Human-Intervention Segmentation via Federated Intent Embedding and Multi-Mask Recommendation

Supplementary Material

1. Related Work

Interactive and intent-aware segmentation. Modern segmentation systems are based on fully convolutional networks [26], encoder–decoder architectures such as U-Net [42], and transformer-based models [10]. These segmentation architectures support dense prediction with strong contextual reasoning. Interactive segmentation methods extend base models by integrating user-provided input to guide or correct predictions during inference. Representative approaches include click-based refinement [25], extreme-point supervision [30], and iterative attention-based feedback [60]. Most interactive segmentation models operate on a per-image basis and require user input for every new prediction, without leveraging information from past interactions [37]. Repeated corrections are therefore necessary, even under consistent user intent [18, 21, 24, 53, 57].

Federated learning and personalization. Federated learning enables decentralized clients to train models without sharing raw data [15, 23, 39]. Applications in vision use adversarial alignment [11], domain discrimination [45], and transformer-based adaptation [16]. These approaches support client generalization but often require parameter updates or synchronization [46, 59, 61]. Personalized variants address client-specific variation but often involve architectural changes or local fine-tuning [5, 9, 43, 44, 61].

Retrieval-augmented and prompt-based segmentation. Retrieval-based methods improve predictions using external support examples or memories, especially in few-shot or exemplar-guided tasks [7, 31, 41]. Prompt-driven models such as SAM [7, 20] enable users to control segmentation via lightweight input. These systems depend on new guidance at each step and do not reuse feedback from earlier interactions [12, 36, 38]. Many retrieval approaches also rely on visual similarity or class labels, rather than capturing semantic intent [1, 4, 6, 7, 31].

Training-free and non-parametric adaptation. Test-time adaptation methods dynamically adapt models to target domains without labels. Domain alignment techniques such as DANN [11], ADDA [45], and DAFormer [16] perform distribution-level adaptation during training. More recent test-time techniques, including CoTTA [50], Slot-TTA [35], and IST [27], use pseudo-labeling, entropy minimization, or continual optimization for inference-time adaptation. Most test-time adaptation methods still require gradient-based updates or auxiliary loss optimization, which limits real-time applicability.

Comparison to the proposed approach. The Human-Intervention Segmentation framework introduced in this work models user corrections as reusable intent representations. Interactive segmentation approaches [37, 53, 54] do not retain feedback across sessions, whereas the proposed system supports intent reuse across inputs. Parameter-free personalization is achieved without weight updates or fine-tuning, unlike federated learning systems [11, 16, 45]. Retrieval is guided by latent intent embeddings rather than visual similarity, differentiating this approach from prompt-driven models [20]. Adaptation occurs without training or optimization, which distinguishes the proposed method from test-time adaptation strategies [27, 35, 50], enabling lightweight and privacy-preserving personalization.

2. Methods

Implementation. All conditioning layers are lightweight, typically consisting of 1–2 linear layers per FiLM block or of a single transformer-style cross-attention block per resolution level. The personalized decoder shares the same backbone as the baseline model for architectural consistency.

2.1. Multi-Output Decoding and Selection

The proposed framework employs a multi-branch decoding structure that generates multiple segmentation outputs in a single inference step. This structure enables real-time, training-free, and user-adaptive inference, allowing the model to incorporate both personalized intent and semantically relevant recommendations. For a given input image I_j, the framework produces:

- P_0: a baseline segmentation trained with ground-truth supervision,
- P_1: a personalized mask conditioned on the user's prior intent,
- $\{\tilde{P}_j^{(m)}\}_{m=1}^{M}$: a set of recommendation masks based on retrieved intent-aligned references.

Branch 1: Ground-Truth-Aligned Prediction. A standard decoder f_p, parameterized by fixed weights θ_M^t, generates the initial prediction $P_0 = f_p(I_j; \theta_M^t)$, which represents a domain-general segmentation output, aligned with objective annotation criteria but not user-specific preferences.

Branch 2: Personalized Intent-Conditioned Prediction. Given a user intent vector $z_j \in \mathbb{R}^d$, previously computed via $z_j = \phi(P_i, Q_i)$, an intent-conditioned decoder f_{his} produces a personalized segmentation $P_1 = f_{\text{his}}(I_j, z_j)$.

This prediction reflects prior user-specific refinements and is generated without any update to θ_M^t.

Branch 3: Multi-Reference Decoding via Federated Recommendation. To capture alternative interpretations from other users, the system retrieves M intent-guided reference subsets $\{\mathcal{R}_m\}_{m=1}^M$, where each subset $\mathcal{R}_m \subset \mathcal{R}_{\text{retrieved}}$ consists of relevant triplets (I_k, P_k, Q_k). A reference-aware decoder f_a then generates $\tilde{P}_j^{(m)} = f_a(I_j \mid \mathcal{R}_m)$ for $m = 1, \ldots, M$, where each $\tilde{P}_j^{(m)}$ encodes a distinct, plausible user-aligned output based on semantic priors extracted from other users' refinements. This design supports multiple valid mask hypotheses for inputs with inherent ambiguity or subjectivity.

Prediction Aggregation and User Selection. All outputs are presented to the user: $\mathcal{P}_j = \{P_0, P_1, \tilde{P}_j^{(1)}, \ldots, \tilde{P}_j^{(M)}\}$, from which a preferred result is selected: $P_j^{\text{final}} = \text{Select}(\mathcal{P}_j)$, based on subjective quality, domain fit, or user preference.

Feedback Loop for Continual Adaptation. The selected output P_j^{final} is optionally used to update the personalization module. The system pairs it with the baseline prediction P_0 to create a new intent embedding $z_j = \phi(P_0, P_j^{\text{final}})$, which is then added to the user's reference set $\mathcal{R}^u$. This mechanism supports continual refinement and long-term personalization in a non-parametric, training-free manner. This decoding structure enables simultaneous inference of diverse predictions conditioned on user-specific or federated knowledge, thereby distinguishing our framework from prior methods that require iterative annotation or per-example model fine-tuning. By separating visual feature extraction and intent conditioning, it enables one-to-many adaptation without compromising real-time performance.

2.2. Privacy-Preserving Intent Sharing

In the proposed federated framework, intent embeddings are shared across clients to enable cross-user personalization. However, to protect user identity and maintain data confidentiality, all shared vectors undergo privacy-preserving transformation before being uploaded to the global repository.

Anonymized Embedding Protocol. Each user $u \in \mathcal{U}$ encodes their local feedback into an intent vector $z_i^u = \phi(P_i^u, Q_i^u) \in \mathbb{R}^d$. Before uploading to the global pool, the vector is perturbed with additive Gaussian noise to enforce differential privacy:

$$\tilde{z}_i^u = z_i^u + \mathcal{N}(0, \sigma^2 \mathbf{I}),$$

where $\sigma = C \cdot \Delta_f / \varepsilon$, Δ_f denotes the global sensitivity of the intent encoder ϕ, C is a scaling constant, and ε is the user-defined privacy budget. This ensures that the contribution of any single intent vector remains indistinguishable under bounded inference.

User Identity Anonymization. To decouple embeddings from user identity, each intent vector $\tilde{z}_i^u$ is stored without metadata or explicit user tags. Instead, only the associated reference triplet (I_i, P_i, Q_i) is retained for decoding, and these are reindexed using randomized hash keys.

Secure Aggregation. The global intent repository $\mathcal{Z} = \bigcup_{u \in \mathcal{U}} \{\tilde{z}_i^u\}$ stores only anonymized and noise-protected vectors. No raw images, segmentation masks, or user labels are ever transmitted or exposed. Only statistical aggregates are used for similarity computation and diversity-aware retrieval, preserving compliance with federated privacy regulations and enabling real-world deployment in sensitive domains.

This mechanism ensures that user intent can be leveraged for personalized recommendations without compromising privacy or revealing identity, making the proposed system compatible with decentralized and high-stakes applications such as clinical imaging and defense analytics.

3. Experiments

Implementation Details. All input images are resized to 256×256. Models are trained for 100 epochs using the AdamW optimizer. The initial learning rate is set to 10^{-3} and is reduced to 10^{-5} when convergence becomes unstable. All experiments are conducted in a federated learning environment: each client device is equipped with an NVIDIA RTX 4070 Ti GPU, while the central server operates on a cluster of ten NVIDIA A5000 GPUs. To support real-time user interaction and feedback collection, we implement a web-based interface using React (v18.3.1) and Flask (v3.0.3), enabling seamless visualization and refinement of segmentation masks.

Inference-Time Personalization Without Test-Time Optimization

A central design choice of the proposed framework is the strict separation between offline learning and deployment-time personalization. All learnable components, including the shared visual encoder, the baseline decoder, the intent encoder ϕ, and the intent-conditioning layers used in the personalized decoder, are optimized during an offline training stage and remain frozen thereafter. At deployment time, the framework performs only forward computation, intent embedding extraction, storage, retrieval, and decoding. No backpropagation, gradient-based adaptation, or per-sample fine-tuning is applied during inference.

More specifically, once a user modifies a baseline prediction P_0 into a corrected mask Q, the system computes

an intent representation

$$z = \phi(P_0, Q),$$

which captures the semantic offset introduced by the correction. This embedding is then reused as a persistent conditioning signal for future images. For a new input I', the personalized output is generated as

$$P_1 = f_{\text{his}}(I', z),$$

without updating any model parameter. In this sense, the proposed method should be understood as an inference-time personalization framework rather than a test-time adaptation method. The practical implication is that user-specific behavior can be accumulated and reused over time while preserving stable latency and avoiding the optimization overhead that is common in adaptation-based approaches.

Prediction Flow and Interpretation of the Intent Variable

The prediction flow of the proposed framework can be interpreted as a transition from a generic segmentation hypothesis to a user-aligned semantic hypothesis. The baseline branch first produces P_0, which reflects the canonical behavior of the segmentation model under standard supervision. A user intervention then transforms P_0 into a refined mask Q, and the pair (P_0, Q) is encoded into the latent intent variable z. The role of z is not to store a target mask itself, but to represent the direction and structure of the user-preferred correction in a reusable latent form.

This interpretation also clarifies the prediction variables appearing throughout the manuscript. The mask P_0 denotes the baseline output of the frozen segmentation model. The mask Q denotes the user-corrected result derived from P_0. The vector z denotes the intent embedding computed from (P_0, Q). The mask P_1 denotes the intent-conditioned prediction for a future image using the stored intent representation. When federated retrieval is enabled, the system further generates a set of recommendation masks $\{\tilde{P}^{(m)}\}_{m=1}^{M}$ based on retrieved intent-aligned references.

Under this view, the overview figure should be read as follows. The upper prediction corresponds to the baseline mask P_0, whereas the lower prediction corresponds to the intent-conditioned output P_1. The corrected mask Q is an explicit input to the intent encoder and is used only to compute z; it is not a ground-truth target assumed to be directly available during future inference. This distinction is important because the framework is designed to reuse human corrections as persistent signals, rather than to assume access to a fully supervised target at deployment time.

Upper-Bound Analysis with Ground-Truth Substitution

The ground-truth-substituted setting is best interpreted as an upper-bound analysis of the intent-conditioning mechanism. In this controlled setting, the user-corrected mask is replaced with the dataset annotation to examine how strongly the decoder responds when the correction signal is maximally informative. This experiment is therefore not intended to redefine the Human-Intervention Segmentation problem, nor does it serve as the basis of the main claim. Instead, it functions as a diagnostic tool for isolating the behavior of the intent-conditioned branch under an idealized correction signal.

The main formulation studied in this work remains the user-guided setting in which $z = \phi(P_0, Q)$ is computed from actual human corrections. In that formulation, the central question is whether user refinements can be transformed into reusable semantic representations and applied to future inputs without additional optimization. The ground-truth-substituted experiment demonstrates that the decoding mechanism can exploit a strong correction prior when such information is available.

This interpretation is supported by the component-wise behavior observed in the ablation study. On ADE20K, the full HIS framework achieves 43.5 mIoU, whereas the baseline prediction P_0 yields 40.2. A conditioning-only variant reaches 41.0, while a retrieval-only variant reaches 40.0. When mismatched or randomly sampled intent embeddings are used, performance further drops to 39.6. These results suggest that the improvement is not explained by generic architectural effects alone. Rather, it emerges from the joint operation of intent encoding, intent-conditioned decoding, and reference-aware recommendation. The upper-bound experiment should therefore be read as supporting evidence for the conditioning mechanism's responsiveness, while the paper's main contribution remains the reuse of human intent captured from intervention history.

Robustness to Imperfect Human Feedback

A practical challenge in personalization is that user corrections are not always equally reliable. Some edits reflect stable and reusable preferences, whereas others may be noisy, image-specific, or exploratory. The proposed framework addresses this issue by treating intent capture as a deliberate act rather than an automatic update after every interaction. In practice, an intent is stored only when a correction is judged to represent a persistent preference. One-off adjustments can simply be discarded, preventing unstable feedback from contaminating the intent repository.

The framework also avoids overcommitting to a single personalized output. Instead of replacing the baseline prediction with a single mandatory user-conditioned mask, it exposes a one-to-many output space comprising the baseline mask, the personalized mask, and several recommendation masks derived from retrieved intent embeddings. This design is particularly important when the user signal is only partially informative or when multiple segmentations may be semantically acceptable. By allowing the user to choose among several candidates, the framework reduces the im-

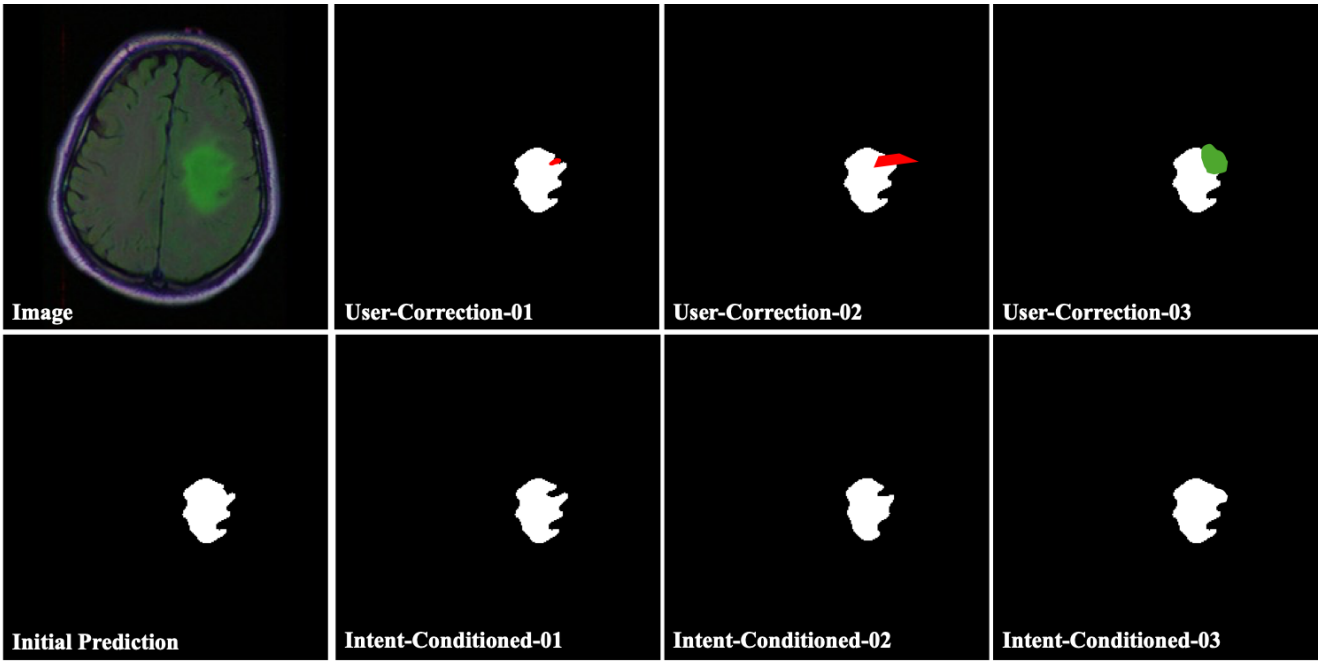

Figure 6. Qualitative examples of intent reuse on unseen images. For each case, the figure shows the baseline prediction P_0, the user-corrected mask Q, the intent-conditioned output P_1, and recommendation masks generated from retrieved intent embeddings. The results demonstrate that a single user intervention can be reused across future inputs without test-time optimization or gradient-based parameter updates.

pact of imperfect intent estimation and converts personalization into a guided selection process rather than a single hard decision.

The quality of the intent encoder remains an important factor in overall performance, since the usefulness of the retrieved and personalized masks ultimately depends on how faithfully ϕ captures the semantic structure of the correction. For this reason, the one-to-many design is not merely an interface convenience, but also a robustness mechanism. It provides a safeguard against encoder errors, ambiguous interventions, and cross-user variability, while preserving the main advantage of reusable personalization.

Backbone Modularity and Practical Scope

Another important property of the proposed framework is that the backbone is kept frozen by design. This choice is motivated by practical deployment considerations: avoiding per-inference optimization improves stability, reduces computational overhead, and allows the system to operate with predictable latency. At the same time, freezing the backbone does not imply that the method is tied to a specific architecture. The personalization mechanism is modular and can be integrated with different segmentation backbones via the same intent-conditioning interface.

This modularity is important when interpreting the scope of comparison in the experimental section. The current implementation employs a compact segmentation backbone to prioritize deployment efficiency and the impact of reusable intent conditioning. As a result, the method is not intended to claim universal superiority over every larger or more specialized segmentation architecture in raw standalone accuracy. Rather, the main claim is that a frozen segmentation backbone can be endowed with persistent, user-aligned behavior through intent embeddings, without sacrificing the computational advantages of inference-only operation. Stronger or domain-specific backbones can therefore be incorporated within the same formulation, suggesting that the contribution of this work lies primarily in the personalization mechanism rather than in a particular choice of base segmentor.

Qualitative Analysis of Intent Reuse

Fig. 6 provides qualitative examples of intent reuse on unseen images. For each example, we show the baseline prediction, the user-amended mask, the intent-conditioned output, and the recommendation masks generated from retrieved intent embeddings. These examples illustrate that a single intervention can be converted into a persistent semantic representation and reused beyond the original image where the correction was made.

Two observations are particularly notable. First, the personalized output tends to preserve the semantic tendency introduced by the user's correction, even on new inputs. This indicates that the embedding does not merely memo-

rize a local mask pattern but instead captures a transferable correction preference. Second, the recommendation masks provide semantically meaningful alternatives rather than arbitrary perturbations, which supports the role of federated retrieval as a source of diverse yet relevant candidate segmentations. To summarize, these results reinforce the central premise of the paper: human correction is more useful when treated as a reusable semantic signal than as a transient per-image refinement cue.

4. Discussion

Intent Persistence Beyond HITL. Previous research on interactive image segmentation has primarily focused on improving per-image prediction accuracy by leveraging iterative user input, such as clicks or strokes. These approaches typically follow a human-in-the-loop (HITL) paradigm, where user feedback influences only the current inference and is discarded afterward. This formulation limits generalization across inputs and often leads to repeated errors, even after similar corrections. The proposed Human-Intervention Segmentation (HIS) framework overcomes this limitation by introducing a persistent adaptation mechanism that encodes user-guided refinements into reusable intent embeddings. These embeddings enable personalized inference without parameter updates or retraining. A single interaction can inform future predictions, providing long-term personalization. In addition, a federated retrieval module expands personalization by suggesting segmentation variants derived from semantically similar interactions by other users, all within a privacy-preserving infrastructure.

Scalability and Future Extensions. The overall system architecture supports a multi-branch decoder that generates a ground-truth-aligned baseline, an intent-conditioned prediction, and a set of recommendation-based outputs in a single pass. This design enables one-to-many decoding, capturing diverse semantic interpretations of the same input. Although this study focuses on static 2D segmentation tasks, the underlying formulation is applicable to broader domains. Future work may explore video segmentation or 3D medical imaging, which require temporal and spatial consistency. Moreover, the representation of intent as a structured embedding opens the possibility of extending beyond segmentation, such as object detection or multimodal editing. Combining this framework with large language models (LLMs) may further enable natural language-based intent communication, offering new directions for personalized and accessible human-AI collaboration.