PainDiffusion: Learning to Express Pain

Quang-Tien Dam, Tri-Tung-Nguyen Nguyen, Yuki Endo, Dinh-Tuan Tran, Joo-Ho Lee

RITSUMEIKAN UNIVERSITY

IROS HANGZHOU 2025

Give it a star

Current pain synthesis methods for patient robots are not automatic

Dental Robot Simulator
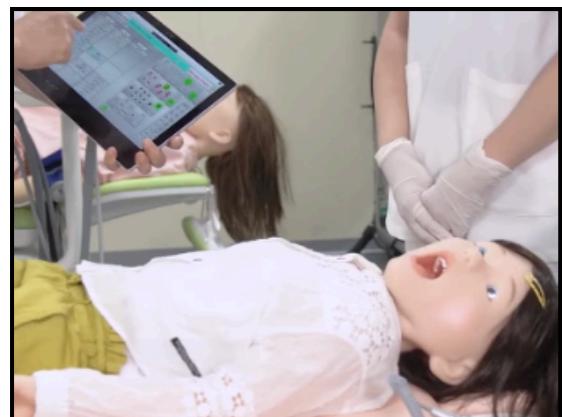
©2025 tmsuk Pedia_roid

**General Patient Robot Simulator**

Robot HAL® S3201

The current patient robots in the market have limited and static emotional expressions.

|	Classify	Synthesize	Rule-based	Learning-based
Moosaei et al (2017)		✅	✅	
Huang et al (2019)	✅			✅
Haque et al (2018)	✅			✅
Lee et al (2021)		✅	✅	
This method		🏅		🏅

To the best of our knowledge, we are the first to tackle pain synthesis with a learning-based approach, which aims to make more realistic patient robots.

Autoregressively and Continuously Generate Reaction Sequences

1. Overall

Past frames — Noisy frames

Stimuli signal

Response config — Strong Happy

Encoder → Model → Generating with Diffusion Forcing [2]

Clean frames

Decoder
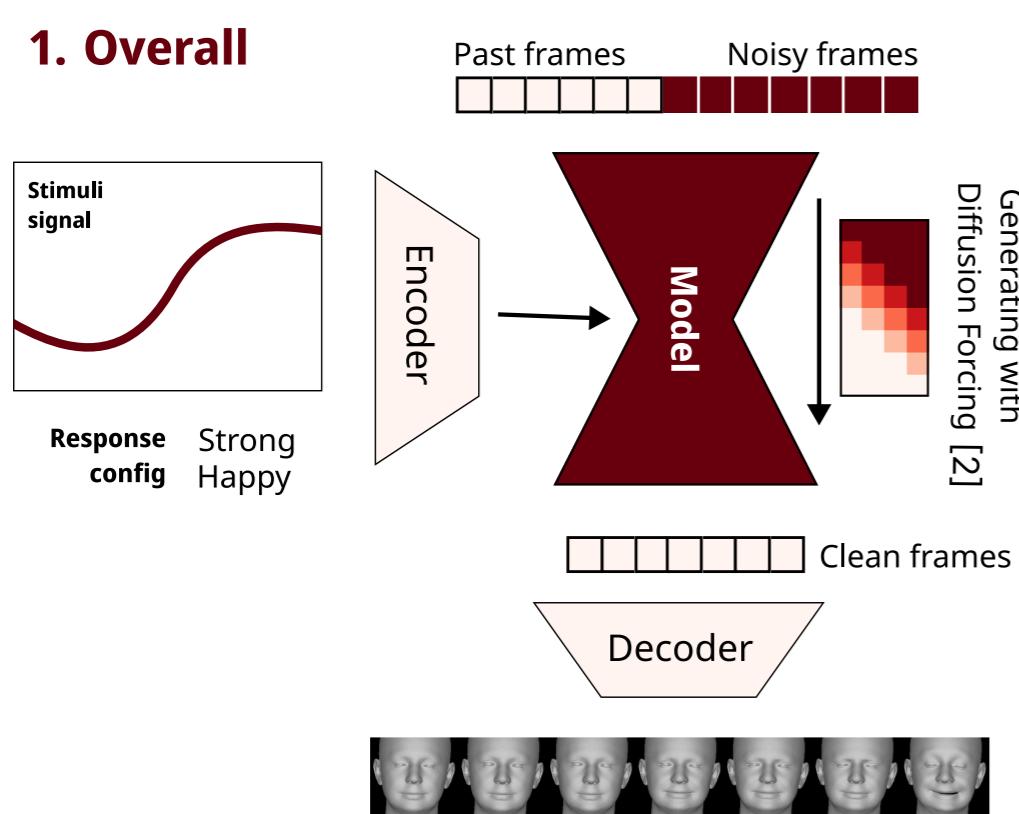
We train PainDiffusion on the Biovid Heatpain Database [1] with Elucidated Diffusion objective [3].

$$\mathbb{E}_{\boldsymbol{y}\sim p_{\text{data}},\boldsymbol{n}\sim\mathcal{N}(\boldsymbol{0},\sigma^2\mathbf{I})}\left[\left\|\boldsymbol{y}_{\text{pred}}-\boldsymbol{y}_{\text{target}}\right\|_2^2\right]$$

$$y_{\text{pred}} = F_\theta\left(c_{\text{in}}\left(\sigma\right)\cdot\left(\boldsymbol{y}+\boldsymbol{n}\right), c_{\text{noise}}\left(\sigma\right)\right)$$

$$y_{\text{target}} = \frac{1}{c_{\text{out}}\left(\sigma\right)}\left(\boldsymbol{y}-c_{\text{skip}}\left(\sigma\right)\cdot\left(\boldsymbol{y}+\boldsymbol{n}\right)\right)$$

2. Unet block

Conditions embeddings

z → Resnet Bl. → Resnet Bl. → Spatial attn. → Temporal attn. → Down/up conv. → z'

Noise level embeddings — Temporal embeddings
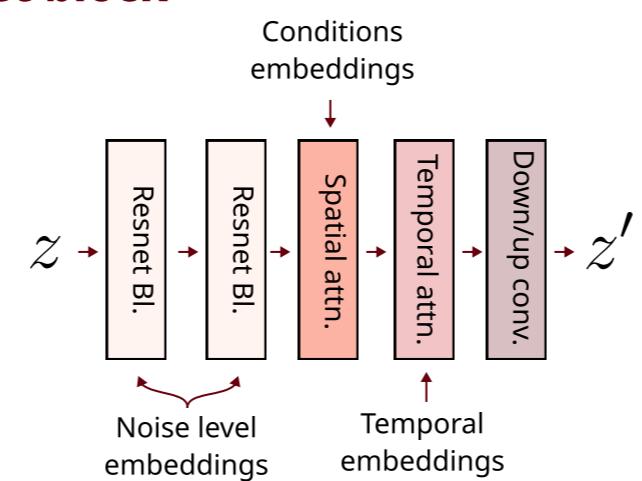
Both noise level and temporal embeddings are from sinusoidal positional encoding to add *positional information in both noise dimension and time dimension*.

To flexibly tune the influence of each controlling signal we use a fine-grain version of classifier-free guidance in inference time.

$$\hat{z} = \left(1+\sum_{c\in C}\lambda_c\right)\cdot F_\theta\left(z,t,C\right)-\sum_{c\in C}\lambda_c\cdot F_\theta\left(z,t,C|_{c=\emptyset}\right)$$

3. Data pipeline

Video, Stimuli → Filter → Sync → Extract AUs → Pain Index; Extract emotion → Emotion Index; → Stimuli signal → C; Extract FLAME exp. → Expression code → z

4. Roll-out strategy

Predicting window

prediction 1
prediction 2
prediction 3
prediction 4

Plan new frames that are coherent with the context frames

Context frames are from previous ending frames

Each prediction considers observation at the moment
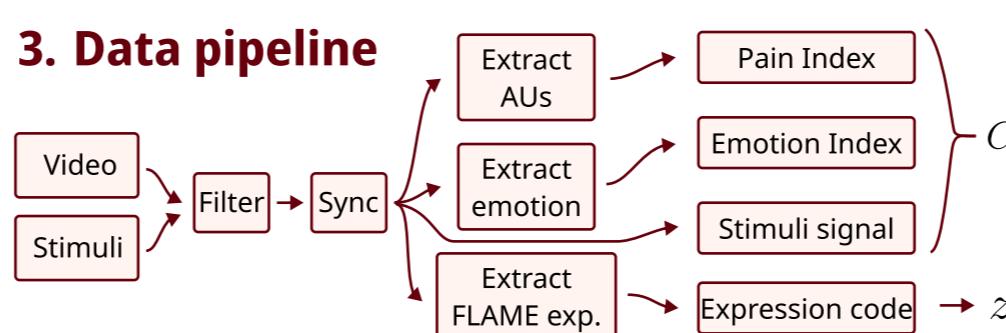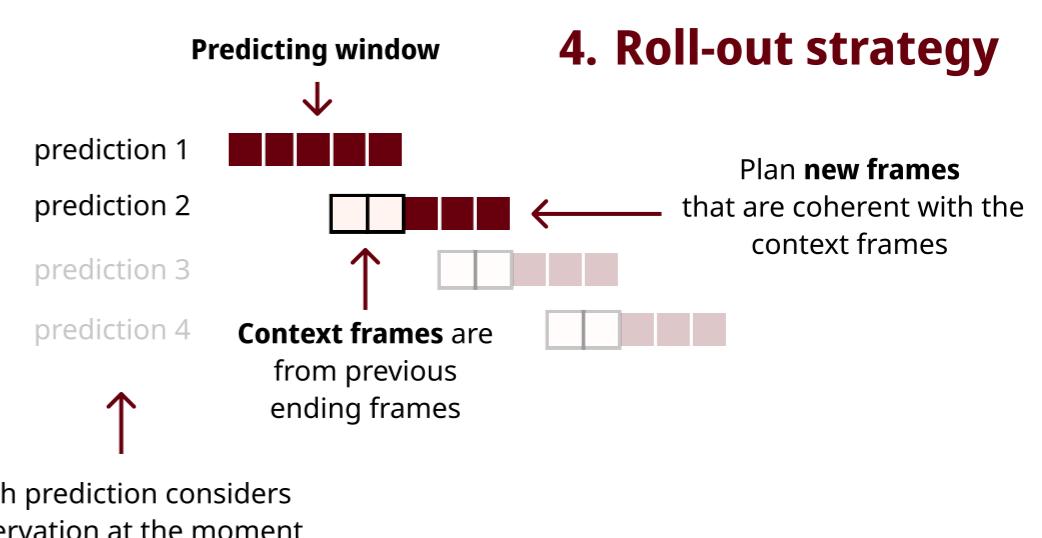
The context frames are added with *little noise level* to give more *freedom in predicting the future frames* while the future frames start from totally noise. The ratio of context frames and the length of the predicting window will affect how fast the model plan new frames.

5. Elbow range of motion exercise

Render and display ← GaussianAvatars ← Generate expression code ← PainDiffusion ← Single joint arm ← Move the arm ← Linear mapping angle to stimuli signal
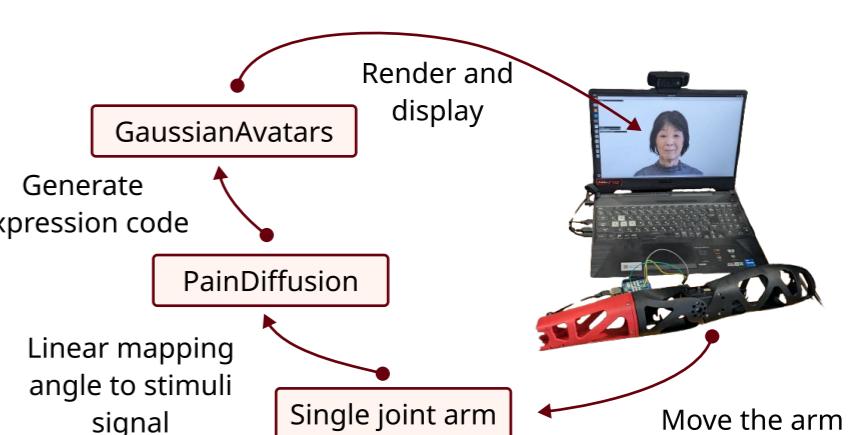
PainDiffusion outperforms the autoregressive discretization method

We run 5 predictions to create the range of graph, the pain index signal from the ground truth recording *falls within* the predicted range of PainDiffusion while autoregressive baseline is higher.

Start of stimuli — End of stimuli

— Groundtruth — Autoregressive — PainDiffusion

Sim (PSPI) (Lower is Better)
Corr (PSPI) (Higher is Better)
Divrs (FLAME Params) (Higher is Better)
Var (FLAME Params) (Higher is Better)
Dist (FLAME Params) (Lower is Better)

Ground truth
Random Training Sample
Nearest Neighbor
FSQ-VAE Autoregressive
PainDiffusion w/ Full-seq
PainDiffusion w/ Diffusion Forcing

Experts chose PainDiffusion 31% of the time in the video preference test
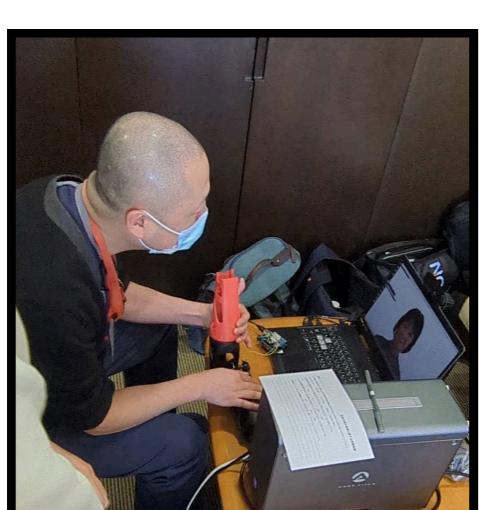
Expressions against stimuli signal is quite real

We asked clinicians to practice rehabilitation elbow range of motion exercise on a single joint arm then rate the expression of the virtual avatar.

Pain Strength — Expression Realism — Similarity to Human
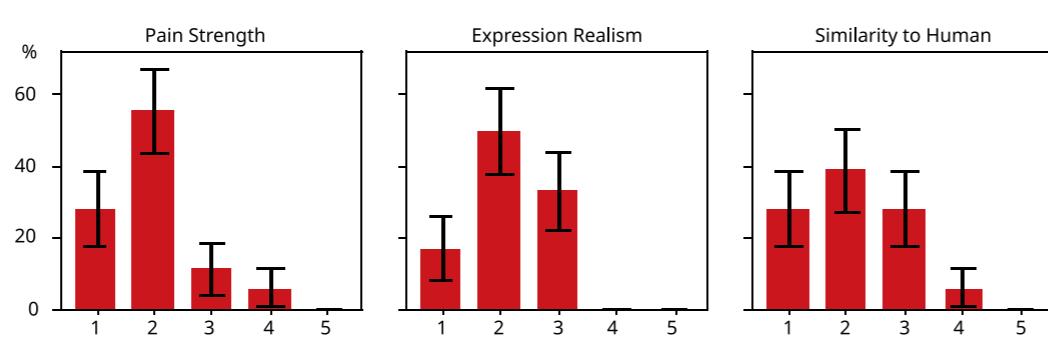
Diversity of motion rated by clinicians

Very diverse
Diverse
Slightly Diverse
Similar
Very similar

Size — Speed — Overall
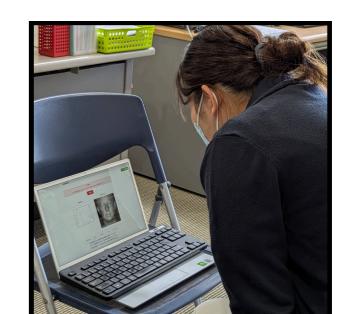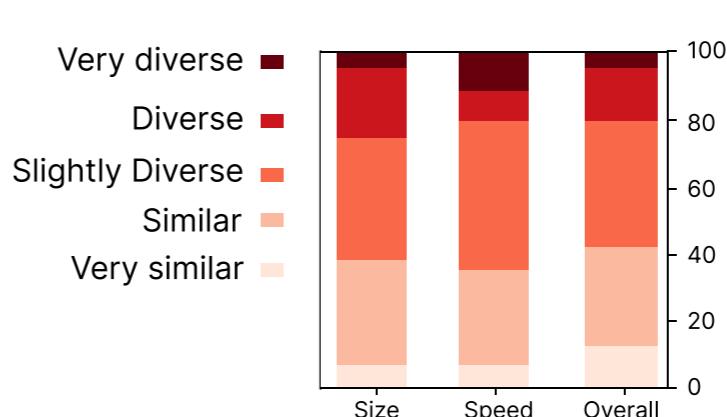
Preference test is between PainDiffusion and groundtruth.

Magnitude of avatar movement across configurations

Average Movement

Stimuli Levels — Emotion — Expressiveness Config

Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, Surprise
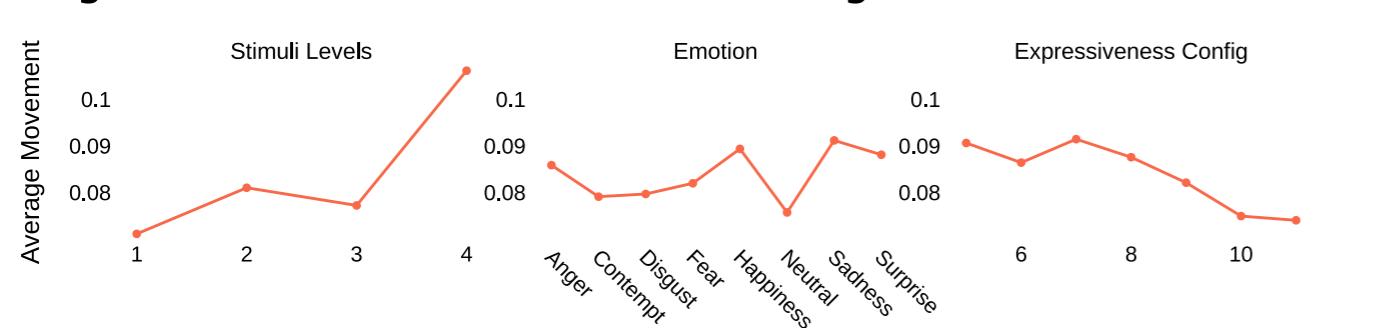
Real examples are also judged as weak

Distribution of Reaction Strength Ratings

Paindiffusion — Groundtruth

Percentage (%)

Very weak, Weak, Slightly strong, Strong, Very strong
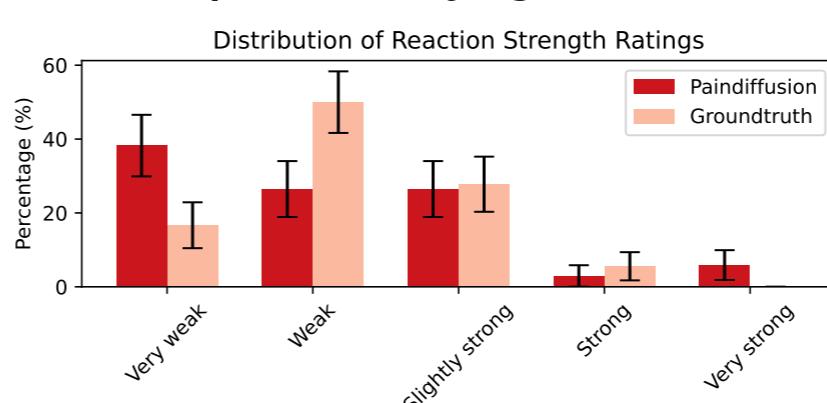
Further insights from clinicians

1. A *calibrated baseline* is needed for better assessing the expression.

2. Cultural elements (e.g. patient record) are necessary to make a more realistic pain expression model.

[1] S. Walter et al. "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system"
[2] B. Chen et al. "Diffusion forcing: Next-token prediction meets fullsequence diffusion"
[3] T. Karras et al. "Elucidating the Design Space of Diffusion-Based Generative Models"