Surgeon Style Fingerprinting and Privacy Risk Quantification via Discrete Diffusion Models in a Vision-Language-Action Framework

Huixin Zhan
Cedars-Sinai Medical Center
700 N. San Vicente Blvd., West Hollywood, CA 90069
Huixin.Zhan@cshs.org

Jason H. Moore
Cedars-Sinai Medical Center
700 N. San Vicente Blvd., West Hollywood, CA 90069
jason.moore@csmc.edu

Abstract

Surgeons exhibit distinct operating styles due to differences in training, experience, and motor behavior — yet current AI systems often ignore this personalization signal. We propose a novel approach to model fine-grained, surgeon-specific fingerprinting in robotic surgery using a discrete diffusion framework integrated with a vision-language-action (VLA) pipeline. Our method formulates gesture prediction as a structured sequence denoising task, conditioned on multimodal inputs including endoscopic video, surgical intent language, and a privacy-aware embedding of surgeon identity and skill. Personalized surgeon fingerprinting is encoded through natural language prompts using third-party language models, allowing the model to retain individual behavioral style without exposing explicit identity. We evaluate our method on the JIGSAWS dataset and demonstrate that it accurately reconstructs gesture sequences while learning meaningful motion fingerprints unique to each surgeon. To quantify the privacy implications of personalization, we perform membership inference attacks and find that more expressive embeddings improve task performance but simultaneously increase susceptibility to identity leakage. These findings demonstrate that while personalized embeddings improve performance, they also increase vulnerability to identity leakage, revealing the importance of balancing personalization with privacy risk in surgical modeling.

1. Introduction

Personalized modeling of surgical behavior has the potential to improve intraoperative decision support, skill assessment, and robot-assisted training. In particular, fine-grained prediction of surgical gestures—low-level action primitives such as "grasp needle" or "position needle"—can reveal patterns in technique that vary across surgeons, tasks, and experience levels. Recent work [17] has leveraged vision and language models for surgical understanding, but most existing systems remain surgeon-agnostic, failing to capture individual variation critical for personalization and performance benchmarking.

Modeling this variation, however, raises unique challenges: How can models account for stylistic behavioral differences between surgeons? Can we learn surgeon-specific representations without exposing identity information? And how can we build robust predictive systems that generalize across users while still capturing meaningful intra-surgeon variability?

To address these challenges, we propose a diffusion-based vision-language-action (VLA) framework for personalized gesture sequence modeling. We formulate gesture prediction as a denoising problem over discrete tokens and condition the reverse process on multimodal inputs: vision from surgical video, language from task-level prompts, and a personalized embedding representing each surgeon's behavioral fingerprint. Importantly, we explore privacy-preserving embedding strategies by using third-party large language models (LLMs) to encode natural language prompts that include both surgeon identity and clinically validated skill scores (e.g., Global Rating Scale, GRS [10]). While these identifiers are visible to the third-party LLM during prompt encoding, they are not directly exposed to the downstream gesture prediction model.

We evaluate our method on the JIGSAWS dataset [9] and show that surgeon-specific embeddings improve structured

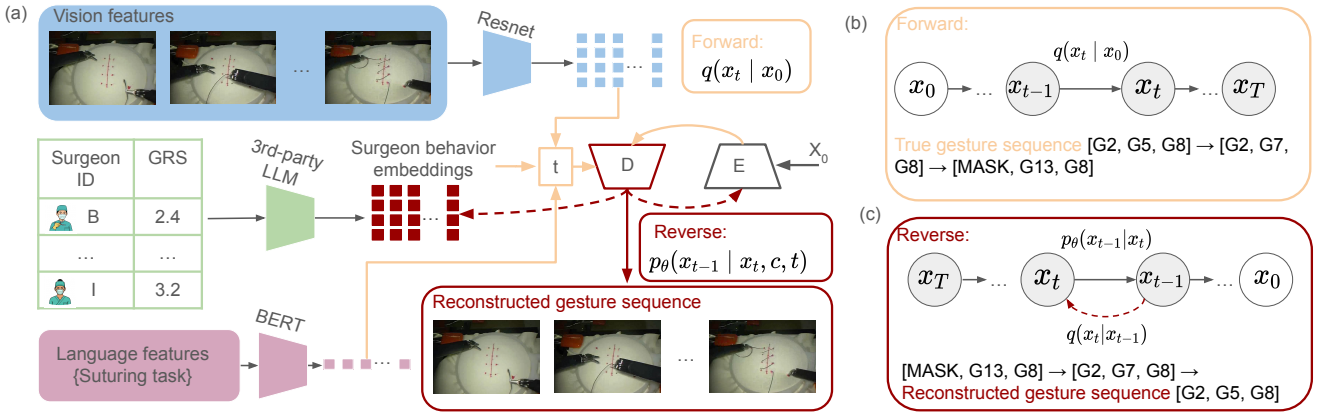


Figure 1. **Personalized gesture sequence prediction using diffusion models.** (a) Vision features are extracted from surgical videos using ResNet, while task-level semantic cues are obtained from the language prompt (e.g., 'suturing task") using BERT. Surgeon-specific behavior embeddings are generated by encoding the surgeon ID and GRS using a third-party large language model (LLM). These visual, linguistic, and behavioral signals are fused to condition the denoising model. (b) In the forward process, categorical noise is gradually applied to the ground truth gesture sequence x_0 via $q(x_t \mid x_0)$, transforming discrete gestures (e.g., [G2, G5, G8]) into corrupted sequences (e.g., [MASK, G13, G8]). (c) The reverse process learns to reconstruct the original sequence from noise by predicting $p_\theta(x_{t-1} \mid x_t, c, t)$, where c denotes multimodal context. This framework enables personalized and condition-aware gesture generation.

Table 1. Performance comparison of gesture prediction models under different surgeon identity representations. The **third-party LLM (ID + GRS)** setting encodes both the surgeon ID and averaged objective GRS using an LLM. The **third-party LLM (ID only)** setting encodes only the surgeon ID using the same language model, while the **non-private baseline** uses standard learnable embeddings. Despite similar predictive performance, the personalized LLM-based embeddings introduce higher re-identifiability risk, demonstrating a trade-off between personalization and privacy.

Task	Surgeon Representation	Top-1 Accuracy	Top-5 Accuracy	Weighted F1-Score
Suturing	Third-party LLM (ID + GRS)	**0.8389**↑	0.9984↓	**0.8447**↑
	Third-party LLM (ID only)	0.8327	**0.9988**	0.8324
	Non-private baseline	0.8240	0.9968	0.8237

sequence denoising performance. To assess potential privacy risks introduced by personalization, we conduct membership inference attacks targeting the learned embeddings. Our results reveal that while LLM+GRS embeddings enhance personalization, they also increase susceptibility to re-identification, highlighting a trade-off between behavioral fidelity and privacy. This work underscores the need for careful evaluation of privacy leakage in personalized surgical AI systems.

Our key contributions are: (1) a diffusion-based formulation for personalized discrete gesture prediction in robotic surgery, (2) a language-model-based embedding scheme that encodes surgeon identity and skill cues via natural language prompts via third-party large language models, (3) a quantitative privacy analysis using membership inference attacks to evaluate identity leakage, and (4) comprehensive experiments regarding both gesture prediction accuracy and stylistic coherence on the JIGSAWS dataset demonstrating both improved personalization and the trade-offs between performance and privacy.

2. Method

We formulate surgeon gesture prediction as a discrete denoising diffusion process. Each gesture sequence $x_0 \in \mathcal{G}^T$, where $\mathcal{G} = \{1, \ldots, K\}$ is the gesture vocabulary, is corrupted over time via a multinomial noise process. The goal is to learn a conditional model that reconstructs x_0 from a noisy sequence $x_t \sim q(x_t \mid x_0)$, given multimodal context.

2.1. Forward Process: Discrete Multinomial Corruption

We define a time-indexed transition matrix $Q_t \in \mathbb{R}^{K \times K}$ that governs the noise schedule. Each diagonal entry $Q_t(i, i)$ decays linearly as:

$$Q_t(i,i) = 1 - \frac{t+1}{T}, \quad Q_t(i, j \neq i) = \frac{1 - Q_t(i,i)}{K-1}$$

This process flips gesture tokens at increasing rates as t progresses, simulating categorical corruption.

2.2. Reverse Process: Personalized Gesture Denoising

We train a transformer-based model [18] to approximate the reverse conditional distribution $p_\theta(x_{t-1} \mid x_t, c, t)$, where the conditioning context c includes:

- Vision features extracted from endoscopic video using a ResNet encoder [11],
- Language features derived from a BERT encoding [3] of the task prompt (e.g., "suturing task"),
- Surgeon embedding $\mathbf{s}_i \in \mathbb{R}^d$ that encodes stylistic behavioral information,
- Timestep embedding t. These components are fused and injected into the denoising model at each timestep.

2.3. Surgeon Embeddings and Personalization

To model inter-surgeon variation, each surgeon is assigned an embedding vector $\mathbf{s}_i$. These embeddings are either learned directly through a trainable embedding layer or derived from natural language prompts—such as "Surgeon ID: i, GRS: 3.4"—using a frozen third-party language model (e.g., Sentence-BERT [14] or MiniLM [19]). By conditioning on these embeddings, the model learns to denoise gesture sequences in a surgeon-specific manner, enabling stylistic profiling and personalized gesture generation. After training, the resulting embeddings can be used for downstream applications such as clustering, retrieval, or surgeon re-identification.

2.4. Training Objective

We sample a timestep $t \sim \mathcal{U}(0, T)$ uniformly and compute a noisy gesture sequence $x_t \sim q(x_t \mid x_0)$. The model is trained to predict the original gesture x_0 using a cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = \mathbb{E}_{x_0,t,x_t} \left[-\log p_\theta(x_0 \mid x_t, c, t) \right]$$

The loss is computed over all gesture positions, and gradients are backpropagated through the entire diffusion trajectory.

2.5. Inference

At test time, we formulate gesture prediction as a generative inference task over discrete sequences. Starting from a fully corrupted gesture sequence $x_T \sim q(x_T \mid x_0)$, which represents near-uniform categorical noise, the model applies the learned reverse process iteratively from $t = T$ to 0. This produces a denoised sequence $\hat{x}_0$ that reconstructs the likely gesture trajectory under the current visual, linguistic, and behavioral context.

This formulation departs from traditional gesture recognition methods, which typically rely on direct sequence classification or frame-wise decoding. Instead, we cast inference as a structured sequence generation problem conditioned on multimodal and surgeon-specific embeddings. The process is inherently probabilistic and allows for uncertainty modeling and sample diversity, enabling the system to generate plausible personalized behaviors.

We evaluate inference performance along two key axes: (i) *gesture prediction accuracy*, which quantifies the model's ability to reconstruct the ground truth gesture sequence, and (ii) *stylistic coherence*, which assesses how closely the generated trajectory reflects the unique behavioral fingerprint of the target surgeon, as captured by their personalized embedding.

2.6. Privacy Risk Analysis via Membership Inference

To quantify the privacy risk associated with personalized embeddings, we conduct a membership inference attack [16]. Given the learned surgeon embeddings $\{\mathbf{s}_i\}$, we simulate an adversary that trains a binary classifier (e.g., XGBoost [2]) to distinguish between in-training (member) embeddings and synthetic out-of-distribution (non-member) ones. High classifier performance—measured by AUC, accuracy, and F1-score—indicates a greater risk of identity leakage.

We compare privacy vulnerability across three embedding strategies. As reported in Section 3, the LLM (ID + GRS) embedding achieves the best gesture prediction performance, but also exhibits the highest susceptibility to membership inference attacks (AUC = 1.000). This reveals a key privacy-performance trade-off: richer personalization yields more discriminative embeddings, which are easier to link back to training data. In contrast, the non-private baseline produces less structured embeddings that offer weaker personalization but lower risk under attack.

These findings emphasize the need for explicit privacy evaluation when deploying personalized models in clinical settings, especially when embedding behavioral signals like skill.

3. Experiments

To evaluate the effectiveness of our proposed framework, we design a series of experiments aimed at answering the following key scientific questions:

- **Q1:** How do different surgeon representation strategies (e.g., learnable ID embeddings, third-party LLMs (ID only), and third-party LLMs (ID + GRS)) impact personalized gesture sequence modeling?
- **Q2:** Can natural language embeddings that incorporate clinically validated metrics (e.g., GRS) enable personalization?
- **Q3:** Do the learned surgeon embeddings capture meaningful and structured variation in skill or behavior across different individuals?
- **Q4:** To what extent do different personalization methods expose the model to privacy leakage under adversarial

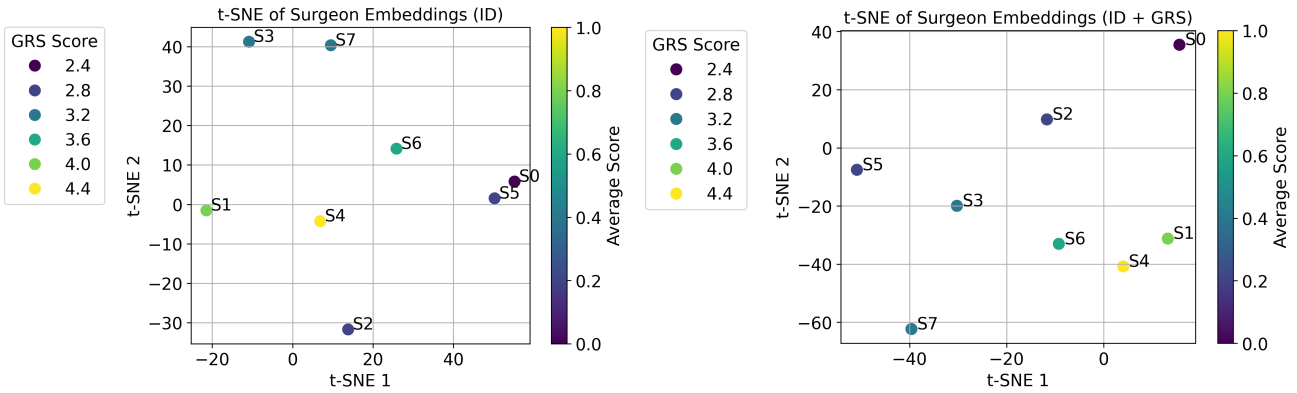


Figure 2. t-SNE of surgeon embeddings. (**Left**) Trained using surgeon ID only. (**Right**) Trained with both ID and GRS supervision. Colors indicate average skill score (GRS).

Table 2. **Privacy evaluation via membership inference attack.** We evaluate privacy leakage under three surgeon embedding strategies by training a binary classifier to determine whether an embedding came from the training set. Higher AUC and precision indicate greater vulnerability to identity leakage. The **third-party LLM (ID + GRS)** strategy enables stronger personalization, but at the cost of higher re-identifiability risk.

Task	Surgeon Representation	Accuracy	Precision	Recall	F1 Score	AUC
Suturing	Third-party LLM (ID + GRS)	**0.997**↑	**0.998**↑	**1.000**↑	**0.988**↑	**1.000**↑
	Third-party LLM (ID only)	0.889	0.500	1.000	0.667	1.000
	Non-private baseline	0.889	0.000	0.000	0.000	0.469

conditions such as membership inference attacks?

We design ablation experiments on surgeon embedding strategies, heatmap visualizations, GRS-informed fingerprinting analyses, and privacy evaluations to systematically investigate each of these research questions. The model is trained for 20 epochs using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 32. Code is available at: https://github.com/huixin-zhan-ai/Surgeon_style_fingerprinting. The results and insights are presented in the following sections.

3.1. Ablation on Surgeon Embedding Strategies.

To assess the impact of surgeon-specific conditioning on gesture prediction, we compare three distinct embedding strategies:

1. **Non-private baseline:** We use a learnable embedding layer that maps each surgeon ID to a unique vector. This approach directly exposes identity and serves as an upper-bound reference in terms of capacity and specificity.
2. **Third-party LLM (ID only):** Surgeon IDs are formatted into short natural language prompts (e.g., "Surgeon ID: 3") and encoded using a publicly available sentence-level language model (e.g., MiniLM). This embedding is kept fixed and projected to the model's hidden dimension via a learnable linear layer.
3. **Third-party LLM (ID + GRS):** To incorporate behavioral priors, we construct natural language prompts that include both the surgeon ID and their averaged GRS, e.g., "Surgeon ID: 3, average skill score: 3.75". These prompts are encoded using a frozen third-party language model (e.g., Sentence-BERT), and projected into the model's hidden space. This strategy captures personalized style in a clinically grounded manner, while avoiding direct integration of raw identifiers into the gesture prediction model.

To assess the impact of surgeon representation strategies on model performance, we compare three variants of our framework: (i) the *third-party LLM (ID + GRS)* method, which encodes both surgeon identity and averaged skill using natural language prompts and a frozen SentenceTransformer; (ii) the *third-party LLM (ID only)* variant, which uses surgeon ID prompts without skill descriptors; and (iii) a *non-private baseline*, where ID-specific embeddings are learned via end-to-end learning.

As shown in Table 1, the third-party LLM (ID + GRS) method achieves the highest Top-1 accuracy (83.89%) and weighted F1-score (0.8447), with near-perfect Top-5 accuracy (99.84%), outperforming the ID-only and non-private baselines. These results reflect the benefit of incorporating

clinically grounded behavioral cues like GRS into surgeon representations.

However, our privacy analysis reveals that this gain in personalization comes at the cost of increased identity leakage under membership inference attacks. This finding reveals a core tension in surgical modeling: richer behavioral embeddings improve prediction, but also make models more susceptible to privacy breaches. Therefore, while third-party LLM embeddings with GRS offer strong personalization capabilities, they also demand careful consideration of privacy risk.

3.2. Qualitative Analysis via Gesture Distribution Heatmaps

To further understand the behavioral distinctions captured by each surgeon embedding strategy, we visualize the predicted gesture distributions across surgeons using heatmaps. Each heatmap cell represents the frequency of a predicted gesture token (e.g., G1–G15) for a specific surgeon, aggregated across all test sequences.

Figure 3 compares the predicted gesture distributions under three settings: (a) the non-private baseline using learnable ID embeddings, (b) the third-party LLM (ID only) condition, and (c) the third-party LLM (ID + GRS) embedding. These visualizations qualitatively reinforce our quantitative results: the third-party LLM (ID + GRS) variant achieves personalization effects comparable to the non-private baseline—despite not relying on direct access to learned identity embeddings.

3.3. Surgeon Fingerprinting and GRS Association

To explore whether our model captures meaningful variations in surgical behavior, we visualize surgeon embeddings using t-SNE. Each point in Figure 2 represents an individual surgeon, and colors indicate their average GRS performance score.

The left panel shows embeddings learned when the model is only provided surgeon IDs. In this setting, the embedding space primarily reflects identity but does not exhibit clear separation by skill level. In contrast, the right panel shows embeddings learned with both surgeon ID and GRS supervision. Here, surgeons with similar skill levels are more coherently grouped, and the embeddings demonstrate smoother gradients along the GRS axis.

This suggests that our personalized representation — guided by visual, linguistic, and behavioral cues — can encode latent skill information in an unsupervised or weakly-supervised fashion. Embedding models that integrate clinical supervision like GRS may enhance both interpretability and downstream personalization.

3.4. Privacy Evaluation via Membership Inference

We assess privacy leakage from different surgeon embedding strategies using a membership inference attack, in which an adversary attempts to determine whether a given embedding originated from the model's training data. Results are shown in Table 2.

The **third-party LLM (ID + GRS)** representation is the most expressive, but also the most vulnerable. It enables highly personalized gesture modeling, but the resulting embeddings are easily distinguishable from synthetic (non-member) embeddings, with an AUC of 1.000, precision of 0.998, and F1-score of 0.988. This indicates that the embedding space reveals strong identity-linked signals, raising concerns about privacy leakage.

The **third-party LLM (ID only)** strategy also exhibits high recall and AUC, but its lower precision (0.500) and F1-score (0.667) suggest a less precise separation between members and non-members. This reflects a weaker entanglement of identity information compared to the GRS-conditioned variant.

The **non-private baseline** shows low attack performance, with zero precision and recall and low AUC (0.469). While this may seem desirable from a privacy standpoint, it more likely reflects noisy or unstructured embeddings rather than true robustness.

Overall, these results highlight a trade-off: more expressive, personalized embeddings improve task performance but increase susceptibility to identity inference attacks. This underscores the importance of quantifying privacy risks when designing surgeon-specific models in surgical AI.

4. Discussion

4.1. Current Vision-Language-Action (VLA) Systems

Recent advances in surgical AI have explored multimodal learning through vision-language-action (VLA) frameworks [4, 20]. These models integrate visual and textual signals to understand surgical phases and gestures. However, most existing systems adopt a one-size-fits-all architecture that models behavior agnostic to individual variation. Works like [5, 13] focus on phase prediction or gesture segmentation, but assume shared behavior across surgeons. As a result, they may generalize poorly in contexts requiring personalization, such as training simulators or adaptive feedback systems.

4.2. Surgeon Style Fingerprinting

Few prior studies explicitly model stylistic differences across surgeons. Earlier efforts such as [8, 21] attempt to correlate motion patterns with skill, but do not offer a generative formulation for personalized prediction. Our work builds on the idea of behavioral fingerprinting by

embedding per-surgeon style representations and generating gesture sequences via discrete diffusion. This complements prior gesture forecasting approaches [12, 15] with a focus on personalized generation. The ability to encode style in a structured embedding opens new possibilities for re-identification, clustering, and personalized curriculum learning.

4.3. Privacy Quantification via Membership Inference

While privacy has received growing attention in medical ML [6, 16], few works examine privacy risks in behavioral embeddings. Most privacy studies focus on image reconstruction or record linkage [7], not personalized sequence generation. Our membership inference analysis reveals a tension between expressiveness and privacy: richer embeddings (e.g., LLM+GRS) perform better but are more susceptible to identity leakage. This is aligned with recent work on privacy-utility trade-offs in embedding models [1]. Future extensions could explore differential privacy or adversarial training to mitigate this risk.

References

[1] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, 2021. 1

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 3

[4] Di Ding, Tianliang Yao, Rong Luo, and Xusen Sun. Visual question answering in robotic surgery: A comprehensive review. *IEEE Access*, 2025. 5

[5] Robert DiPietro, Colin Lea, Anand Malpani, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision (ECCV)*, pages 36–52. Springer, 2019. 5

[6] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. 1

[7] Matthew Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1322–1333, 2015. 1

[8] Ilja Funke, Simon T Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 101–108. Springer, 2019. 5

[9] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, page 3, 2014. 1

[10] Jean D Gray. Global rating scales in residency education. *Academic Medicine*, 71(1):S55–63, 1996. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[12] Mohammadmahdi Honarmand, Muhammad Abdullah Jamal, and Omid Mohareri. Vidlpro: A video-language pretraining framework for robotic and laparoscopic surgery. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024. 1

[13] Le Ma, Hangyeol Kang, Nadia Magnenat-Thalmann, and Katarzyna Wac. Transsg: A spatial-temporal transformer for surgical gesture recognition. In *Computer Graphics International Conference*, pages 151–165. Springer, 2024. 5

[14] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[15] Chang Shi, Yi Zheng, and Ann Majewicz Fey. Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8017–8024. IEEE, 2022. 1

[16] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 3, 1

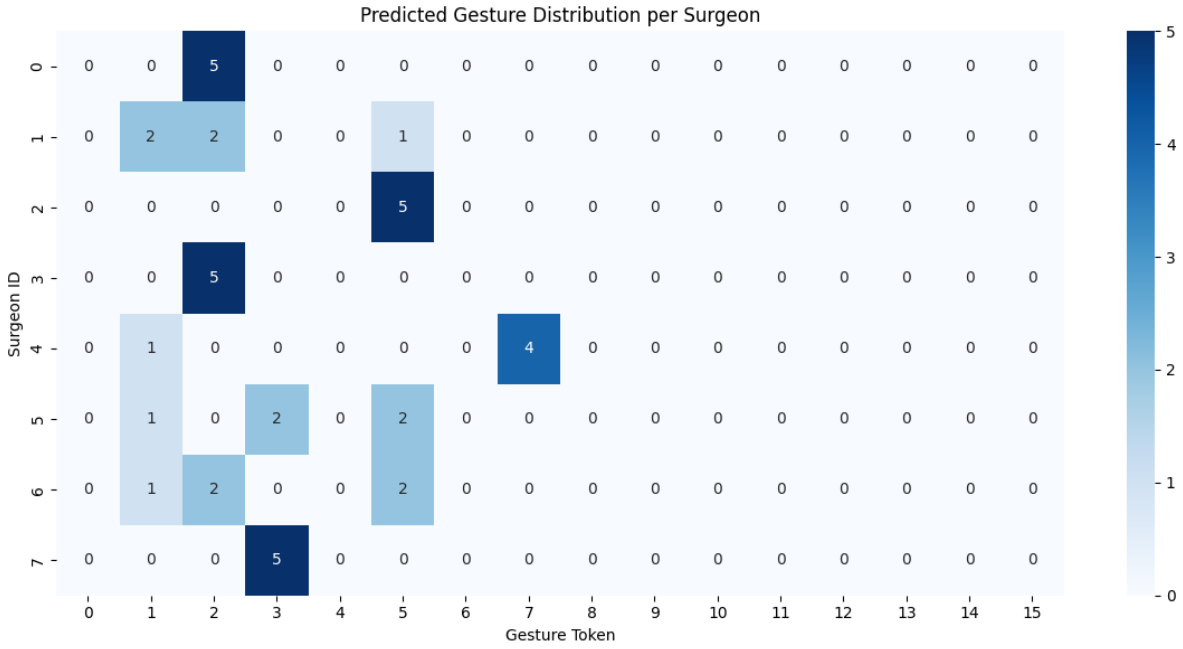
[17] Beatrice van Amsterdam, Matthew J Clarkson, and Danail Stoyanov. Gesture recognition in robotic surgery: a review. *IEEE Transactions on Biomedical Engineering*, 68(6), 2021. 1

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

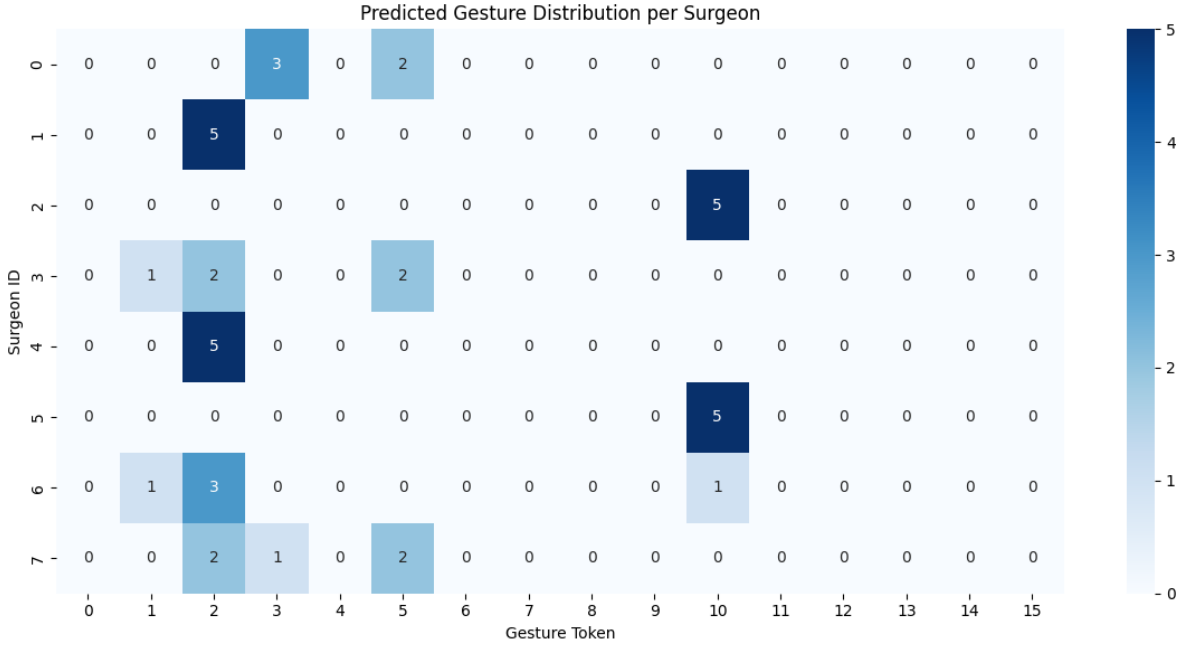
[19] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020. 3

[20] Sadra Zargarzadeh, Maryam Mirzaei, Yafei Ou, and Mahdi Tavakoli. From decision to action in surgical autonomy: Multi-modal large language models for robot-assisted blood suction. *IEEE Robotics and Automation Letters*, 2025. 5

[21] Aneeq Zia and Irfan Essa. Automated assessment of surgical skills using frequency analysis. *IEEE Transactions on Biomedical Engineering*, 65(9):2155–2166, 2018. 5

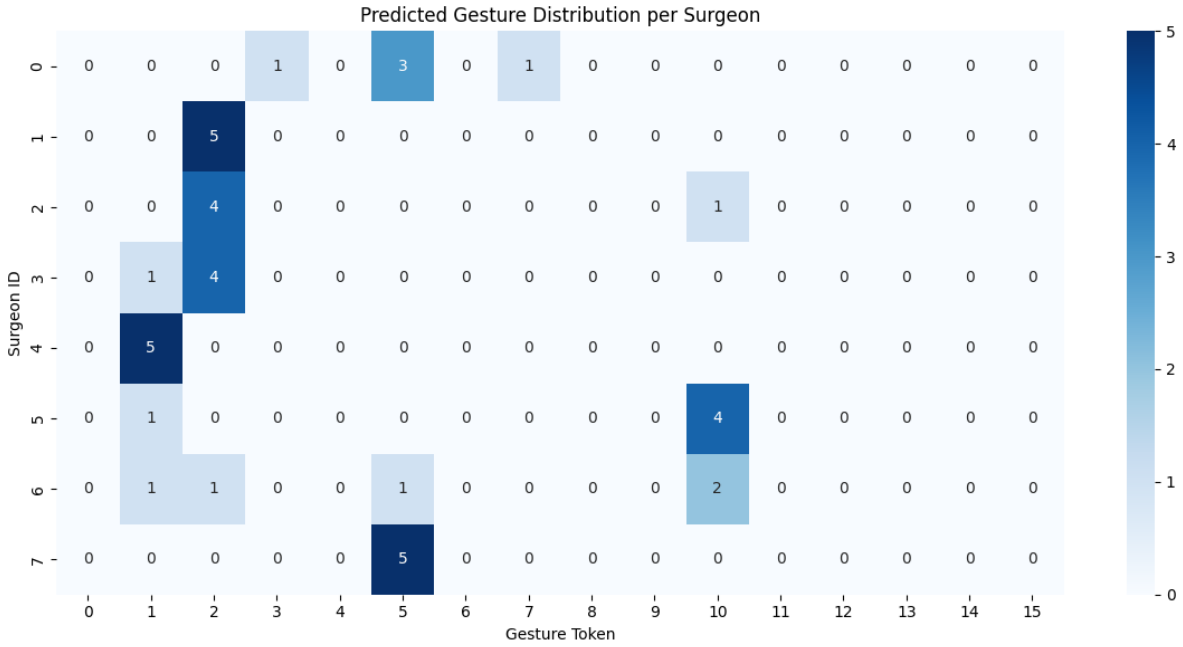


(a) Non-private baseline



(b) Third-party LLM (ID only)

Figure 3. Predicted gesture token distributions across surgeons. Rows represent surgeon IDs and columns denote gesture tokens.



(c) Third-party LLM (ID + GRS)

Figure 3. (continued) Panel (c) shows structured, personalized distributions under ID + GRS embedding.