Guiding Human-Object Interactions with Rich Geometry and Relations

Supplementary Material

1. Overview

This supplementary material provides detailed insights into our method, covering evaluation details (Section 2), extended experiments on guidance mechanisms, T2M-BEHAVE Dataset, and novelty assessment (Section 3), performance comparisons with baselines (Section 4), user study results (Section 5), and qualitative results demonstrating the model's robustness (Section 6). Additionally, we include comprehensive details on motion representation and architecture for clarity and reproducibility (Section 7), along with a discussion on hand motion synthesis and future extensions (Section 8).

2. Evaluation Details

Text and HOI Feature Extraction. Currently, there are no publicly available feature extractors specifically designed to evaluate human-object interaction motions. To address this limitation, we take inspiration from T2M [1] and adopt a similar evaluation framework. Our method converts the textual descriptions into feature vectors with a frozen CLIP text encoder. At the same time, the generated HOI sequences are processed using an HOI feature extractor based on a bidirectional GRU (BiGRU) model. Moreover, we modify the input dimensions of the BiGRU model according to the representation of HOI sequences. Specifically, the length of the representation is 216, with 72 dimensions allocated to the human's 24 joints, 132 dimensions representing 6D continuous rotations, 9 dimensions for the object's rotation matrix, and 3 dimensions for object transformations. By minimizing the distance between features of the matched text-HOI pairs, this approach establishes a strong alignment between the textual descriptions and HOI motion sequences.

Contact Percentage. Following the method in CHOIS [2], we compute the Contact Percentage by calculating the minimum distance between hand joints and object vertices at each frame. A 5 cm threshold is used to determine contact. The Contact Percentage is then the ratio of frames with contact to the total number of frames, representing the proportion of time during which hand-object contact occurs.

Collision Percentage. In line with OMOMO [3], we compute the Collision Percentage by querying the object's Signed Distance Function (SDF) at each time step for each vertex on the reconstructed human mesh. A 4 cm threshold is applied to detect collisions, where a collision is counted if the signed distance is negative and its absolute value exceeds 4 cm. The Collision Percentage is calculated as the ratio of frames with collisions to the total number of frames in the sequence.

Motion Deviation. Following the approach in ARCTIC [4], we define the Motion Deviation (MDev) metric to assess the consistency of motion between the hand and object vertices in a HOI sequence. MDev measures the difference in movement directions between the hand and object vertices over consecutive frames within a window (m, n). Let h_i^t and o_j^t represent the hand and object vertices at frame t, respectively. A contact window (i, j, m, n) is defined as the longest period during which h_i^t and o_j^t remain within a threshold $\alpha = 5\,\text{cm}$ for all frames between m and n, with no contact at frames $m-1$ and $n+1$. The MDev is then calculated as:

$$\text{MDev} = \frac{1}{n-m} \sum_{t=m+1}^{n} \left\| \left(\hat{\mathbf{h}}_i^t - \hat{\mathbf{h}}_i^{t-1} \right) - \left(\hat{\mathbf{o}}_j^t - \hat{\mathbf{o}}_j^{t-1} \right) \right\|,$$

where MDev quantifies the movement consistency between the contacting hand and object vertices. The final MDev value is the average of MDev values across all detected windows, with the result expressed in millimeters (mm).

3. Additional Experiments

Impact of Guidance Mechanism at Different Noise Levels. In this section, we examine the impact of the diffusion timestep selection on the effectiveness of IDF Guidance. To assess how different timestep settings affect generation performance, we conducted a series of experiments. As summarized in Table 1, the results show that initiating IDF Guidance during the final 10 timesteps yields the best results. This approach outperforms both applying IDF Guidance throughout all 1000 timesteps and applying it solely during the final timestep, highlighting the importance of timing the guidance effectively for optimal results.

Effect of Guidance Iterations. This section examines the influence of the number of guidance iterations on the performance of IDF-guided denoising. Specifically, we perform k iterations of L-BFGS optimization at each denoising step, where k is treated as a tunable hyperparameter. The experiments are conducted during the last 10 timesteps of the denoising process to refine the generation results. As summarized in Table 2, increasing k to 10 demonstrates the most significant improvement, achieving the best balance across metrics such as Precision, FID, and Motion Deviation. This outcome highlights the importance of iterative

Methods	Precision↑	FID↓	$C_{\%}$	MDev↓
w/o Guidance	0.879	5.726	0.424	7.020
$t \leq 0.001T$	0.888	5.159	0.458	6.435
$t \leq 0.005T$	0.896	5.129	0.463	5.927
$t \leq 0.01T$	**0.902**	**5.119**	**0.466**	**5.815**
$t \leq 0.1T$	0.890	5.399	0.464	6.590
$t \leq 0.5T$	0.879	5.586	0.430	6.938
$t \leq T$	0.879	5.590	0.431	6.954

Table 1. Quantitative results showing the impact of guidance timesteps on generation performance during diffusion. Initiating IDF Guidance during the final 10 timesteps ($t \leq 0.01T$) consistently achieves the best performance across metrics. Precision refers to the R-precision top-3 metric.

Methods	Precision↑	FID↓	$C_{\%}$	MDev↓
w/o Guidance	0.879	5.726	0.424	7.020
$k = 1$	0.883	5.597	0.430	6.837
$k = 5$	0.900	5.347	**0.474**	6.482
$k = 10$	**0.902**	**5.119**	0.466	**5.815**

Table 2. Effect of guidance iterations (k) on generation performance during the final 10 denoising steps. Increasing k enhances performance, with $k = 10$ achieving the best results.

$R(\cdot)$	Params	Precision↑	FID↓	$C_{\%}$	MDev↓
MDM	28.79M	0.887	5.529	0.370	6.204
MDM	30.89M	0.890	5.358	0.431	7.401
MDM	33.00M	0.896	5.384	0.410	6.703
Ours	29.36M	**0.902**	**5.119**	**0.466**	**5.815**

Table 3. Performance comparison of relation models $R(\cdot)$. Ours, based on the VDT model with spatial and temporal attention mechanisms, achieves superior precision, FID, and MDev metrics compared to the MDM model, despite having comparable or smaller parameter sizes.

refinement in aligning distributions, as second-order optimization enables more precise adjustments and accelerates convergence towards the target solution.

Influence of the Relation Model. We evaluated the impact of the Relation Model by replacing the original VDT model [5] with the MDM model [6] and increasing the number of Transformer encoder layers in MDM to expand its parameters during training. The results, summarized in Table 3, indicate that VDT consistently outperforms MDM across all metrics. This advantage is attributed to VDT's spatial and temporal self-attention mechanisms, which enable it to effectively capture complex dependencies, leading to higher-quality outputs.

Method	FS ↓	$C_{\%}$	$Coll_{\%}$	MDev↓
GT	0.071	0.410	0.370	8.989
HOI-Diff[7]	0.182	0.191	0.259	24.807
Ours	**0.174**	**0.239**	**0.195**	**10.784**

Table 4. T2M-BEHAVE [7] cross-benchmark tests: 24.6% lower collisions vs HOI-Diff (0.195 vs 0.259), despite dataset's compact scale.

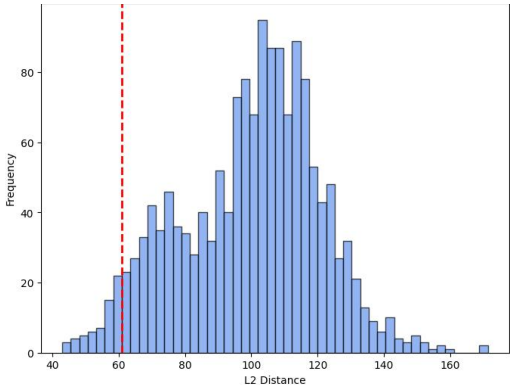

Figure 1. Novelty analysis. L2 distances of 512 test cases vs. top-3 training neighbors. Red lines mark intra-trainset distance, demonstrating generation beyond training data.

Results on T2M-BEHAVE Dataset. As FullBodyManipulation (FBM) is much larger than other text-conditioned HOI datasets, both our work and CHOIS [2] adopt FBM only for benchmarking. We have included evaluations on T2M-BEHAVE [7] in Table 4. The results highlight inherent dataset characteristics, with ground-truth (GT) collision metrics registering at 0.370. While the dataset's compact scale and lack of dedicated interaction-focused evaluation protocols limit comprehensive benchmarking, our method demonstrates improved physical plausibility by reducing collisions to 0.195 compared to HOI-Diff's 0.259. The motion quality metrics (FS: 0.174 vs. 0.182; MDev: 10.784 vs. 24.807) indicate our approach maintains reasonable fidelity.

Retrieval vs Generation Following CG-HOI [8], Figure 1 first detects top-3 nearest training samples for each of 512 test cases, then plots an L2 distance histogram. It also marks the intra-trainciteset distance with a red line. It shows the generated motions mostly fall outside the intra-trainset distance, which confirms our model produces novel motions rather than retrieval.

4. Model Complexity and Inference Efficiency

Table 5 compares computational costs for 10-second motion generation on an RTX 4090 GPU, with all models trained for 300,000 steps. Our method combines a motion diffusion model (17.98M parameters) and a relation diffusion model

Metric	InterGen	MDM	HOI-Diff	CHOIS	Ours
Params(M)	54.24	17.91	47.74	13.25	47.34
Inference time(s)	2.0	0.36	3.68	2.93	8.1

Table 5. Comparison of model complexity and inference efficiency.

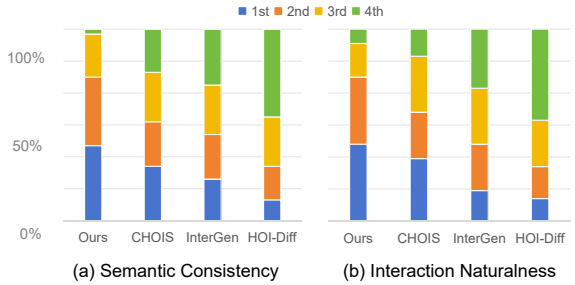

Figure 2. User study. We generated HOIs for 15 captions using 4 methods and asked 20 users to rank them by text alignment and realism. Our method outperforms others in both aspects.

(29.36M parameters). With a batch size of 32, our method achieves an average of 8.1 seconds per sequence, remaining viable compared to baselines (MDM: 0.36s, HOI-Diff: 3.68s).

5. User study

We conducted a user study comparing four HOI generation methods across 15 text captions, with twenty participants ranking the results based on two evaluation criteria: (1) Semantic Consistency (alignment between animations and text descriptions) and (2) Interaction Naturalness (naturalness of poses and object interactions). As shown in Figure 2, our method outperformed three baseline methods in both metrics, demonstrating superior text-visual correspondence and biomechanical plausibility.

6. Additional Qualitative Results

In this section, we provide additional HOI generation results across diverse settings, highlighting the versatility and effectiveness of our approach.

Consistent Motions across Different Objects. We assess our method's ability to generate consistent motions across different scenarios, such as "lift the object, move the object, and put down the object" across different objects. Figure 3 illustrates the interaction results for a monitor, trashcan, plastic box, large box, and chair. These examples highlight ROG's ability to produce semantically accurate and consistent interactions across a range of objects.

Diverse Motions on the Same Object. We evaluate the ability of ROG to generate diverse motions for the same object. Taking a clothes stand as an example, Figure 4 showcases a range of motions, including "lift the object", "put down the object", and "pull the object". These results demonstrate the flexibility of our approach in producing semantically distinct and contextually appropriate interactions with a single object.

7. Method Details

Motion representation Details. The motion data is represented in a 288-dimensional vector per frame, comprising two primary components: human motion and object-related information. The 204-dimensional human motion data captures both the global 3D coordinates of 24 body joints (24×3) and the 6D rotational parameters for 22 joints (excluding palm joints, 22×6). The remaining 84 dimensions describe object interactions through three key elements: the object's global position (3), its 3×3 rotation matrix (9), and the spatial coordinates of 24 predefined key points on the object's surface (24×3).

Architecture Details. Figure 5 illustrates the details of our relation model, including the IDF encoder and VDT layers.

8. Hand Motion Synthesis and Future Extensions

While our framework currently omits fine-grained hand motion synthesis, this limitation primarily stems from the lack of comprehensive hand-object interaction data in mainstream HOI benchmarks such as FullBodyManipulation and BEHAVE. Existing datasets predominantly focus on coarse full-body dynamics, leaving finger-level articulations underexplored. By extending the Interactive Distance Field (IDF) to incorporate hand-specific keypoints (e.g., fingertip positions), our framework can seamlessly adapt to richer representations. This flexibility demonstrates the broad applicability of our method across different motion scales, enabling the generation of both body movements and hand manipulations when trained on appropriate data.

References

[1] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, L. Cheng. Generating diverse and natural 3D human motions from text. In *CVPR*, 2022. 1

[2] J. Li, A. Clegg, R. Mottaghi, J. Wu, X. Puig, C. Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 1, 2

[3] J. Li, J. Wu, C. Liu. Object Motion Guided Human Motion Synthesis. In *ACM TOG*, 2023. 1

[4] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. Black, O. Hilliges. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *CVPR*, 2023. 1

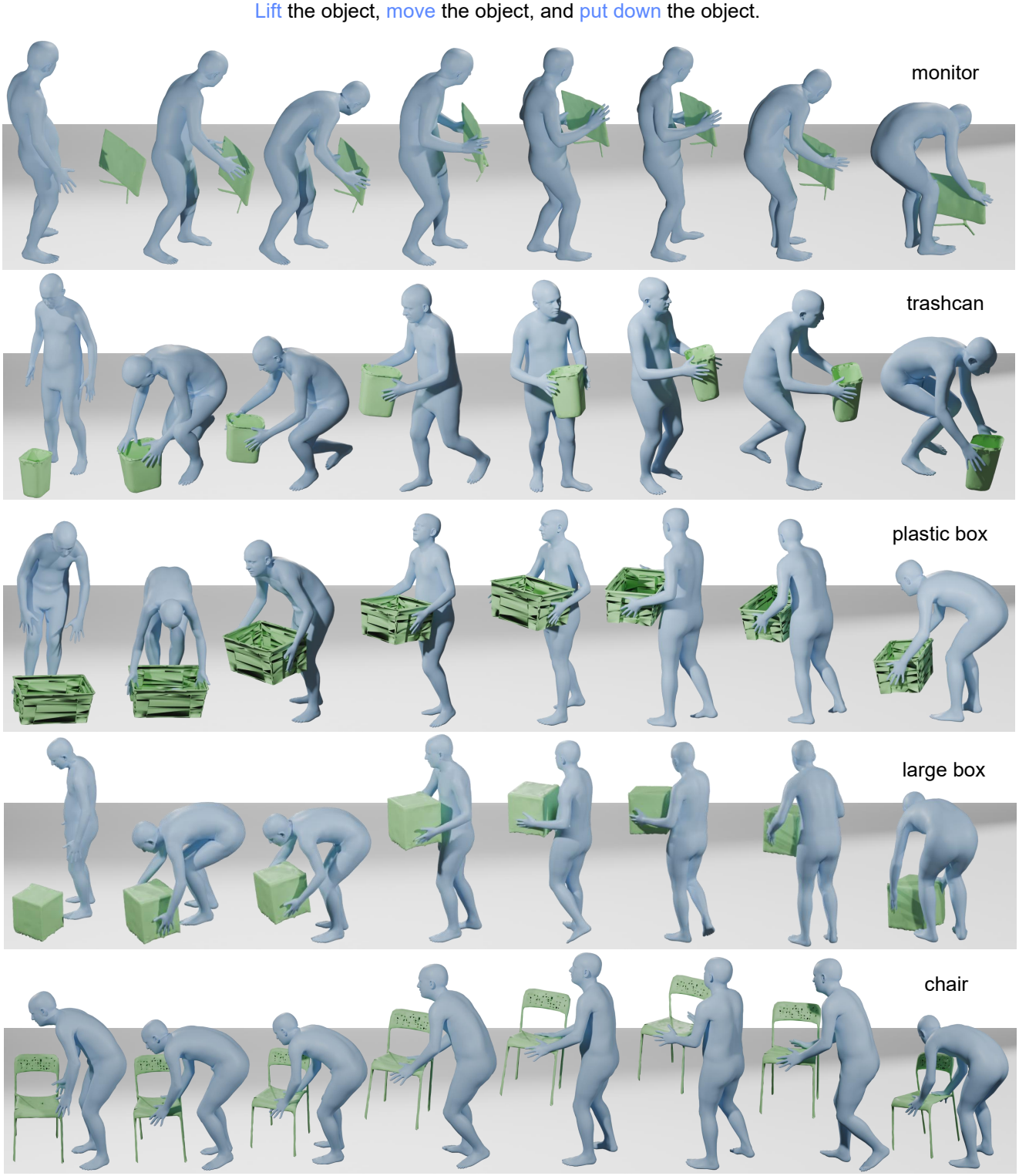

Figure 3. Our model generates semantically accurate and consistent human-object interactions across various objects.

[5] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, M. Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In arXiv:2305.13311, 2023. 2

[6] G. Tevet, S. Raab, B. Gordon, Y. Shafir, A. Bermano, D.

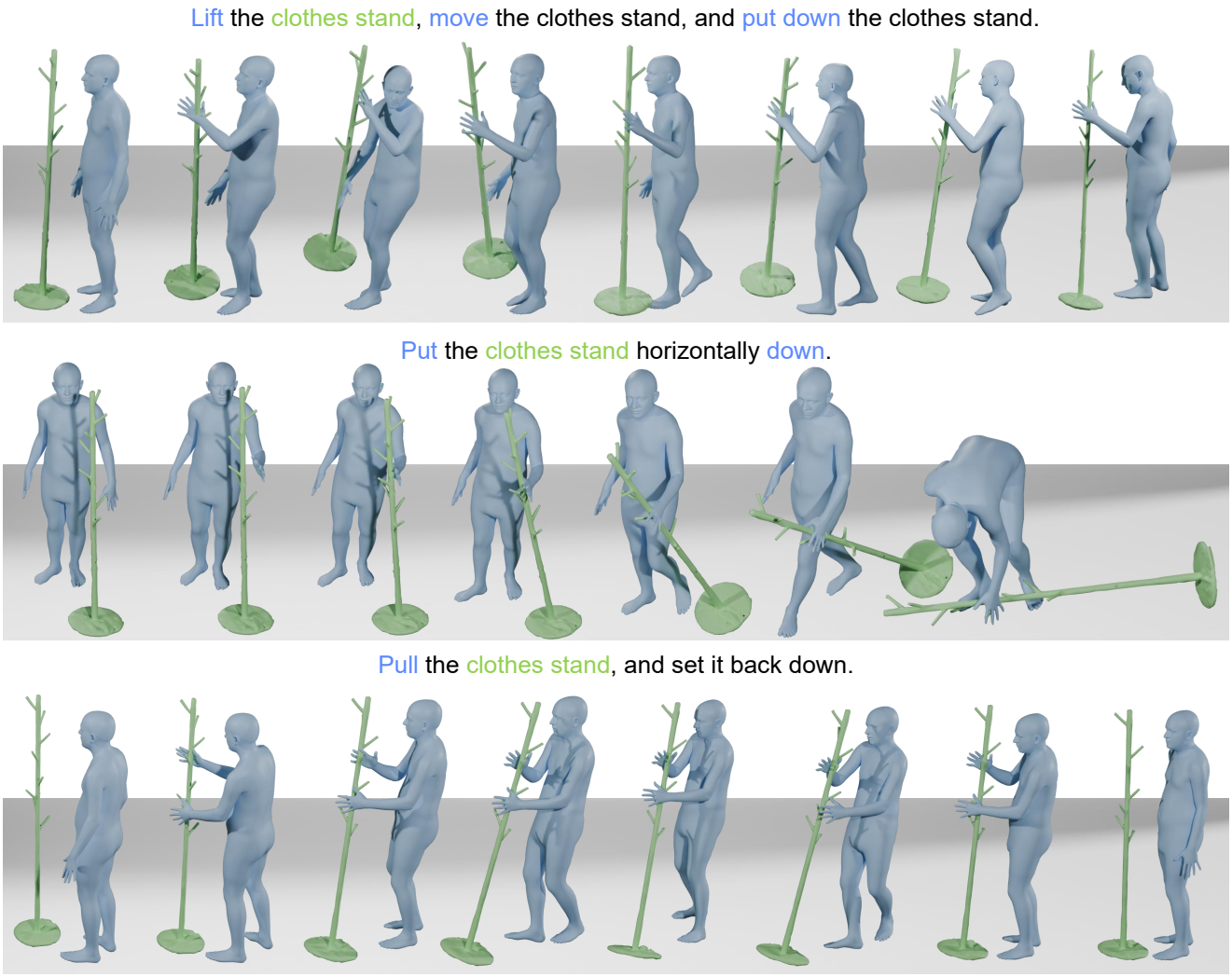

Figure 4. Our model generates diverse and semantically distinct human-object interactions for a single object.

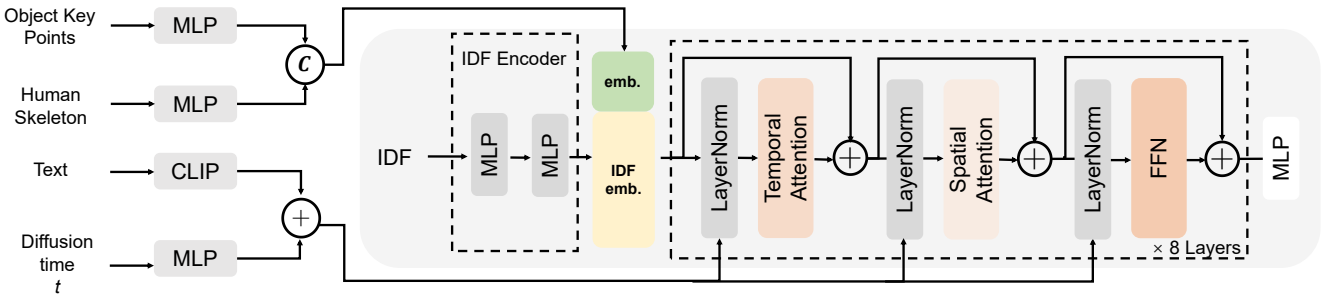

Figure 5. Overview to our Relation Model's architecture. It consists of an IDF encoder and multiple VDT layers.

Cohen-Or. Human Motion Diffusion Model. In *ICLR*, 2023. 2

[7] X. Peng, Y. Xie, Z. Wu, V. Jampani, D. Sun, H. Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In arXiv:2312.06553, 2023. 2

[8] C. Diller, A. Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024. 2