{\theta_t^k(q) - \theta_{t,\min}^k}{\theta_{t,\max}^k - \theta_{t,\min}^k + \epsilon} \right]^\gamma, \qquad k \in \{T, S\}, \tag{15}$$

where $\theta_{t,\min}^k$ and $\theta_{t,\max}^k$ are computed over the spatio-temporal latent locations of the current prediction. We use $\gamma = 3$ in our implementation.

Importantly, A_t^k is not an externally specified foreground or segmentation mask. Instead, it is induced by the model's own conditional geometry: locations whose predicted field direction changes

more strongly under the corresponding condition contrast receive larger intervention weights. This yields separate semantic and structural allocation fields A_t^T and A_t^S without requiring an additional localization model.

3.2 Motion-Budgeted Semantic Steering

We next regulate how strongly, and for how long, the target-specific semantic response is amplified. Let λ_T denote the global semantic guidance scale and let $\tau_t \in [0,1]$ denote normalized sampling progress.

Motion-conditioned intervention budget. Let $m(P)$ denote a motion-complexity statistic computed from the input pose sequence P. We map it to a normalized intervention budget

$$\rho(P) = f(m(P)), \qquad 0 < \rho_{\min} \le \rho(P) \le 1, \tag{16}$$

where f is monotonically non-increasing with respect to motion complexity. Thus, more complex motion is assigned a more conservative intervention budget. The exact motion statistic and mapping function used in the final implementation are reported in Appendix B.

The same budget $\rho(P)$ controls two complementary aspects of semantic intervention: its temporal exposure and its allowable directional deviation.

Temporal exposure. Let $h(\cdot)$ be a canonical temporal guidance profile centered at c with base half-width w_0. We define its motion-conditioned realization as

$$s_\rho(\tau_t) = h\left(\frac{\tau_t - c}{\rho(P) w_0}\right), \tag{17}$$

where $h(u) = 0$ outside its support. Decreasing $\rho(P)$ therefore contracts the interval during which strong semantic guidance remains active.

When the scaled support remains within the sampling interval, a change of variables gives

$$\int_0^1 s_\rho(\tau)\, d\tau = \rho(P) \int_0^1 s_1(\tau)\, d\tau, \tag{18}$$

so that $\rho(P)$ directly controls the cumulative temporal exposure of semantic guidance.

Combining temporal exposure with the response-induced allocation field gives the local semantic gain

$$g_t^T(q) = 1 + (\lambda_T - 1) s_\rho(\tau_t) A_t^T(q). \tag{19}$$

Since $s_\rho(\tau_t) \in [0,1]$ and $A_t^T(q) \in [0,1]$,

$$1 \le g_t^T(q) \le \lambda_T. \tag{20}$$

Directional budget. The native target-conditioned prediction forms an anchor-relative angle $\theta_t^T(q)$ with $v_B(q)$. Rather than allowing unrestricted angular extrapolation, we allocate a motion-dependent fraction of the remaining safe angular range. Let $\theta_{\text{safe}} < \pi/2$ denote the non-opposing angular boundary used by the method, and define the available additional angular headroom as

$$\Delta\theta_{t,\max}(q;P) = \rho(P) \left[\theta_{\text{safe}} - \theta_t^T(q)\right]_+, \tag{21}$$

where $[a]_+ = \max(a, 0)$.

The amplified semantic response requests an angular increment

$$\Delta\theta_t^{\text{req}}(q) = \left(g_t^T(q) - 1\right) \theta_t^T(q). \tag{22}$$

We therefore use the budgeted angle

$$\hat{\theta}_t(q) = \theta_t^T(q) + \min\left\{\Delta\theta_t^{\text{req}}(q), \Delta\theta_{t,\max}(q;P)\right\}. \tag{23}$$

For locations satisfying $\theta_t^T(q) < \theta_{\text{safe}}$, Eq. 23 guarantees

$$\hat{\theta}_t(q) \le \theta_{\text{safe}} < \frac{\pi}{2}, \tag{24}$$

while locations already outside the safe angular region receive no additional angular amplification.

We then construct the unit direction orthogonal to the generic semantic anchor:

$$u_t^{\perp}(q) = \frac{\hat{v}_J(q) - \cos\theta_t^T(q)\,\hat{v}_B(q)}{\left\|\hat{v}_J(q) - \cos\theta_t^T(q)\,\hat{v}_B(q)\right\|_2 + \epsilon}. \tag{25}$$

The steered semantic prediction is reconstructed on the anchor-relative angular plane as

$$v_t^T(q) = \|v_B(q)\|_2 \left[\cos\hat{\theta}_t(q)\,\hat{v}_B(q) + \sin\hat{\theta}_t(q)\,u_t^{\perp}(q)\right]. \tag{26}$$

For locations inside the safe region, Eq. 24 implies

$$\left\langle \widehat{v_t^T}(q), \hat{v}_B(q) \right\rangle = \cos\hat{\theta}_t(q) > 0, \tag{27}$$

i.e., the steered semantic prediction remains non-opposing to the generic structure-anchored prediction. This is a directional property only; it does not by itself constitute a guarantee of structural correctness.

3.3 Structural Recomposition and Magnitude Consistency

Semantic steering modifies the target-specific response while the structural conditions must remain explicitly represented in the final prediction. We therefore recompose the steered semantic prediction with the independently estimated structural response $r_t^S = v_J - v_{\bar{S}}$.

Using the structural allocation field from Eq. 15, we form

$$z_t(q) = v_t^T(q) + \lambda_S A_t^S(q) r_t^S(q), \tag{28}$$

where λ_S controls the global strength of the structural response.

The recomposition in Eq. 28 can alter the local magnitude of the predicted field. To separate directional steering from uncontrolled magnitude growth, we restore the magnitude of the generic structure-anchored prediction:

$$v_t^{\text{out}}(q) = \frac{z_t(q)}{\|z_t(q)\|_2 + \epsilon}\|v_B(q)\|_2. \tag{29}$$

Ignoring the numerical stabilizer ϵ, this yields the exact magnitude property

$$\|v_t^{\text{out}}(q)\|_2 = \|v_B(q)\|_2. \tag{30}$$

Thus the final operation changes the direction of the recomposed prediction without introducing an additional token-wise magnitude gain.

Outside the active semantic-guidance interval, we directly use the original joint prediction v_J. The method therefore modifies only the guidance rule and leaves the underlying backbone, model parameters, and native sampler unchanged.

4 Experiments

Implementation details. Wan-Animate (Cheng et al., 2025) is our primary backbone. We additionally evaluate PoseAnything (Wang et al., 2026b) as a second backbone to test whether the guidance principle transfers beyond one architecture, which is reported in the appendix. We compare with Vanilla CFG, LTX, ADG, CFG++, CFG-Ctrl and APG. Guidance scales and method-specific hyperparameters are selected according to a fixed protocol determined before the final comparison. We filter 1,000 clips from Motion-X Zhang et al. (2025b) as the test set. The complete settings are reported in Appendix B.4.

Evaluation protocol. Semantic alignment is measured by CLIP Score and T2V Score. MPJPE measures adherence to the target pose sequence. VBench2 Human Fidelity is reported through all three constituent dimensions, including Human Clothes, Human Identity, and Human Anatomy, so that the anatomy weakness is not hidden behind an aggregate score. VideoScore2 Visual Quality and Physical Common Sense Consistency evaluate whether stronger conditional intervention degrades perceptual quality or physical plausibility. Additional explanation on the choice of the metrics are reported and analyzed in the appendix.

Table 1: **Quantitative comparison on Wan-Animate.** Cloth., Ident., Anat., and HF denote Human Clothes, Human Identity, Human Anatomy, and the aggregate Human Fidelity score from VBench2. VQ and Phys. denote Visual Quality and Physical Common Sense Consistency from VideoScore2. Best results are shown in **bold** and second-best results are underlined.

Method	Semantic		Structure	VBench2 Human Fidelity				VideoScore2	
	CLIP ↑	T2V ↑	MPJPE ↓	Cloth. ↑	Ident. ↑	Anat. ↑	HF ↑	VQ ↑	Phys. ↑
Vanilla	28.9235	0.7050	195.621	0.9852	0.8760	0.8484	0.9032	3.3349	3.0608
LTX	28.1572	0.7077	188.692	0.9733	0.8345	0.8301	0.8793	3.1791	2.9652
ADG	28.6184	0.7106	207.813	0.9707	0.8144	**0.8902**	0.8918	3.1522	2.8813
CFG++	28.9402	0.7108	201.246	0.9851	0.8705	0.8531	0.9029	3.3066	3.0308
CFG-Ctrl	28.9650	0.7139	169.195	0.9832	0.8850	0.7982	0.8888	3.3472	3.0128
APG	29.0128	0.7082	193.053	0.9927	0.8839	0.8519	0.9095	3.3489	2.9985
Ours	**29.1972**	**0.7142**	**159.324**	**0.9964**	**0.9156**	0.8473	**0.9198**	**3.5056**	**3.1631**

4.1 Quantitative Results on Wan-Animate

Tab. 1 evaluates whether stronger target-specific semantic intervention can be achieved while retaining the structural and perceptual constraints already established by the human-video generator. Across the evaluated guidance methods, Ours obtains the highest CLIP score (29.1972) and T2V score (0.7142), indicating the strongest numerical semantic alignment under both metrics. Importantly, this semantic improvement is not accompanied by weaker pose adherence. Ours also achieves the lowest MPJPE (159.324), compared with 169.195 for the next-best method, CFG-Ctrl. This result is particularly relevant to our formulation: regulating the target-specific conditional response does not require trading semantic alignment against the input pose trajectory. The same pattern extends beyond semantic and pose metrics. On VBench2 Human Fidelity, Ours achieves the highest Human Clothes (0.9964) and Human Identity (0.9156) scores, resulting in the highest aggregate Human Fidelity score of 0.9198. Human Anatomy is the only reported dimension on which Ours does not obtain the best result: ADG achieves 0.8902, whereas Ours obtains 0.8473. We report this component explicitly rather than relying only on the aggregate Human Fidelity score. Ours also obtains the highest VideoScore2 Visual Quality (3.5056) and Physical Common Sense Consistency (3.1631). These results suggest that the improvement in semantic conditioning is not limited to text–video similarity, but co-occurs with stronger pose adherence, identity and clothing fidelity, visual quality, and physical plausibility. Overall, the results support the central motivation of our method: for a trajectory already constrained by human structure and reference appearance, target-specific semantics can be strengthened more effectively by regulating the conditional response than by globally amplifying a single guidance discrepancy.

4.2 Qualitative Results on Wan-Animate

Figure 2 presents qualitative comparisons on three representative human–object interaction sequences with substantially different motion patterns and scene dynamics. The examples cover luge, calf roping, and white-water kayaking, allowing us to examine not only visual quality, but also whether guidance preserves the activity-specific relationship between the human, the interacted object, and the surrounding scene.

Across the three cases, Ours more consistently preserves the structural configuration specified by the source motion while retaining the target-specific interaction semantics. In the luge example, the generated subject remains stably seated on the sled and follows the characteristic reclined configuration of the reference motion. Several alternative guidance strategies introduce larger changes in body configuration, interacted-object appearance, or scene content. In the calf-roping example, Ours maintains a coherent spatial relationship among the rider, horse, and calf across the displayed frames, whereas some baselines exhibit stronger appearance changes or less stable interaction geometry. These observations are consistent with the lower MPJPE and higher semantic-alignment scores reported in Table 1.

The white-water kayaking example further illustrates the importance of controlling the intervention in visually difficult, high-motion regions. As highlighted by the enlarged regions, several competing

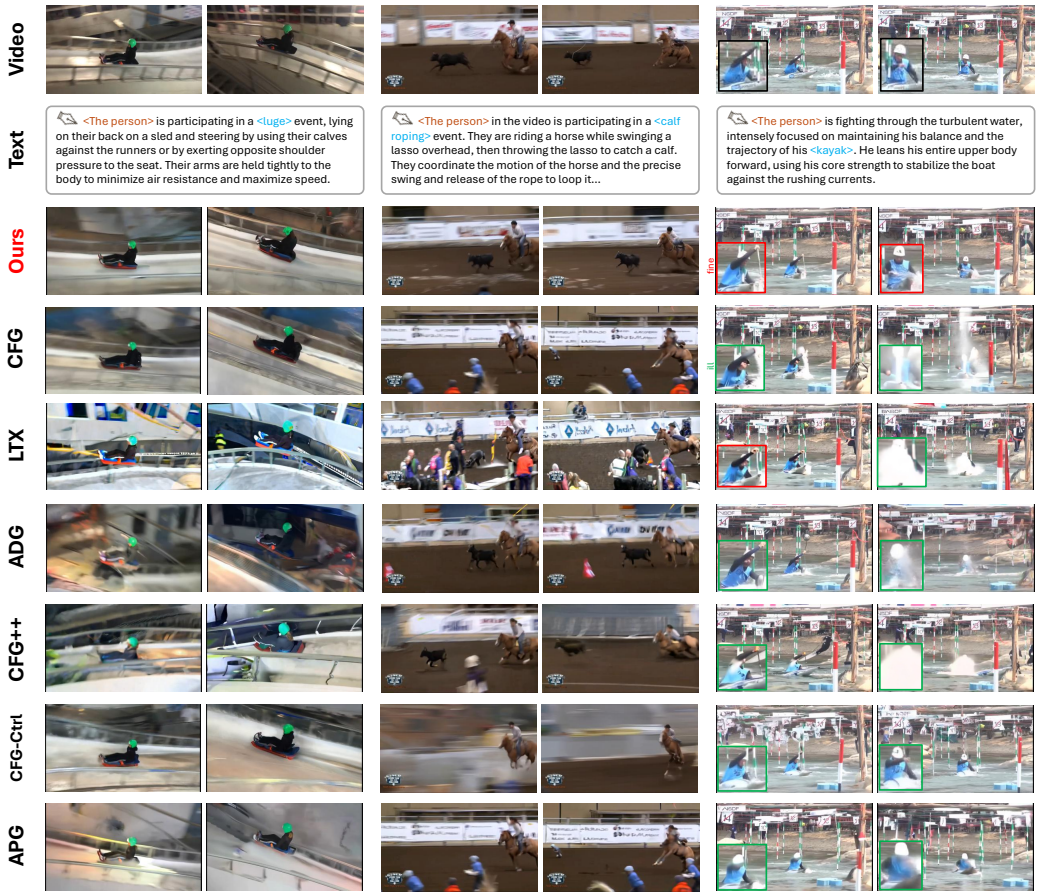


Figure 2: **Qualitative comparison on Wan-Animate.** We compare Ours with CFG, LTX, ADG, CFG++, CFG-Ctrl, and APG on representative luge, calf-roping, and kayaking sequences.

guidance methods exhibit noticeable subject deformation, disappearance, or locally over-amplified artifacts across frames. Ours retains a more identifiable human–kayak configuration while preserving the surrounding turbulent-water scene. These examples qualitatively support the central motivation of our formulation: target-specific semantic guidance is most useful when it strengthens the intended interaction without unnecessarily perturbing the human structure and visual context.

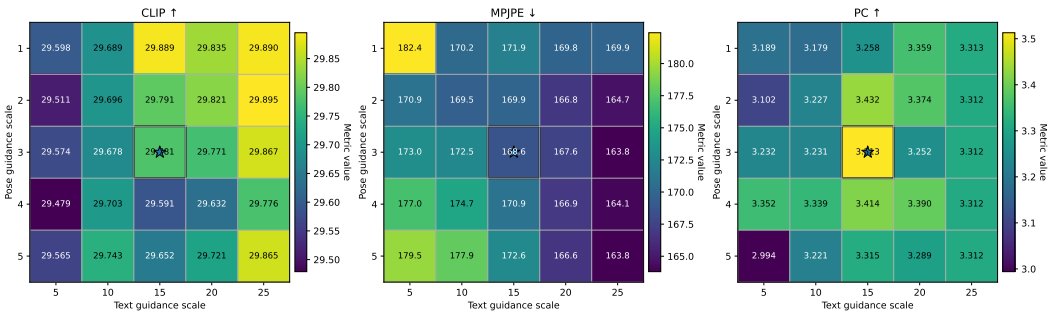


Figure 3: **Guidance-scale sensitivity on Wan-Animate.** We sweep the pose guidance scale $\lambda_{\text{pose}} \in \{1, 2, 3, 4, 5\}$ and the text guidance scale $\lambda_{\text{text}} \in \{5, 10, 15, 20, 25\}$. The star marks the default operating point $(\lambda_{\text{pose}}, \lambda_{\text{text}}) = (3, 15)$ used in the main experiments.

4.3 GUIDANCE-SCALE SENSITIVITY

We randomly choose another 1,000 clips from Motion-X as validation set, and analyze sensitivity to the pose and text guidance scales using CLIP, MPJPE, and VideoScore2 Physical Common Sense Consistency (PC), as shown in Fig. 3. Increasing λ_{text} generally improves CLIP and reduces MPJPE, while PC peaks at an intermediate text scale and then declines. The pose scale also shows a non-monotonic effect, indicating that guidance-scale selection is inherently multi-objective.

We therefore use $(\lambda_{\text{pose}}, \lambda_{\text{text}}) = (3, 15)$ as the default operating point. It achieves the highest PC score (3.513), while remaining close to the best CLIP (29.781 vs. 29.895) and MPJPE (168.6 vs. 163.8). This setting provides a favorable balance between semantic alignment, pose adherence, and physical plausibility.

Table 2: **Comparison of temporal guidance policies on Wan-Animate.** We select representative metrics covering semantic alignment, pose adherence, and VBench2 Anat., Cloth., Ident., and HF. Best results are shown in **bold** and second-best results are underlined.

Policy	Semantic	Structure	VBench2 Human Fidelity			
	CLIP ↑	MPJPE ↓	Anat. ↑	Cloth. ↑	Ident. ↑	HF ↑
Uniform	24.66	169.14	0.7862	0.9871	0.8632	0.8788
Step Function	26.28	264.23	0.8146	0.9866	0.8698	0.8903
Fixed Window	25.02	165.73	0.8366	0.9791	0.8774	0.8977
TV-CFG	25.73	200.70	**0.8590**	0.9839	0.8871	0.9100
Ours	**29.20**	**159.32**	0.8473	**0.9964**	**0.9156**	**0.9198**

4.4 TEMPORAL ALLOCATION OF SEMANTIC GUIDANCE

The guidance scale determines the overall strength of semantic intervention, but not how this intervention should be distributed along the generation trajectory. We therefore fix the global text guidance scale at $\lambda_{\text{text}} = 15$ and compare Uniform, Step Function (Kynkäänniemi et al., 2024), Fixed Window, TV-CFG (Jin et al., 2025a), and our temporal allocation policy under the same backbone and test subset. To keep the analysis compact, we select representative metrics covering semantic alignment, pose adherence, and human fidelity, while reporting all three VBench2 Human Fidelity components together with their aggregate score.

As shown in Table 2, temporal allocation substantially affects generation quality even when the global guidance strength is fixed. Ours achieves the highest CLIP score (29.20) and the lowest MPJPE (159.32), together with the best Human Clothes (0.9964), Human Identity (0.9156), and aggregate Human Fidelity (0.9198). Human Anatomy is the only reported dimension on which Ours does not lead, where TV-CFG obtains 0.8590 compared with 0.8473 for Ours. These results indicate that temporal allocation is not merely a scheduling detail: different distributions of the same global guidance strength lead to distinct trade-offs in semantic alignment, pose adherence, and human fidelity. Together with the guidance-scale analysis, this motivates controlling both the magnitude of semantic intervention and its exposure along the generation trajectory.

5 CONCLUSION

We study guidance for human-video generators whose trajectories are already constrained by pose, face, and reference appearance. Instead of amplifying a single conditional discrepancy, we formulate semantic and structural effects as axis-wise conditional responses and regulate them through a bounded steering process that controls local intervention, directional deviation, temporal exposure, and output magnitude. On Wan-Animate, our method achieves the best numerical results among the evaluated guidance methods on semantic alignment, pose adherence, aggregate Human Fidelity, visual quality, and physical consistency, while Human Anatomy remains the main dimension on which it does not lead. Cross-backbone experiments further evaluate transfer beyond the primary architecture. These results support viewing guidance in structurally conditioned generation as controlled conditional intervention rather than globally scaled residual amplification.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, manuscript organization, and research-assistance tasks. All AI-assisted text, claims, code, and experimental suggestions were reviewed by the authors. The authors take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the mathematical formulation and principal implementation choices in the main paper, with complete guidance hyperparameters, baseline reimplementation settings, preprocessing details, additional ablations, and evaluation protocols in the appendix. Code and configuration files will be released anonymously as supplementary material when permitted by the submission process.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3(1):3, 2024.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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.

Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng ... Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. In *ICCV*, 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*, 2024.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. *arXiv preprint arXiv:2509.22007*, 2025a.

Cheng Jin, Zhenyu Xiao, Chutao Liu, and Yuantao Gu. Angle domain guidance: Latent diffusion requires rotation rather than extrapolation. *arXiv preprint arXiv:2506.11039*, 2025b.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

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*, pp. 366–384. Springer, 2024.

OpenAI. Gpt-4v(ision) system card, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Tianyu Sun, Zhoujie Fu, Zihui Gao, Bang Zhang, and Guosheng Lin. Rgb-d video generation for improving human-to-robot object handover prediction. *IEEE Robotics and Automation Letters*, 2026.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. Cfg-ctrl: Control-based classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11437–11447, June 2026a.

Ruiyan Wang, Teng Hu, Kaihui Huang, Zihan Su, Ran Yi, and Lizhuang Ma. Poseanything: General pose-guided video generation with part-aware temporal coherence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23193–23203, 2026b.

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 68(10):200103, 2025.

Ziyi Xu, Ziyao Huang, Juan Cao, Yong Zhang, Xiaodong Cun, Qing Shuai, Yuchen Wang, Linchao Bao, and Fan Tang. Anchorcrafter: Animate cyber-anchors selling your products via human-object interacting video generation. *IEEE Transactions on Visualization and Computer Graphics*, 2026.

Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, et al. Human motion video generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, pp. 74896–74910. PMLR, 2025a.

Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. *arXiv preprint arXiv:2501.05098*, 2025b.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Overview of the Appendix

In the appendix, we provide additional theoretical derivations, extensive ablation studies, and qualitative evaluations to further validate our method and broaden the scope of our work. The appendix is organized to follow the logical progression of our methodological interventions:

- **Appendix A:** Provides necessary proofs and mathematical derivations for the formation of our proposed response-steering.
- **Appendix B:** Presents all necessary explanation on the implementation details of our experiments.
- **Appendix C:** The results of cross-backbone evaluation on PoseAnything Wang et al. (2026b), another DiT-based pipeline for human video generation.
- **Appendix D:** Details the supplemental qualitative ablation studies to support our method.
- **Appendix E:** Showcases the robustness and scalability of our method under extreme guidance scales ($\lambda \in \{20, 30\}$), accompanied by quantitative sensitivity curves.
- **Appendix F:** Concludes with extended discussions on all other necessary supplemental discussions.

A Boundedness Properties of the Intervention

This section establishes several elementary but useful properties of the response-steering formulation in Sec. 3. The purpose is not to claim global convergence or structural correctness, but to make explicit how the proposed intervention is bounded in guidance strength, angular deviation, temporal exposure, and output magnitude.

Proposition 1 (Bounded structure-anchored intervention)**.** *Assume that*

$$0 \le A_t^T(q) \le 1, \qquad 0 \le s_\rho(\tau_t) \le 1, \qquad 0 < \rho(P) \le 1, \tag{31}$$

with semantic guidance scale $\lambda_T \ge 1$, and let $\theta_{\text{safe}} < \pi/2$. Then the semantic steering rule in Sec. 3.2 satisfies the following properties.

(i) ***Bounded local semantic gain.*** *The token-wise semantic gain*

$$g_t^T(q) = 1 + (\lambda_T - 1) s_\rho(\tau_t) A_t^T(q) \tag{32}$$

satisfies

$$1 \le g_t^T(q) \le \lambda_T. \tag{33}$$

(ii) ***Bounded directional deviation.*** *For any location satisfying $\theta_t^T(q) < \theta_{\text{safe}}$, the budgeted angle*

$$\hat{\theta}_t(q) = \theta_t^T(q) + \min\left\{\Delta\theta_t^{\text{req}}(q), \rho(P)\left[\theta_{\text{safe}} - \theta_t^T(q)\right]\right\} \tag{34}$$

obeys

$$\theta_t^T(q) \le \hat{\theta}_t(q) \le \theta_{\text{safe}} < \frac{\pi}{2}. \tag{35}$$

Consequently, the reconstructed semantic direction remains non-opposing to the generic structure-anchored prediction:

$$\left\langle \widehat{v_t^T}(q), \hat{v}_B(q) \right\rangle = \cos\hat{\theta}_t(q) > 0. \tag{36}$$

(iii) ***Linear control of temporal exposure.*** *Let the motion-conditioned temporal profile be*

$$s_\rho(\tau) = h\left(\frac{\tau - c}{\rho(P) w_0}\right), \tag{37}$$

where $h(u) = 0$ outside its compact support. If the scaled support remains within $[0, 1]$, then

$$\int_0^1 s_\rho(\tau)\, d\tau = \rho(P) \int_0^1 s_1(\tau)\, d\tau. \tag{38}$$

(iv) ***No output-magnitude amplification.*** *Let the final prediction be*

$$v_t^{\mathrm{out}}(q) = \frac{z_t(q)}{\|z_t(q)\|_2 + \epsilon}\|v_B(q)\|_2, \qquad \epsilon \geq 0. \tag{39}$$

Then

$$\|v_t^{\mathrm{out}}(q)\|_2 = \frac{\|z_t(q)\|_2}{\|z_t(q)\|_2 + \epsilon}\|v_B(q)\|_2 \leq \|v_B(q)\|_2. \tag{40}$$

In the idealized case $\epsilon = 0$ *and* $z_t(q) \neq 0$*,*

$$\|v_t^{\mathrm{out}}(q)\|_2 = \|v_B(q)\|_2. \tag{41}$$

Proof. **(i) Bounded local semantic gain.** Since

$$0 \leq s_\rho(\tau_t)A_t^T(q) \leq 1 \tag{42}$$

and $\lambda_T - 1 \geq 0$, multiplying the inequality by $\lambda_T - 1$ gives

$$0 \leq (\lambda_T - 1)s_\rho(\tau_t)A_t^T(q) \leq \lambda_T - 1. \tag{43}$$

Adding 1 to each term immediately yields Eq. 33.

The assumption $A_t^T(q) \in [0, 1]$ follows directly from its construction. Specifically, the min–max normalized angular response lies in $[0, 1]$, and raising it to any positive power $\gamma > 0$ preserves this range. If all local angles are identical, the numerator in Eq. 15 vanishes and the allocation field is zero.

(ii) Bounded directional deviation. For a location satisfying $\theta_t^T(q) < \theta_{\mathrm{safe}}$, define

$$\Delta_t = \theta_{\mathrm{safe}} - \theta_t^T(q) > 0. \tag{44}$$

Because $0 < \rho(P) \leq 1$,

$$0 < \rho(P)\Delta_t \leq \Delta_t. \tag{45}$$

Moreover, $\Delta\theta_t^{\mathrm{req}}(q) \geq 0$ because $g_t^T(q) \geq 1$. Therefore,

$$0 \leq \min\{\Delta\theta_t^{\mathrm{req}}(q), \rho(P)\Delta_t\} \leq \Delta_t. \tag{46}$$

Adding $\theta_t^T(q)$ gives

$$\theta_t^T(q) \leq \hat{\theta}_t(q) \leq \theta_t^T(q) + \Delta_t = \theta_{\mathrm{safe}}, \tag{47}$$

which proves Eq. 35.

By construction, $u_t^\perp(q)$ is orthogonal to $\hat{v}_B(q)$. Ignoring the numerical stabilizer in the normalization of $u_t^\perp$, the reconstructed semantic direction is therefore

$$\widehat{v_t^T}(q) = \cos\hat{\theta}_t(q)\hat{v}_B(q) + \sin\hat{\theta}_t(q)u_t^\perp(q). \tag{48}$$

Taking its inner product with $\hat{v}_B(q)$ yields

$$\left\langle \widehat{v_t^T}(q), \hat{v}_B(q) \right\rangle = \cos\hat{\theta}_t(q). \tag{49}$$

Since $\hat{\theta}_t(q) \leq \theta_{\mathrm{safe}} < \pi/2$, we have $\cos\hat{\theta}_t(q) > 0$, which proves Eq. 36. For the degenerate case $\theta_t^T(q) = 0$, the same result follows by continuity.

(iii) Linear control of temporal exposure. Using the change of variables

$$u = \frac{\tau - c}{\rho(P)w_0}, \qquad d\tau = \rho(P)w_0\, du, \tag{50}$$

and assuming that the scaled support does not intersect the boundaries of $[0, 1]$, we obtain

$$\int_0^1 s_\rho(\tau)\, d\tau = \rho(P)w_0 \int h(u)\, du. \tag{51}$$

For $\rho(P) = 1$,

$$\int_0^1 s_1(\tau)\, d\tau = w_0 \int h(u)\, du. \tag{52}$$

Combining the two expressions yields Eq. 38.

(iv) Output-magnitude control. Taking the Euclidean norm of the final prediction gives

$$\|v_t^{\text{out}}(q)\|_2 = \left\| \frac{z_t(q)}{\|z_t(q)\|_2 + \epsilon} \|v_B(q)\|_2 \right\|_2 \tag{53}$$

$$= \frac{\|z_t(q)\|_2}{\|z_t(q)\|_2 + \epsilon} \|v_B(q)\|_2. \tag{54}$$

Since $\epsilon \geq 0$,

$$\frac{\|z_t(q)\|_2}{\|z_t(q)\|_2 + \epsilon} \leq 1, \tag{55}$$

which proves Eq. 40. If $\epsilon = 0$ and $z_t(q) \neq 0$, the multiplicative factor is exactly one, yielding Eq. 41. □

Analysis. Proposition 1 does not imply that the generated video is structurally correct or that the chosen intervention is optimal. Instead, it characterizes the quantities that the guidance rule controls explicitly. The local semantic amplification cannot exceed the global guidance scale, semantic steering is prevented from entering the opposing angular half-space for locations initially inside the safe region, the shared budget controls cumulative high-guidance exposure linearly, and the final recomposition cannot increase the magnitude beyond that of the generic structure-anchored prediction. These properties motivate viewing the method as a bounded conditional intervention rather than unrestricted CFG extrapolation.

B Implementation and baseline details

B.1 Dataset Settings

As we have mentioned in the main paper, we chose Motion-X as the experimental dataset, because of its diversity in scene and complexity in motion. We examined the motion type distribution of its subsets, and calculate the scene/motion type per subset. To make sure that our test set would cover most varieties, we select the most samples from the most diversified subset HAA500, while only keep a some proportion of IDEA400 which is the least diversified. We curate a test set comprising 1,000 video clips from the Motion-X dataset Lin et al. (2023). As detailed in Table 3, these clips are sampled across seven subsets. When we carry out the training and testing, we follow other previous works Zhang et al. (2025a); Wang et al. (2025); Xue et al. (2025) and take the first frame of each video clip as the reference image. If the first frame does not contain a person, then we re-select another frame that contains the person in that video. Furthermore, to maximize the evaluation breadth and prevent bias toward repetitive motions, we strictly enforce a constraint of sampling only one clip per specific action category.

Within each subset, we also follow the official guide and choose motions with different complexity. For example, we select the videos of sports, household behaviors, communications, leisure activities, etc. from HAA500. As for subsets such as KungFu and Music, we could only select videos of KungFu or instrument performances. We ensure that samples with various "subject composition" could be included, so that our method could be tested across different scenarios.

As for validation set that we use to make parameter screening, we follow the same criteria and sample another 1,000 video clips to compose the val set. Notably, we strictly prevent any intersection between the test set and val set. We also make sure that no same type of action would be included in both dataset, especially for HAA500 subset and IDEA400 subset. This is to prevent overfitting and potential unfair comparison.

Note on Annotation Quality: During the curation process, we observed that a portion of the original text annotations in the Motion-X dataset contains inaccuracies or semantic misalignments with the visual content. For future research utilizing this dataset, we highly recommend employing advanced Vision-Language Models (e.g., GPT-4V OpenAI (2023)) via Visual Question Answering (VQA) pipelines to conduct an initial filtering and refinement of the text prompts prior to experimental application.

Table 3: **Details of the Motion-X Test Set Split.** We report the total available video clips and the number of sampled clips for our evaluation test set across seven subsets.

Motion-X Subset	Total Clips	Sampled Clips
HAA500	6,944	500
Animation	522	50
HuMMan	971	50
KungFu	1,032	50
Music	3,394	100
Perform	923	50
IDEA400	12,025	200
Total	**25,811**	**1,000**

B.2 Metric Selection

Nowadays, various benchmark or metric strategies have been proposed, and we consider the following aspect: semantic alignment, pose structure, human fidelity, physics common sense. Since our method falls in the category of human video generation, the study of pose structure and human fidelity should be surely included. And considering that we aim to propose a method with better semantic representation, we choose the traditional CLIP score, and the newer T2V-metric Lin et al. (2024), which is deemed more alike to human perception.

Besides some out-dated metrics Unterthiner et al. (2018) that gave been mostly substituted, there are some other recent metrics that claim to have a more comprehensive measurement of the generated human video, but some of them are strictly limited within their concept space. For example, the VBench family Huang et al. (2024); Zheng et al. (2025) also has "human action" and "human interaction", but the former only recognize a few action type in its action pool, and the latter only works for human-human interaction. Therefore, we took the displayed metrics into consideration.

B.3 Wan-Animate Implementation

Our method is training-free and operates entirely at inference time on the frozen Wan-Animate backbone, without additional fine-tuning or parameter updates. At each active guidance step, we evaluate three prediction branches: the fully conditioned prediction $v_J = v_\theta(x_t, t; T, S, R)$, the generic-semantic prediction $v_B = v_\theta(x_t, t; T_0, S, R)$, and the structure-ablated prediction $v_{\tilde{S}} = v_\theta(x_t, t; T, S_0, R)$. We use T_0 = "a person is performing an action". The structural condition $S = (P, F)$ contains pose and face inputs; for S_0, pose latents are replaced by zeros and face inputs are set to -1, while the target text and reference conditions remain unchanged.

For both semantic and structural responses, the token-wise angular discrepancy is min–max normalized and sharpened with exponent $\gamma = 3$. We use a structural response scale of 3.0 and cap semantic steering at 1.17 radians. The temporal guidance profile linearly increases during the first 20% of sampling, remains at its maximum from 20% to 70%, and decreases linearly thereafter. The final prediction is normalized to the token-wise magnitude of v_B. We use Flow DPM-Solver++ with 20 inference steps, schedule shift 5.0, BF16 inference, and FP32 computation for the response geometry. All test and validation experiments are conducted on $8\times$ NVIDIA H200 GPUs.

B.4 Baseline Reimplementation Details

All guidance baselines are reimplemented on the same frozen Wan-Animate backbone and evaluated under identical sampling and conditioning settings. Unless otherwise specified, we set the guidance scale to 3, matching the default scale used by our method. For methods whose guidance parameters have a different numerical meaning from standard CFG, we follow their native parameterization rather than directly matching the scale value.

For ADG, we use a guidance scale of 4, following the reference setting provided by the official implementation. For CFG++, the guidance coefficient λ is defined on a substantially smaller numerical

range than the scale used by conventional CFG; we therefore follow its native parameter domain and set $\lambda = 0.6$, consistent with the official example setting.

CFG-Ctrl introduces additional control-specific parameters. Its Wan-Video configuration recommends `cfg_scale` $= 5$, `smc_cfg_lambda` $= 5.0$, `smc_cfg_k` $= 0.2$, zero warmup, and 50 sampling steps. In our Wan-Animate implementation, however, `cfg_scale` $= 5$ produces severe temporal flickering, while `cfg_scale` $= 2$ yields stable generation. We therefore use `cfg_scale` $= 2$ while retaining `smc_cfg_lambda` $= 5.0$, `smc_cfg_k` $= 0.2$, zero warmup, and 50 sampling steps. All reported results use these fixed settings across the evaluation set.

Table 4: **Guidance settings used for baseline reimplementations.** Unless otherwise specified, the guidance scale is set to 3, matching our default setting for pose track.

Method	Guidance setting
Vanilla CFG	scale $= 3.0$
APG	scale $= 3.0$
LTX	scale $= 3.0$
ADG	scale $= 4.0$
CFG++	$\lambda = 0.6$
CFG-Ctrl	cfg scale $= 2.0$, $\lambda = 5.0$, $K = 0.2$

C Cross-Backbone Evaluation on PoseAnything

We further evaluate our method on PoseAnything (Wang et al., 2026b) to examine whether the proposed guidance principle transfers beyond Wan-Animate to a structurally different human-video generator. This experiment is intended as a focused cross-backbone evaluation rather than a second exhaustive baseline study. Based on our previous Wan-Animate comparison, we select APG as the strongest representative competing guidance method and compare it with Ours and Vanilla CFG. All methods use the same PoseAnything backbone, inputs, preprocessing, and inference settings, without additional training or parameter updates.

C.1 Quantitative Results

Table 5 reports semantic alignment, pose adherence, and VBench2 Human Fidelity. Ours achieves the highest CLIP score (29.5681) and the lowest MPJPE (170.319), compared with 29.1512/197.403 for Vanilla CFG and 29.1826/182.586 for APG. The lower MPJPE indicates that the improvement in semantic alignment is accompanied by stronger adherence to the input pose sequence.

For Human Fidelity, all three methods obtain the same Clothes score of 1.0000. Ours achieves the highest Identity score (0.8771) and aggregate Human Fidelity score (0.9366), while APG obtains the highest Anatomy score (0.9466), followed by Ours at 0.9328. Overall, Ours performs best on CLIP, MPJPE, Identity, and aggregate Human Fidelity, providing additional evidence that the response-steering formulation transfers beyond the primary Wan-Animate backbone.

C.2 Qualitative Results

Figure 4 presents qualitative comparisons on three representative actions: basketball dribbling, high-five interaction, and one-arm push-up. These examples cover object-related motion, two-person interaction, and an asymmetric full-body pose, providing complementary tests of semantic and structural guidance.

Across the displayed frames, Ours more consistently preserves the action-specific configuration of the reference video while maintaining the target semantics. In the basketball example, Ours retains the dribbling configuration and subject appearance more faithfully. In the high-five example, Ours better preserves the relative configuration of the two participants and the intended hand interaction. The one-arm push-up example further tests structural adherence under a challenging asymmetric pose, where Ours maintains the characteristic one-arm support and body configuration more closely across the shown frames.

Table 5: **Cross-backbone quantitative evaluation on PoseAnything.** We compare Ours with Vanilla CFG and APG, which is selected as the strongest representative competing baseline from our previous Wan-Animate evaluation. Cloth., Ident., Anat., and HF denote Human Clothes, Human Identity, Human Anatomy, and aggregate Human Fidelity from VBench2. Best results are shown in **bold** and second-best results are underlined.

Method	CLIP ↑	MPJPE ↓	Cloth. ↑	Ident. ↑	Anat. ↑	HF ↑
Vanilla CFG	29.1512	197.403	**1.0000**	0.8612	0.9147	0.9253
APG	29.1826	182.586	**1.0000**	0.8126	**0.9466**	0.9197
Ours	**29.5681**	**170.319**	**1.0000**	**0.8771**	0.9328	**0.9366**

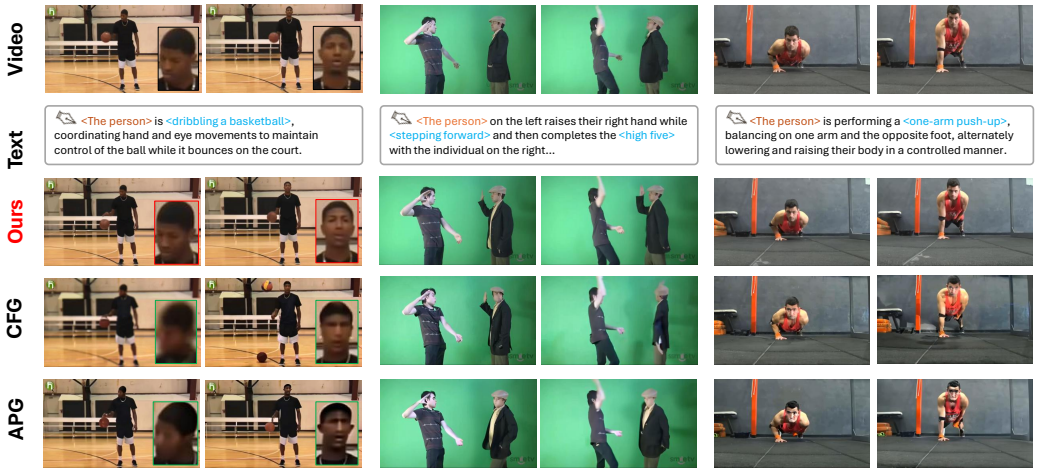


Figure 4: **Qualitative comparison on PoseAnything.** We compare Ours with Vanilla CFG and APG on basketball dribbling, high-five interaction, and one-arm push-up sequences. The top two rows show reference video frames and the corresponding text descriptions, followed by generations from Ours, Vanilla CFG, and APG. Ours more consistently preserves the action-specific human configuration while maintaining the requested semantics. Insets in the basketball example enlarge the face region for appearance comparison.

These qualitative observations are consistent with the quantitative results in Table 5: Ours achieves stronger semantic alignment and lower pose error than both Vanilla CFG and APG, while also obtaining the highest aggregate Human Fidelity. Together, the PoseAnything experiments support the transferability of the proposed training-free guidance formulation across different structurally conditioned human-video generators.

D Supplemental Ablation Studies

D.1 Quantitative Results for Semantic Steering

Table 6 summarizes the CLIP Score and MPJPE performance for semantic steering value. We implement the screening test on the validation set. This result gives us the support for previous claim, that the semantic steering too large or too small slows the convergence of the diffusion model.

D.2 Qualitative Analysis of Temporal Guidance Policies

To complement the quantitative comparison of temporal guidance policies in the main paper, we provide additional qualitative examples in Fig. 5. All methods use the same global text guidance scale $\lambda_{\text{text}} = 15$ and differ only in how semantic guidance is distributed along the sampling trajectory. We compare Uniform, Step Function, Fixed Window, TV-CFG, and Ours on two representative sequences with different motion and scene characteristics.

Table 6: **Impact of Semantic Steering.** Comparison of CLIP Scores at standard guidance scales.

Scale (λ)	Configuration	MPJPE	CLIP Score ↑
$\lambda = 15$	1.57	171.44	28.59
	1.37	163.71	28.62
	1.17	**159.32**	**29.20**
	1.07	161.77	29.03
	0.97	163.58	27.19

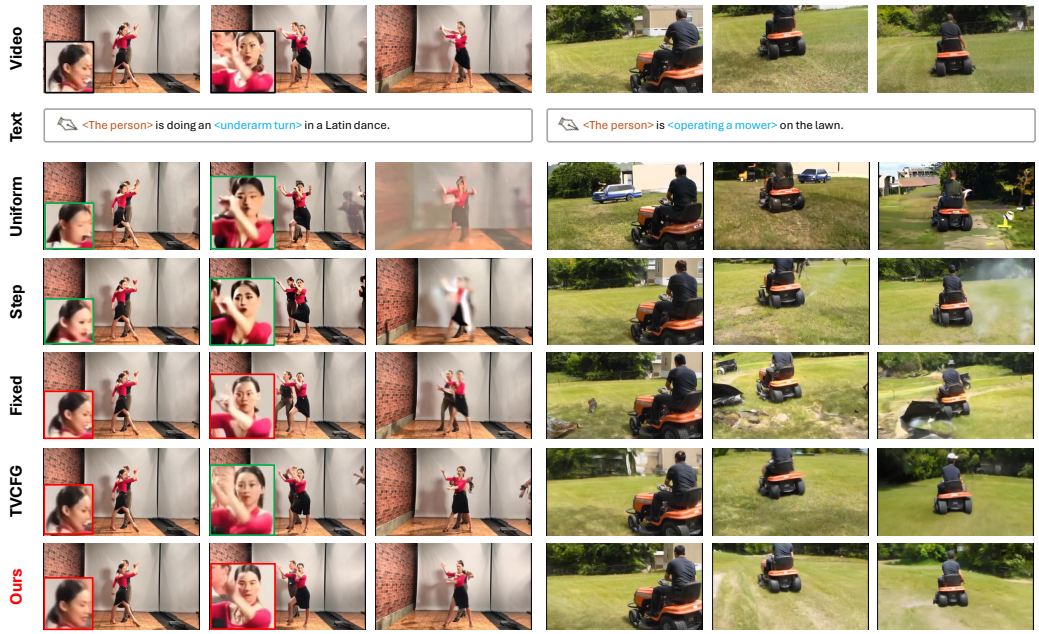


Figure 5: **Qualitative comparison of temporal guidance policies.** All methods use the same global text guidance scale $\lambda_{\text{text}} = 15$. On the Latin-dance sequence (left), different temporal policies produce noticeable differences in structural and facial behavior during rapid motion. On the lawnmower sequence (right), the policies differ in human–object configuration and scene consistency across the displayed frames. Ours more consistently preserves human structure, subject appearance, and interaction geometry under the same global guidance strength.

In the Latin-dance sequence, the rapid arm and body motion exposes clear differences in structural and facial stability. Uniform and Step Function exhibit stronger blur or appearance variation in several displayed frames, while Fixed Window and TV-CFG preserve the global action more effectively but still show local changes in facial appearance and body structure. Ours maintains a more consistent subject appearance and preserves the characteristic turning motion across the shown frames.

The lawn-mower sequence provides a complementary case in which the human–object configuration and background should remain stable while the subject moves through the scene. Several alternative temporal policies introduce noticeable changes in the mower geometry, local background content, or overall appearance across the displayed frames. Ours more consistently preserves the rider–mower configuration and surrounding scene structure.

These visual comparisons complement the quantitative results in Table 2. They illustrate that temporal allocation affects not only the strength of semantic conditioning, but also how strongly the guided trajectory perturbs human appearance, interaction geometry, and scene structure during generation. The results therefore support our motivation for regulating when strong semantic intervention is applied rather than using a uniformly amplified guidance signal throughout sampling.

E ROBUSTNESS UNDER HIGH GUIDANCE SCALES

We further examine the behavior of different guidance strategies under unusually large semantic guidance scales. Unlike the operating point used in our main experiments, this analysis intentionally increases the text guidance scale to $\lambda_T = 20$ and $\lambda_T = 30$ as qualitative stress tests. The goal is not to optimize performance at these settings, but to examine how rapidly visual and structural artifacts emerge when the magnitude of semantic intervention is substantially increased.

E.1 QUALITATIVE STRESS TEST

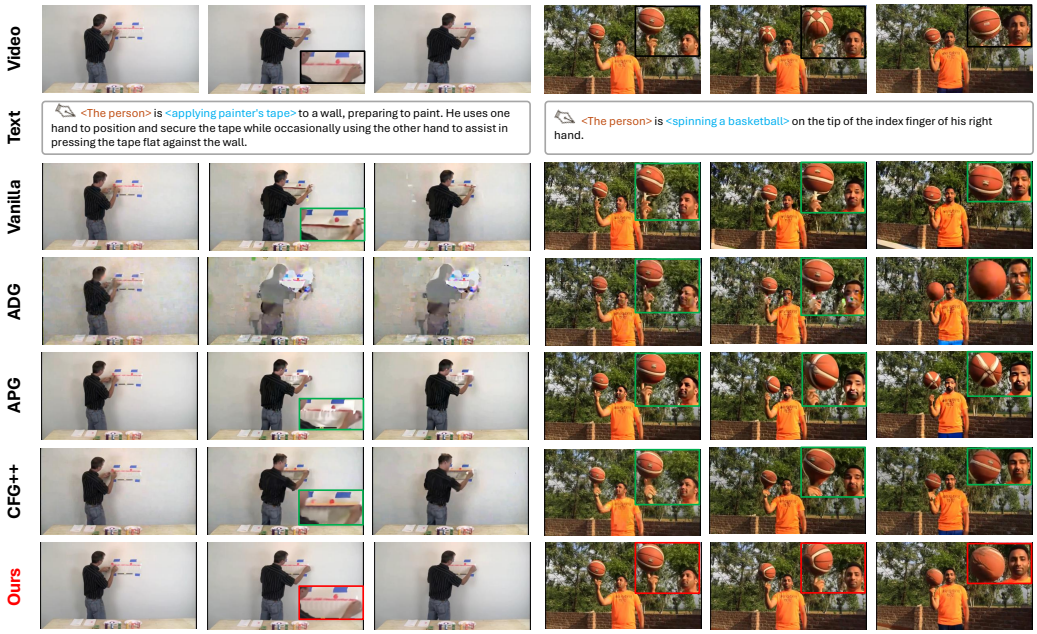


Figure 6: **Qualitative comparison at a high guidance scale ($\lambda_T = 20$).** At this elevated guidance strength, several baseline methods begin to exhibit stronger appearance shifts, local blurring, or background deformation. Ours retains more stable object geometry and scene structure in the highlighted regions while preserving the requested semantic content.

High guidance scale ($\lambda_T = 20$). As shown in Fig. 6, increasing the semantic guidance scale substantially beyond the default operating range exposes different robustness characteristics across guidance methods. Vanilla CFG and CFG++ exhibit stronger color and appearance changes in several examples, while ADG shows more noticeable local deformation in challenging regions. In comparison, Ours preserves the displayed human–object configuration and background structure more consistently. This behavior is consistent with our bounded steering formulation, which limits angular deviation and normalizes the magnitude of the recomposed prediction rather than allowing unrestricted extrapolation.

Extreme guidance scale ($\lambda_T = 30$). At $\lambda_T = 30$, the differences become more pronounced (Fig. 7). Several baselines exhibit severe local distortion, loss of object shape, or degradation of interaction geometry under the amplified semantic intervention. Ours remains visually more stable in the displayed cases, although degradation can still occur at such an extreme operating point. We attribute this relative robustness to the combination of response-induced local allocation, bounded angular steering, and magnitude stabilization, which together constrain how strongly the target-specific conditional response can perturb the underlying structured prediction.

These experiments offer a stress test for different guidance scale scenarios. Together with the guidance-scale sensitivity analysis in the main paper, they indicate that the proposed response regula-

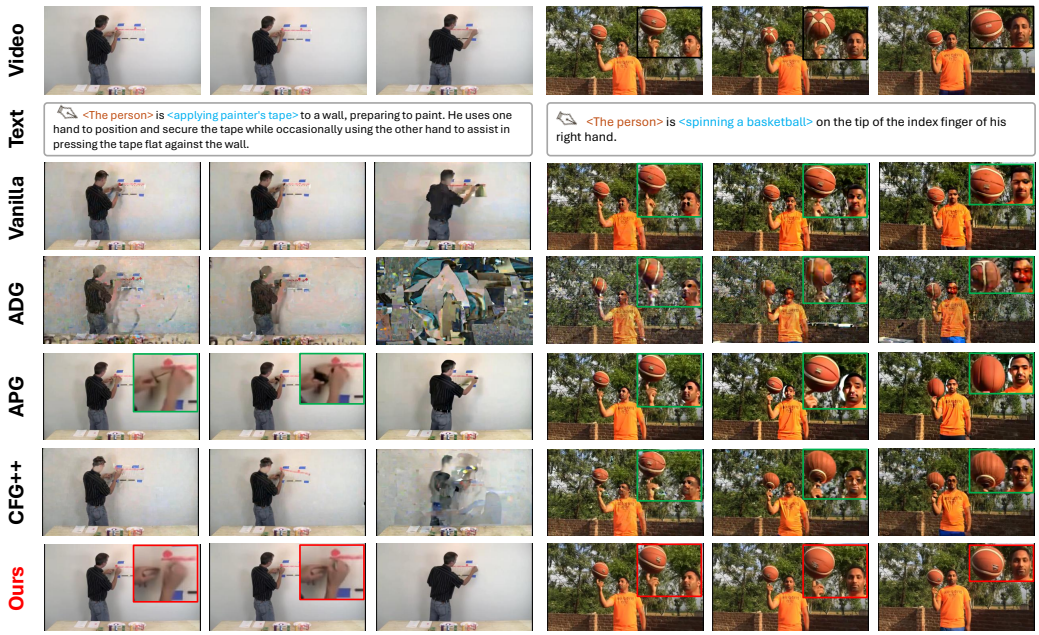


Figure 7: **Extreme guidance-scale stress test ($\lambda_T = 30$).** Further increasing the semantic guidance strength causes substantially stronger visual and geometric degradation for several competing guidance methods. Ours remains comparatively stable in the displayed examples, retaining more recognizable object geometry and human–scene structure under the same high-scale setting.

tion degrades more gracefully when semantic intervention is pushed far beyond the default operating regime.

E.2 Quantitative Stability Profile across Scales

To provide a rigorous numerical evaluation of our method's robustness, we plot the **MPJPE** (Mean Per-Joint Position Error) trend across $\lambda \in \{5, 15, 20, 30\}$ in Fig. 8.

Observation: The Bifurcation of Stability. The quantitative results reveal a clear performance bifurcation beyond the standard scale of $\lambda = 15$. For baselines such as *Vanilla* and *CFG++*, the MPJPE exhibits a steady upward drift, signaling early-stage structural distortion. More drastically, *ADG* suffers from an exponential error explosion at $\lambda = 30$ ($> 580mm$), corresponding to the total manifold collapse visualized in Fig. 7.

The Resilience of Our Method. In sharp contrast, our method is the only framework that maintains a stable, and even slightly improving, error profile as semantic pressure increases. While other methods succumb to **Manifold Inflation**, our method rectifies the latent trajectory. This allows for deeper semantic penetration (as evidenced by our qualitative results) without the "stability tax" typically paid by traditional CFG-based methods.

F Extended Discussions and Limitations

F.1 Latency Analysis

Although our method has given a relatively strong statistical performance on several metrics, the requirement for more computation within one CFG loop would lower the inference latency. As for on a single card NVIDIA H200 GPU, the inference second per frame is 2.52. As for our method, this number climbs up to 4.07.

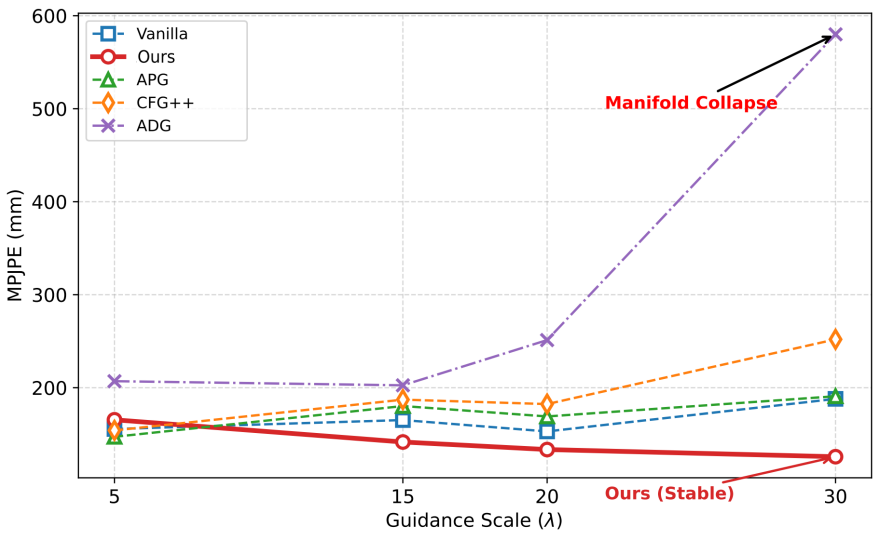


Figure 8: **Structural-error trend across guidance scales.** MPJPE is reported across increasingly strong guidance settings as a supplementary robustness diagnostic.

F.2 Method Limitation

Our method is initially a post-training classifier free guidance strategy, therefore it is different from those human video generation methods Zhang et al. (2025a); Hu et al. (2025); Xu et al. (2026) that try to formulate human actions, no matter before or during training. Therefore, there would be some pipelines that are more effective in statistical results than our method. But our method does not require extra adjustment to other pipelines, and requires little computing resources to re-implement it on other baselines. This guarantees a strong generalization ability for our method.