Abstract:

1. It states the problem is "… without adapting prompt semantics to the evolving pose state or explicitly leveraging structured spatio-temporal pose context." Simply not doing something isn't the right way of stating the problem. What is missed or the gaps caused by not doing something should be stated as the problem.
2. As stated at the end of the abstract, the proposed method "… enables semantic context to influence both attention-based encoding and velocity-driven motion refinement, yielding improved generalization under occlusion, irregular motion, and domain shift." Make sure you have experiments designed to evaluate these aspects: influence on attention-based encoding and velocity-driven motion refinement, occlusion, irregular motion, and domain shift.

Introduction:

3. The introduction should include the following aspects: background/applications, existing methods, statement of the open challenges (ONLY the ones to be addressed in this paper), an overview of the key ideas, and a summary of the rest of the paper.
4. The current introduction is lengthy and provides a vague statement of the problems and how the problems are addressed by the existing methods, as well as the open challenges.
5. The use of words must be careful. Apply and propose mean different levels of novelty; employ and introduce also mean different levels of novelty.

Method:

6. The Architecture Overview needs to give a complete view of the method. Figure 2 requires further explanation.
7. Figure 3 needs to be redrawn to show the idea of Laplacian pooling.
8. Each symbol used in the description must be explained.
9. The sections need reorganization. Try to follow the steps shown in Figure 2 (overall architecture).

title to be decided

Anonymous Authors[1]

Abstract

Recent work introduces semantic conditioning and prompt learning—inspired by vision–language models—to inject high-level cues such as action labels or natural-language descriptions into 3D pose estimation. However, existing prompt-based approaches treat prompts as static semantic priors without adapting prompt semantics to the evolving pose state or explicitly leveraging structured spatio-temporal pose context. We address these limitations by introducing pose-adaptive prompting for 3D HPE. Instead of fixed prompts, we dynamically modulate prompt embeddings based on the current pose hypothesis. Then, we reframe prompt learning as explicit spatio-temporal context adaptation by extracting structured pose context along two complementary dimensions: to separate global body configuration from local joint articulation, and (ii) multi-scale Gaussian temporal context pooling to disentangle short-term kinematics from long-range motion intent. The resulting spatio-temporal context generates additive prompt adaptations prior to pose–prompt interaction. Integrated into the existing 3D pose estimation architecture, pose-adaptive prompting enables semantic context to influence both attention-based encoding and velocity-driven motion refinement, yielding improved generalization under occlusion, irregular motion, and domain shift.

1. Introduction

Monocular 3D human pose estimation (3D HPE) remains fundamentally ill-posed due to depth ambiguity, self-occlusion, motion blur, and viewpoint degeneracy. Although recent spatiotemporal architectures based on graph convolutions or transformers achieve strong benchmark performance, many monocular 3D HPE models resolve ambiguity primarily through implicit feature statistics learned from training data—often collapsing to dominant dataset modes—rather than through explicit, interpretable reasoning over motion or structure (Kocabas et al., 2019; Li et al., 2022). Recent large-scale pretraining approaches mitigate this bias but still rely on latent, non-controllable representations (Zhu et al., 2023).

Motivated by the success of vision–language models(Radford et al., 2021; Zhou et al., 2022), recent work has explored semantic conditioning and prompt learning to incorporate high-level cues into 3D HPE, drawing inspiration from CLIP-style prompting where textual inputs steer visual representations via joint embedding spaces or learned prompt tokens. These approaches show that textual semantics—such as action labels, body-part descriptions, or pose-related language—can provide additional guidance and bias predictions toward semantically consistent configurations (Zheng et al., 2023; Feng et al., 2024; 2025; Xu et al., 2024; Delmas et al., 2024a;b; Zhang et al., 2024; Michel et al., 2022). Despite these advances, existing prompt-based approaches for 3D HPE exhibit several structural limitations. First, prompts are typically treated as static semantic priors: prompt tokens or textual embeddings are learned as global parameters or provided as fixed inputs, rather than being generated or adapted from the current pose estimate. This mirrors early prompt-learning strategies in vision–language models, where learned context tokens remain fixed once trained and are not explicitly conditioned on individual inputs (Zhou et al., 2022; Gao et al., 2024). In pose-specific methods, ActionPrompt uses fixed action-dependent prompts (Zheng et al., 2023), FinePOSE employs globally learned part-aware prompts reused across samples (Xu et al., 2024), and PoseLLaVA relies on externally provided instruction text that is invariant to pose uncertainty or articulation difficulty (Feng et al., 2025). As a result, prompt semantics remain largely unchanged across different pose hypotheses, occlusion patterns, or motion regimes.

Second, prompt–pose interaction is predominantly realized through indirect fusion mechanisms. FinePOSE(Xu et al., 2024) and ActionPrompt(Zheng et al., 2023) integrate prompts via prompt–pose cross-attention, where pose features attend to fixed prompt tokens to receive semantic guidance. Other approaches rely on downstream alignment or projection, mapping pose features and prompt embed-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

dings into a shared latent space using similarity or contrastive objectives (Chen et al., 2025; Zhang et al., 2023; Yang et al., 2024). While effective for injecting high-level semantics, these designs treat prompts as external guidance channels: the pose representation adapts to the prompt, but the semantic content of the prompt tokens themselves is not explicitly updated as a function of the evolving pose state or uncertainty.

Third, existing prompt learning does not explicitly leverage structured spatio-temporal pose context to drive prompt adaptation. Human pose is inherently graph-structured and temporally multi-scale, yet prompt integration is generally performed in latent feature space without conditioning prompt semantics on joint topology or temporal structure. Multimodal LLM-based approaches process pose and language as separate modalities without explicitly encoding pose-graph structure into prompt semantics (Feng et al., 2025), while alignment-based methods similarly lack mechanisms that decompose pose context into spatial or temporal components (Chen et al., 2025). Text–pose representation works such as PoseScript(Delmas et al., 2023) and PoseFix(Delmas et al., 2024a) focus on semantic description or correction, rather than using pose-graph or temporal structure to modulate prompt representations. Consequently, existing prompt-based approaches are limited in their ability to adapt semantic guidance to localized joint ambiguities or to motion-dependent uncertainty that evolves over time.

To address these limitations, we introduce pose-adaptive prompting, which reframes prompt learning for 3D HPE as explicit spatio-temporal context adaptation rather than static semantic conditioning. Instead of treating prompts as fixed priors, we dynamically adapt prompt embeddings according to the current pose state. Specifically, we extract structured spatial and temporal context from the input pose sequence and use this context to generate additive prompt adaptations that modulate the learnable prompt embeddings before text encoding. Unlike prior methods that fuse pose features with fixed prompt embeddings via cross-attention, our approach conditions the prompt embeddings themselves on spatio-temporal pose context prior to fusion, ensuring that semantic information entering the interaction stage is already tailored to the current pose hypothesis.

To enable pose-adaptive prompting, we introduce explicit pose context pooling along two complementary dimensions. As in Figure 1, spatially, we apply spectral Laplacian context pooling on the human kinematic graph to learn frequency-selective representations that separate global body configuration from local joint articulation. Temporally, we employ multi-scale Gaussian context pooling to disentangle short-term kinematics from long-range motion intent. The resulting spatio-temporal context is mapped to per-slot prompt adaptations, encouraging semantic specialization across

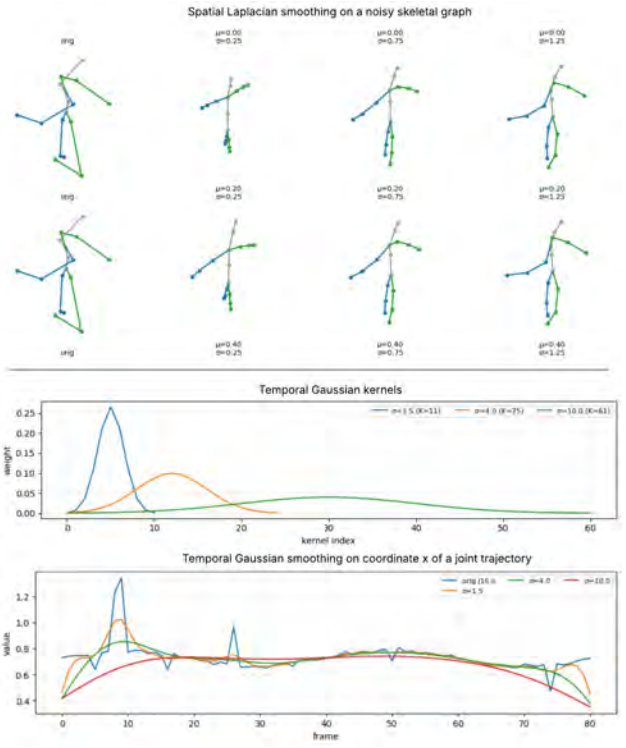

Figure 1. Caption

body parts and preventing prompt collapse. Integrated into the GAtFuN (Pham et al., 2026) architecture—which combines spatio-temporal multi-head self-attention with a velocity-based graph refinement branch—this design allows semantic context to modulate both feature encoding and motion evolution. By conditioning the shared latent representations used by both branches, pose-adaptive prompts enable the model to override dataset-specific correlations when visual evidence is weak, leading to improved robustness and generalization under occlusion, irregular motion, and domain shift.

Our contributions are summarized as follows:

- We introduce pose-adaptive prompting for 3D HPE, transforming static prompts into pose-aware semantic operators.
- We propose spectral Laplacian spatial context pooling on the pose graph to separate global configuration from local articulation.
- We introduce multi-scale Gaussian temporal context pooling to disentangle short-term kinematics from long-range motion intent.
- We integrate structured spatio-temporal prompt adaptation into GAtFuN, enabling semantic context to modulate both attention-based encoding and velocity-driven motion refinement.

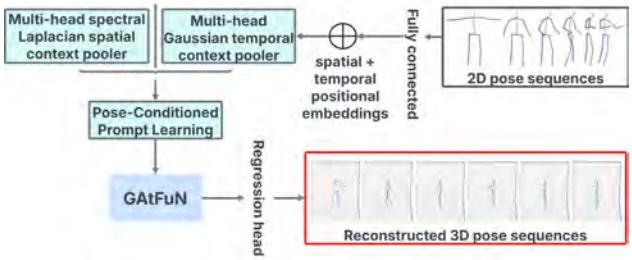

Figure 2. Overall Architecture

2. Related Work

3. Method

3.1. Architecture Overview

Given a monocular 2D pose sequence $x \in \mathbb{R}^{B\times T\times N\times 2}$ (batch B, frames T, joints N), we predict 3D joint coordinates $Y \in \mathbb{R}^{B\times T\times N\times 3}$. Our model extends GAtFuN by introducing (i) spectral Laplacian spatial context pooling, (ii) multi-scale Gaussian temporal context pooling, and (iii) pose-adaptive prompting that conditions prompt embeddings on pooled pose context. All other components (spatio-temporal MHSA, graph attention motion refinement, and adaptive motion fusion) follow the GAtFuN(Pham et al., 2026) design, as in Figure 2.

We first embed the 2D joints into a D-dimensional feature space using a learnable projection:

$$X = \phi(x) \in \mathbb{R}^{B\times T\times N\times D}. \tag{1}$$

To preserve spatio-temporal structure, we add learnable positional embeddings for joints and frames:

$$X \leftarrow X + e^{S}_{\text{pos}} + e^{T}_{\text{pos}}, \tag{2}$$

where $e^{S}_{\text{pos}} \in \mathbb{R}^{1\times 1\times N\times D}$ and $e^{T}_{\text{pos}} \in \mathbb{R}^{1\times T\times 1\times D}$.

3.2. Spectral Laplacian Spatial Context Pooling

To extract structured spatial context that respects the kinematic topology, we operate on the human pose graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with adjacency A and Laplacian L. Let L be the combinatorial Laplacian with eigendecomposition:

$$L = U\Lambda U^{\top}, \tag{3}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$ are graph frequencies.

Given pose features $X \in \mathbb{R}^{B\times T\times N\times D}$, we project to the spectral domain:

$$\tilde{X} = U^{\top}X. \tag{4}$$

We define H_s spatial pooling heads, each learning a frequency-selective Gaussian mask:

$$w_h(\lambda) = \exp\left(-\frac{(\lambda-\mu_h)^2}{2\sigma_h^2}\right), \quad h = 1, \ldots, H_s, \tag{5}$$

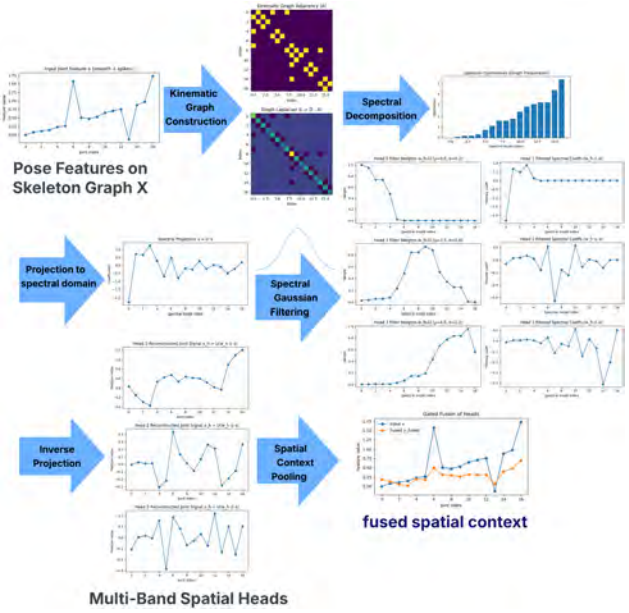

Figure 3. Laplacian Pooling

where μ_h are learnable center frequency of the band and σ_h controls the bandwidth. By learning multiple heads with different (μ_h, σ_h), the model decomposes pose features into complementary spatial scales. In the spectral domain of the pose graph, low Laplacian eigenvalues encode smooth, global body structure, while high eigenvalues capture localized joint articulation. Accordingly, the filter center μ_h selects the spatial frequency band emphasized by each head—small μ_h focusing on global configuration and larger μ_h on local articulation—while the bandwidth σ_h controls the scale of aggregation, with smaller values isolating specific frequencies and larger values integrating broader spatial context.

Each head reconstructs filtered features and mixes them with a residual gate:

$$X^{(h)} = X + \alpha_h\Big(U\big(w_h \odot \tilde{X}\big) - X\Big), \tag{6}$$

where α_h is learnable and $\odot$ applies w_h across the spectral dimension. We then pool across joints to obtain spatial context vectors:

$$v_s^{(h)} = \text{Pool}_{t,n}\big(\rho(X^{(h)})\big) \in \mathbb{R}^{B\times D_c}, \tag{7}$$

where $\rho(\cdot)$ is a linear projection to a context dimension D_c.

3.3. Gaussian Temporal Context Pooling

To provide explicit temporal scale separation, we introduce Gaussian temporal pooling heads (Figure 4. For each head $h = 1, \ldots, H_t$, we use a learnable Gaussian kernel over

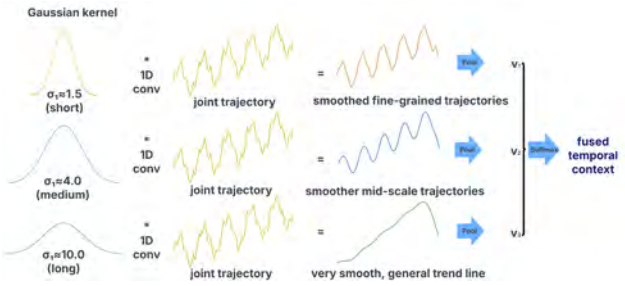

Figure 4. Gaussian Pooling

temporal offsets:

$$g_h(\tau) = \exp\left(-\frac{(\tau - \mu_h)^2}{2\sigma_h^2}\right), \tag{8}$$

where parameter σ_h controls the receptive field of each temporal pooling head by determining how rapidly the Gaussian kernel decays over time (and μ_h optionally shifts the center). A smaller σ_h yields a narrow kernel, capturing short-term kinematics and rapid motion changes. In contrast, a larger σ_h produces a wider kernel that aggregates information across a longer temporal window, encoding long-range motion intent and coarse temporal trends. Learning multiple heads with different σ_h values enables explicit separation of short- and long-term temporal context, improving robustness to noise and irregular motion.

We apply depthwise temporal filtering:

$$X_t^{(h)} = \sum_{\tau} g_h(\tau)\, X_{t-\tau}. \tag{9}$$

Finally, we pool to obtain temporal context vectors:

$$v_t^{(h)} = \mathrm{Pool}_{t,n}\big(\eta(X^{(h)})\big) \in \mathbb{R}^{B\times D_c}, \tag{10}$$

where $\eta(\cdot)$ is a linear projection.

3.4. Context Gating and Slot-Wise Pose Descriptors

We aggregate multi-head spatial and temporal contexts using learned gates:

$$v_s = \sum_h \beta_h^s v_s^{(h)}, \quad v_t = \sum_h \beta_h^t v_t^{(h)}, \tag{11}$$

$$\beta^s = \mathrm{Softmax}(a^s),\ \beta^t = \mathrm{Softmax}(a^t), \tag{12}$$

and form a joint context vector $z = [v_s; v_t]$. We then produce slot-specific descriptors for each semantic slot k using lightweight MLPs:

$$z_k = \mathrm{MLP}_k(z). \tag{13}$$

3.5. Pose-Adaptive Prompting

We introduces a set of learnable prompt embeddings $P = \{p_k\}_{k=1}^K$, where each prompt corresponds to a semantic slot (e.g., person, speed, action, body, arm, leg). Each prompt has token length L and embedding dimension D: $p_k \in \mathbb{R}^{L\times D}$. Following CLIP-style prompting, each p_k is formed by concatenating t frozen token embeddings from a pretrained text encoder with a learnable context. Using slot descriptors $\{z_k\}$, we compute prompt adaptations Δp_k and form adapted prompts p_k. Concretely, for each semantic slot k, we predict an additive adaptation $\Delta p_k \in \mathbb{R}^{(L-t)\times D}$ from pooled spatio-temporal pose context z_k:

$$p_k = \mathrm{Concat}(\tilde{p}_k,\, r_k + \Delta p_k) \tag{14a}$$

$$\Delta p_k = f_k(z_k), \quad r_k \in \mathbb{R}^{(L-t)\times D}, \tag{14b}$$

$$\tilde{p}_k = \mathrm{CLIP}(t_k)[:t] \in \mathbb{R}^{t\times D}. \tag{15}$$

where $f_k(\cdot)$ is a slot-specific MLP. The prompt tensor is broadcast across the batch: $P \in \mathbb{R}^{B\times K\times L\times D}$ and fed to the frozen CLIP text transformer (optionally followed by a lightweight trainable encoder), yielding pooled prompt embeddings $P' \in \mathbb{R}^{B\times D}$, which can be broadcast and added to pose features:

$$X \leftarrow X + \mathrm{Broadcast}(P'). \tag{16}$$

The pose-conditioned prompt embeddings P' subsequently participate in pose–prompt fusion to preserve the benefits of cross-attention while ensuring the semantic representation itself is spatial-temporal state-aware rather than fixed.

3.6. Pose Modeling and Motion Refinement

The rest of the model follows the GAtFuN (Pham et al., 2026) architecture, including dual-branch motion fusion between a pose modeling stream based on spatio-temporal MHSA and a motion refinement stream based on spatio-temporal graph attention applied to joint velocities. Our pose-adaptive prompt learning is injected early into the shared latent representation via pose–prompt cross-attention, allowing semantic context to modulate both spatial attention in the pose modeling branch and motion propagation in the refinement branch. More details are provided in the supplementary material. After iterative spatio-temporal refinement and adaptive motion fusion, we regress the final 3D joint positions Y.

4. Results

4.1. Implementation

4.2. Ablation Study

4.3. Computation Cost

4.4. Qualitative Analysis

5. Conclusion

References

Chen, Y., Xie, X., and Li, F. Vision-language model guided pose knowledge mining for human pose estimation. *Journal of Computational Design and Engineering*, 12(9): 32–45, 2025.

Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., and Rogez, G. Posefix: correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15018–15028, 2023.

Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., and Rogez, G. Posescript: Linking 3d human poses and natural language. *IEEE transactions on pattern analysis and machine intelligence*, 2024a.

Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., and Rogez, G. Poseembroider: Towards a 3d, visual, semantic-aware human pose representation. In *European Conference on Computer Vision*, pp. 55–73. Springer, 2024b.

Feng, D., Guo, P., Peng, E., Zhu, M., Yu, W., and Wang, P. Posellava: Pose centric multimodal llm for fine-grained 3d pose manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2951–2959, 2025.

Feng, Y., Lin, J., Dwivedi, S. K., Sun, Y., Patel, P., and Black, M. J. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2093–2103, 2024.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Kocabas, M., Karagoz, S., and Akbas, E. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1077–1086, 2019.

Li, W., Liu, H., Tang, H., Wang, P., and Van Gool, L. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13147–13156, 2022.

Michel, O., Bar-On, R., Liu, R., Benaim, S., and Hanocka, R. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13492–13502, 2022.

Pham, Y., Yuan, X., and Zhuang, C. Motion-aware graph fusion network for 3d human pose estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2026.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Xu, J., Guo, Y., and Peng, Y. Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 561–570, 2024.

Yang, J., Zeng, A., Zhang, R., and Zhang, L. X-pose: Detecting any keypoints. In *European Conference on Computer Vision*, pp. 249–268. Springer, 2024.

Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., and Liu, Z. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.

Zhang, X., Wang, W., Chen, Z., Xu, Y., Zhang, J., and Tao, D. Clamp: Prompt-based contrastive learning for connecting language and animal pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23272–23281, 2023.

Zheng, H., Li, H., Shi, B., Dai, W., Wang, B., Sun, Y., Guo, M., and Xiong, H. Actionprompt: Action-guided 3d human pose estimation with text and pose prompting. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2657–2662. IEEE, 2023.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022.

Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., and Wang, Y. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15085–15099, 2023.