

Towards Highly-Constrained Human Motion Generation with Retrieval-Guided Diffusion Noise Optimization

Supplementary Material

This supplementary material documents additional implementation details (Section A), experimental details (Section B), results and analyses (Section C), task configuration for quantitative and qualitative evaluation (Section D) and further discussions (Section E) for the proposed method. The video results are also provided in the supplementary material.

A. Implementation Details

LLM-based Relational Task Parsing. We implement automatic relational task parsing using the large language model DeepSeek-R1 [2]. The instruction comprises a task format and reasoning rules, along with an example scenario case that the LLM reasons for itself based on the given task description. We feed the textual description or optionally the code implementation of the combined constraint function to the LLM, asking it to identify the difficult constraint c_D and its relationship with other constraints. The instruction is shown in Fig. 9. The retrieval confidence score is manually specified for the random noise fitting set C_1 .

Semantic Check. We implement semantic check based on keyword compliance. We consider two types of keywords: action verbs and descriptive adverbs for spatial relations. Specifically, we filter out motion samples that (1) contain contradictory adverbs, and (2) lack the specified action verb when the task is defined by numerical constraints. For example, given the prompt “a man walks forwards”, motion samples with annotation *backwards* are filtered out.

Choices of Masks. To determine the mask type, we run Eq. (13) for both 2^{N_T} temporal masks and 2^{N_S} spatial masks, and pick the one with the smallest loss. Given the task description, we can also directly parse the mask type that is more reasonable. Currently, we avoid spatiotemporal masks as they can easily degrade the content preserved in the diffusion noise z_1 and z_2 .

Soft Mask Blending. To refine sharp boundaries in the binary mask candidate M , we further optimize a soft value mask in *sigmoid* representation. First, we transform the optimized M into an approximate sigmoid form as $M = \sigma(M')$ so that $z' = \sigma(M')z_1 + (1 - \sigma(M'))z_2$. We then perform another round of optimization on M' based on Eq. (13) in the main paper. In this way, the mask M becomes smoother and its values still lie in $[0, 1]$.

Reward Function. Eq. (14) in the main paper represents a general form of the reward function and the default value for λ_k is set to 1.0. Depending on each individual generation task, we can adjust the weight λ_k and omit cer-

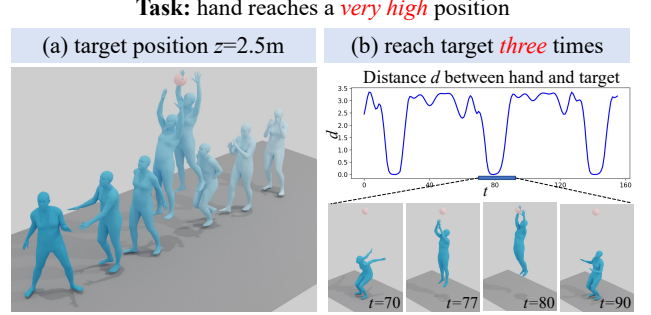


Figure 6. Qualitative examples for Task: *hand reaches a very high position* with different types of constraints. (a) Spatial constraint: the target is located at $z = 2.5$ along the walking path. (b) Numerical constraint: the target is located at the origin and the goal is to reach the target position three times.

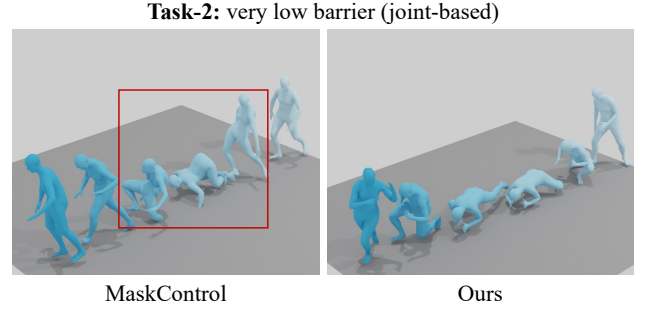


Figure 7. Comparison with MaskControl for Task-2 in the joint-based form. MaskControl generates implausible poses given very challenging constraints. The barrier is not visualized for clarity.

tain terms for simplicity. In our experiments, for Task-1, the reward function is designed as $\mathcal{R} = \mathcal{L}_{\text{jitter}} + \mathcal{L}_{\text{decorr}} + 10\mathcal{L}_{\text{foot skate}}$. For Task-2, the reward function is designed as $\mathcal{R} = \mathcal{L}_{\text{jitter}} + \mathcal{L}_{\text{decorr}}$. For Task-3, the reward function is designed as $\mathcal{R} = \mathcal{L}_{\text{foot skate}} + 0.5\mathcal{L}_{\text{semantic}}$. Following DNO [4], we adopt a simplified semantic score for *raising hand* as an indicator function for whether the hand height is above a threshold. More precise text-action similarity score can also be adopted. For the reward function, we use a rectified form of foot skating by defining the ground plane as the lowest foot height in the generated motion sequence.

B. Experimental Details

Evaluation Metrics. We calculate the metric of maximum scene penetration for Task-1 and Task-2. For Task-1 *very narrow gap*, it is defined as the sum of maximum horizon-

tal penetration depths of the body surface vertices into each wall. For Task-2 *very low barrier*, it is defined as the sum of maximum vertical penetration depths of the body surface vertices into the overhead barrier and the ground. For Task-3, the success rate is defined as the percentage of samples with the correct number of walking steps. The number of walking steps for the generated motion is calculated by counting rising or falling edges of the foot velocity when crossing a predefined threshold. For semantic success rate, similar to DNO [4], we further check whether the generated motion contains the action *raising hand*. We measure whether there is at least a frame in the generated motion in which the hand is above the shoulder. For the ablation study in Table 4, the local foot skating is calculated where the walking distance z is within the range $4 < z < 5$. The ground height is rectified and the foot height threshold is set to 0.025 m.

Baseline Details. For Unconstrained MDM-RoHM, we exclude shape parameters from RoHM representation and derive a 284-dimensional pose feature. The root trajectory is recovered from the relative trajectory representation and the joints and vertices are both recovered from SMPL parameters. We train the motion diffusion model with 600,000 steps with an initial learning rate of 0.0001. The DDIM-50 model is used for generation.

C. Additional Experiments and Analyses

More Qualitative Examples. In Fig. 6, we present results for the task *hand reaches a very high position* under different types of constraints, i.e., spatial and numerical constraints. The numerical constraint for reaching the target three times is implemented as a differentiable function that counts prominent minima in the distance between hand and target, as illustrated in Fig. 6 (b).

More Quantitative Results. Previous works [4, 5] typically report results on very limited handcrafted tasks for evaluating zero-shot constraint-based motion generation. While it is a common challenge to design a comprehensive benchmark for this field, we construct a larger benchmark comprising nine cases derived from this work, covering varied constraint types, task scenes and control joints. The benchmarking tasks include cases in Table 1, Table 2, Fig. 3 and Fig. 6, and further include the task *very high hurdle* in which a human jumps over a high hurdle. The full result is shown in Table 6. With the proposed retrieval guidance, we achieve lower constraint error on most of the tasks compared to ProgMoGen+DNO.

Effect of Relational Task Parsing. It is useful for relational task parsing to include difficult constraint c_D with its tightly related constraints into the retrieval set C_R . For example, retrieving jointly for the difficult constraint *walk five steps* together with the constraint *walk four meters* yields more effective guidance for optimizing the whole task than

| Constr. Type | Task Description | Ctrl Joint | C.Err(DNO) | C.Err(Ours) ↓ |
|-----------------------|-----------------------------|------------|---------------|-----------------|
| strict spatial | very narrow gap | full body | 0.0162 | 0.0050 |
| | very low barrier | full body | 0.000115 | 0.000049 |
| | very high hurdle | full body | 0.000 | 0.000 |
| | reach high position | hand | 0.0013 | 0.0015 |
| | reach high position | foot | 0.0013 | 0.0007 |
| numerical | walking steps w/ action | foot | 0.282 | 0.0003 |
| | number of claps | hand | 3.827 | 0.480 |
| | number of target touches | hand | 9.526 | 4.190 |
| combined | overhead barrier+walk steps | full body | 2.250 | 0.313 |

Table 6. Quantitative comparison on the expanded benchmark.

| Task-2: very low barrier | | | |
|---|------------|----------|----------|
| Reward \mathcal{R} | Foot Skate | Max Acc. | C.Err. |
| MotionCritic | 0.254 | 0.217 | 0.000047 |
| $\mathcal{L}_{\text{jitter}} + \mathcal{L}_{\text{decorr}}$ | 0.228 | 0.194 | 0.000049 |
| $\mathcal{L}_{\text{jitter}} + \mathcal{L}_{\text{foot skate}} + \mathcal{L}_{\text{decorr}}$ | 0.228 | 0.192 | 0.000059 |

Table 7. Comparison on the different reward function design.

retrieving based on *walk five steps* only. From Fig. 3 of the main paper, we see that the LLM can well infer the retrieval subset C_R and the target position constraint is correctly identified to be included. In practice, during retrieval, we set a smaller weight for the target position constraint in order not to overshadow the most difficult one c_D .

Reward Function Design. We also experiment on MotionCritic [9] which is a motion quality reward learned from preference alignment. We use $\mathcal{R} = \sigma(-s)$ to transform the reward into loss where s is the reward score and σ is the sigmoid function, and set the weight λ to 1.0. The result is shown in Table 7. We find that it is not as effective as common heuristic quality check functions for pruning invalid masked noise compositions. Also, from Table 7, we see that the design of reward function is not highly sensitive.

Comparison with MaskControl. We also conduct qualitative comparison with MaskControl [8]. Since MaskControl only supports joint-based constraints, we compare on the joint-based version of Task-2. In this modified task, all joints are constrained to be below $h = 0.5$ m to avoid the overhead barrier. As in Fig. 7, MaskControl generates implausible poses when guided by very difficult constraint functions such as a planar barrier. In contrast, our method generates physically plausible and natural poses.

Text Guidance for Numerical Constraints. For Task-3, we observe that injecting numbers into prompts does not improve success rates, showing that the current generation models still have difficulty generating accurate action counts. Recently, ATOM [3] enables control over action frequency with text prompt by finetuning on preference annotations. However, it only supports a limited range of numerical values and cannot handle customized constraints as in our highly-constrained generation tasks.

Analysis on Motion Diversity. During masked noise optimization, the mask is optimized to adapt to different randomly sampled noise z_0 , introducing diversity for the generated motions. In the supplementary video results, we show that for the task *very narrow gap*, the generated motions exhibit various local details under the same retrieval, such as speed difference while passing through the gap.

Analysis on Retrieved Motions. By minimizing constraint error for the constraint subset C_R , the retrieved samples possess certain motion skills for tackling these difficult constraints. However, the retrieved samples may not necessarily satisfy all the constraints in C_R , which still needs to be solved in the final phase of diffusion noise optimization. For example, for the task *walk and clap hands wide for four times*, the retrieved motion contains the pattern of clapping four times, but lacks the required motion amplitude (see the visualization in the supplementary video results).

D. Detailed Task Configuration

We provide detailed constraint functions for each task in the quantitative evaluation and qualitative examples. As counting-based numerical constraints (e.g. the number of walking steps) are challenging to design by an end user, following ProgMoGen [5], we first ask an LLM to provide a reasonable form for the constraint function, and then manually refine it and set appropriate parameters.

D1. Tasks for Quantitative Evaluation

For **Task-1** *very narrow gap*, the goal is to pass through a narrow gap formed by two walls and reach the target position. The total constraint function consists of four parts: (1) `loss_narrow_gap` represents a narrow gap of 0.4 meters in width and 3 meters in length, constraining horizontal positions of all body joints and SMPL vertices to the range $-0.2 < x < 0.2$ when the walking distance is in the range $0.5 < z < 3.5$; (2) `loss_target_pos` constrains the pelvis joint to reach the target position $z = 5$; (3) `loss_self_collision` is a simplified term to avoid collision between arms and the body, constraining the distances between arm/wrist joints and spine joints to be larger than 0.2 m; (4) `loss_ground_contact` constrains the foot to be close to the ground at beginning and end frames. The text prompt is “a man walks forwards”. The sequence length for generated motions is set to 100 and 150 frames for evaluating on two scenarios of fast and slow walking.

For **Task-2** *very low barrier*, the goal is to pass through a low overhead barrier and reach the target position. The total constraint function consists of four parts: (1) `loss_barrier` represents a barrier of 0.5 meter in height and 1 meter in length, constraining vertical positions of all body joints and SMPL vertices to the range $0 < y < 0.5$ when the walking distance is in the range $2 < z < 3$; (2) `loss_target_pos` constrains the pelvis and head

joints to reach $z = 5$; (3) `loss_y_tstart` constrains the pelvis height to be 0.9 m at the beginning frame and (4) `loss_y_tend` constrains the pelvis height to be 0.9 m at the end frame, encouraging a normal standing pose on both ends. The text prompt is “a man walks forwards”. The motion sequence length is set to 100.

For **Task-3** *assign number of walking steps*, the goal is to reach the target position with a specified number of steps. The total constraint function consists of three parts: (1) `loss_walk_step` constrains the number of walking steps to be 6; (2) `loss_target_pos` constrains the pelvis joint to reach $z = 4$; (3) `loss_foot_both_ends` constrains both feet to be on the ground with zero velocity at beginning and end frames. The text prompt is “a man walks forwards and raises up hands at the same time”. The motion sequence length is set to 100. Specifically, `loss_walk_step` is implemented as a differentiable peak counting function for foot velocity, where each prominent peak in the smoothed foot velocity signal roughly corresponds to one stepping action. It operates by summing the sigmoid-weighted outputs for all frames identified as prominent peaks:

$$\text{step_number} = \sum_t \sigma(T(v[t] - v[t+1])) \cdot \sigma(T(v[t] - v[t-1])) \cdot \sigma(T(v[t] - \theta_v)) \quad (15)$$

where v is the foot velocity, σ is the sigmoid function, θ_v is the velocity threshold and T is the scaling factor. We set a large $T = 10000$ to get an accurate step number estimation for comparing against the required number during optimization. We calculate the total number of walking steps on both feet.

D2. Tasks for Qualitative Examples

Qualitative results for Task-1 and Task-3 are shown in Fig. 1 of the main paper, and a qualitative example for Task-2 is shown in Fig. 3 of the main paper. Details of the remaining tasks from Fig. 3 of the main paper are provided below.

For the task *walk with foot reaching a very high position*, the goal is to reach the target destination $z = 5$ and also reach a position of 1.8 m high located at $z = 2.5$. The total constraint function consists of three parts: (1) `loss_reach_high` constrains one of two feet to reach the high position; (2) `loss_target_pos` constrains the pelvis joint to reach $z = 5$ at the end frame; (3) `loss_foot_contact_both_ends` constrains two feet to be on the ground at beginning and end frames.

For the task *walk and clap hands wide for four times*, the goal is to walk to the target position while performing a specified number of hand claps. The total constraint function consists of four parts: (1) `loss_clap_times` constrains the character to clap its hands four times, which

is implemented as a differentiable function that counts the prominent minima of the distance between the hands; (2) `loss_clap_wide` constrains the inter-hand distance to increase rapidly near the minima point and exceed a threshold; (3) `loss_target_pos` constrains the pelvis joint to reach the target $z = 4$ at the end frame; (4) `loss_hand_contact` constrains the distance between the hands to be close to zero at prominent minima points. Specifically, `loss_clap_times` is implemented as a mean square error function to compare groundtruth clap times with the generated motion, and the clap times of the generated motion can be obtained as:

$$\text{clap_times} = \sum_t \sigma(T(d[t+1] - d[t])) \cdot \sigma(T(d[t-1] - d[t])) \cdot \sigma(T(\theta_d - d[t])) \quad (16)$$

where d is the distance between two hands, σ is the sigmoid function, θ_d is the distance threshold, and T is the scaling factor. We set $T = 10000$ for an accurate clapping number estimation.

For the task *walk to avoid overhead barrier in five steps*, the goal is to reach the target position with a specified number of steps and at the same time avoid an overhead barrier of 1.0 m high located at the region of $2 < z < 3$. The constraint functions are: (1) `loss_barrier` constrains all the joints and SMPL vertices to be lower than 1.0 m and above the ground for the walking distance $2 < z < 3$; (2) `loss_walk_steps` constrains the character to walk for five steps. See its form in Eq. (15); (3) `loss_target_pos` constrains the pelvis joint to reach the target $z = 4$ at the end frame; (4) `loss_foot_both_ends` constrains the feet to be on the ground at beginning and end frames.

E. Additional Discussions

Limitations and Future Work. For our method, the foot skating remains at a similar level but larger than ProgMo-Gen+DNO [5, 4]. Also, some generated motions of our method may contain unnatural motion transition as in Fig. 8 (left). More delicate attentional-layer-based masked noise composition can be designed to further improve motion quality. Second, the reward guidance is not imposed on the final generated motion, limiting the capability of precise text alignment (see Fig. 8 (right)). The semantic alignment term (e.g. VQA score [6]) can be included as a constraint function when text alignment should be strictly enforced. Last, the retrieval configuration involves semantic consistency check, top candidate picking, and coefficients for the retrieval and reward function, which may require some tuning. As this work introduces a general framework for retrieval-based diffusion noise optimization, its re-

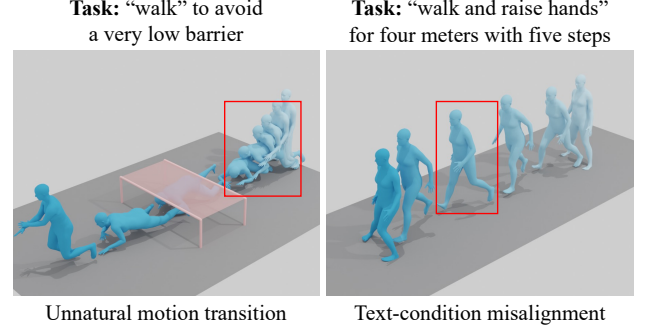


Figure 8. Failure cases. Some generated motions of our method may contain unnatural motion transition (left) or may not fully adhere to the text condition (right).

trieval module can be a focus of future improvement. We can further improve its performance by refining retrieval candidates via re-ranking, using larger in-the-wild motion datasets [1], adopting motion clipping and constraint re-writing, and extending to more than one retrieval samples to handle tasks with multiple difficult constraints. While higher-level semantics in motion generation is generally best to be described by text condition itself, one potential application of our method is for abstract actions if they are difficult to be solved by text alone. Such actions can be decomposed into textual and non-textual constraints, where the proposed retrieval guidance can be employed to resolve the challenging non-textual ones.

Differences from other related methods. The motion editing method such as native DNO [4] requires a high-quality reference motion, which is not available in the highly-constrained generation task. STMC [7] is a purely text-driven spatial and temporal composition method and cannot handle customized constraints. Different from these approaches, we provide a unified framework for determining what to retrieve and finding the composition of the retrieved noise and random noise. Moreover, we especially focus on zero-shot goals based on a set of constraint functions, under the training-free framework. For control tasks based on complex joint trajectories or key-frames, methods built upon conditional training, such as MaskControl [8], are more specialized.

References

- [1] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Juntao Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13336–13348, 2025.
- [2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning ca-

pability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

- [3] Haonan Han, Xiangzuo Wu, Huan Liao, Zunnan Xu, Zhongyuan Hu, Ronghui Li, Yachao Zhang, and Xiu Li. Atom: Aligning text-to-motion model at event-level with gpt-4vision reward. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22746–22755, 2025.
- [4] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1334–1345, 2024.
- [5] Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2024.
- [6] Boming Miao, Chunxiao Li, Xiaoxiao Wang, Andi Zhang, Rui Sun, Zizhe Wang, and Yao Zhu. Noise diffusion for enhancing semantic faithfulness in text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23575–23584, 2025.
- [7] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track time-line control for text-driven 3d human motion generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1911– 1921. IEEE Computer Society, 2024.
- [8] Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9955–9965, 2025.
- [9] Haoru Wang, Wentao Zhu, Luyi Miao, Yishu Xu, Feng Gao, Qi Tian, and Yizhou Wang. Aligning human motion generation with human perceptions. In *The Thirteenth International Conference on Learning Representations*.

In a human motion generation task given as a set of constraints C, you are asked to perform relational task parsing by reasoning about the order of difficulty and relationship between constraints. So first, you should identify the most difficult constraint c_d , and then parse the relationship between c_d and other constraints which the relation type are in {'connected', 'conflicting', 'none'}. The type 'connected' means two constraints are tightly related to each other. The type 'conflicting' means two constraints are conflicting, for example, a constraint requiring very low head height in the middle frame and very high head height at the final frame would be considered conflicting. The type 'none' means there are no strong relationship between them. So your answer should return c_d , c_d 's connected constraints and c_d 's conflicting constraints.

Here is an example: [\[An example scenario that LLM devises for itself.\]](#)

Also, the counting-based objective is considered to be more difficult than spatiotemporal constraints unless the spatiotemporal constraint is extremely challenging.

Now, I will give you a set of constraints.

Example Scenario: Constraint Set C

Let's imagine we are generating a motion for a character to perform a "Dive and Roll" maneuver under a low obstacle. The constraints are:

- c1 (Start): Standing upright, hands at sides.
- c2 (Middle): Head height must be below 0.5 meters (to pass under a table).
- c3 (Middle): Body must be fully extended in a dive posture.
- c4 (End): End in a standing upright position.
- c5 (End): Right hand must be touching the right knee.

Step 1: Identify the Most Difficult Constraint (c_d)

To identify the most difficult constraint, we look for the one that is most specific, physically demanding, and likely to be the primary driver of the overall motion structure.

- c1 (Start): A common, neutral pose. Not difficult.
- c2 (Middle): Very specific spatial requirement (low head height). This forces a major deviation from a normal standing pose.
- c3 (Middle): A specific, dynamic full-body posture. Difficult, but common in motions like dives.
- c4 (End): A common, neutral pose. Not difficult by itself.
- c5 (End): A specific, but relatively simple, end-state contact constraint.

Analysis: While c3 is difficult, c2 (head height < 0.5m) is arguably the most difficult. It is an extreme spatial limitation that dictates the entire trajectory of the spine and head. Satisfying c2 requires a deep crouch, a dive, or a roll, and it directly conflicts with the starting and ending upright postures. The motion essentially exists to satisfy this single, hard constraint.

Therefore, the most difficult constraint is: $c_d = c2$ (Middle): Head height must be below 0.5 meters.

Step 2: Parse Relationships with c_d

Now, we will analyze the relationship between c_d (c2) and every other constraint.

c_d (c2) vs. c1 (Start: Standing upright)

Relationship: Conflicting

Reasoning: The start pose is upright (head height ~1.7m), but the middle frame requires an extremely low head height (0.5m). The motion must resolve this large, rapid spatial displacement, making the constraints inherently conflicting in their required states.

c_d (c2) vs. c3 (Middle: Body fully extended in a dive)

Relationship: Connected

Reasoning: These two constraints apply to the same temporal segment (Middle). A fully extended dive posture typically raises the head and torso. To satisfy the low head height (c2) during a dive (c3), the entire dive trajectory must be angled sharply downwards. The solution for one is entirely dependent on the other; they are tightly coupled and must be solved together.

c_d (c2) vs. c4 (End: Standing upright)

Relationship: Conflicting

Reasoning: This is similar to the conflict with c1. The motion must go from a very low head height in the middle frame back to a high head height in the final frame. This creates a strong kinematic conflict that the motion must resolve, likely through a rapid recovery like a roll or a powerful leg extension.

c_d (c2) vs. c5 (End: Right hand touching right knee)

Relationship: None

Reasoning: The low head height constraint c2 is a mid-motion spatial requirement. The hand-knee contact c5 is a final pose configuration. There is no direct physical or logical link between them. The motion can satisfy the low head height and then, during the recovery to the standing pose, easily achieve the hand-knee contact without the two constraints interfering or depending on each other.

Final Answer

Based on the relational task parsing of the provided example constraint set C:

Most Difficult Constraint (c_d):

c2 (Middle): Head height must be below 0.5 meters.

c_d 's Connected Constraints: [c3]

c3 (Middle): Body must be fully extended in a dive posture.

c_d 's Conflicting Constraints: [c1, c4]

c1 (Start): Standing upright, hands at sides.

c4 (End): End in a standing upright position.

Figure 9. Automatic LLM-based relational task parsing. The above shows the instruction comprising a task description, reasoning rules, and an example case that the LLM devises and reasons itself which is shown in the below.