Supplementary Table S1. Implementation details and reproducibility settings

Component	Specification
Visual backbone	Pretrained EfficientNetV2-S
Audio backbone	EfficientNetV2-S, single-channel input
Transformer architecture	Zorro-inspired masked attention
Embedding dimension	512
Transformer layers	8
Attention heads	8
Attention head dimension	64
Feed-forward hidden dimension	2,048
Fusion tokens	6 learnable tokens
Positional embedding	Learnable 3D positional embedding for video
Dropout	0.2
Video sampling	8 uniformly spaced frames across each recording
Image resolution	256 × 256 pixels
Visual augmentation	Horizontal flip (0.3), shift-scale-rotate (0.3), color jitter (0.2), grid dropout (0.3), grid distortion (0.1), Gaussian blur (0.2), Gaussian noise (0.2)
Validation/test augmentation	Resize only
Optimizer	AdamW
Learning rate	1×10^{-4}
Batch size	4
Learning-rate schedule	Cosine annealing with warmup
Early stopping	Validation balanced accuracy, patience 8
Max Epochs	64
Random seed	831
Cross-validation	5 outer × 5 inner folds
Ordinal classification	CORN loss, 3 classes
Outer-fold prediction	Average of five inner-model class probability vectors

Supplementary Table S2. Software environment

Software	Version
Python	3.10.12
PyTorch	2.4.0
CUDA	12.5
timm	1.0.15
Albumentations	2.0.8
coral-pytorch	1.4.0

Supplementary Table S3. Patient-balanced performance

A. Recording-level and patient-balanced performance

Pain scale	Model	Recording-level BA	Patient-balanced BA (95% CI)	Recording-level QWK	Patient-balanced QWK (95% CI)
PAINAD	Image-only	0.785	0.778 (0.731–0.825)	0.760	0.749 (0.692–0.801)
PAINAD	Video-only	0.808	0.801 (0.759–0.844)	0.793	0.783 (0.733–0.826)
PAINAD	Video+Audio	0.817	0.810 (0.766–0.854)	0.801	0.793 (0.738–0.840)
NRS	Image-only	0.737	0.726 (0.685–0.765)	0.666	0.657 (0.613–0.699)
NRS	Video-only	0.732	0.726 (0.689–0.764)	0.736	0.731 (0.689–0.769)
NRS	Video+Audio	0.700	0.697 (0.659–0.734)	0.686	0.686 (0.643–0.728)

B. Paired differences in recording-level performance

Pain scale	Comparison	BA difference (95% CI)	QWK difference (95% CI)
PAINAD	Video-only − Image-only	+0.023 (−0.001–0.047)	+0.033 (−0.001–0.068)
PAINAD	Video+Audio − Video-only	+0.009 (−0.012–0.030)	+0.008 (−0.022–0.037)
NRS	Video-only − Image-only	−0.005 (−0.044–0.034)	+0.070 (0.029–0.111)
NRS	Video+Audio − Video-only	−0.032 (−0.084–0.019)	−0.050 (−0.110–0.011)

C. Recording-level class-specific recall with patient-clustered 95% confidence intervals

Pain scale	Model	No pain recall (95% CI)	Mild pain recall (95% CI)	Moderate-to-severe recall (95% CI)
PAINAD	Image-only	0.823 (0.723–0.920)	0.817 (0.758–0.870)	0.716 (0.623–0.796)
PAINAD	Video-only	0.815 (0.716–0.912)	0.833 (0.778–0.884)	0.775 (0.698–0.840)
PAINAD	Video+Audio	0.831 (0.732–0.928)	0.866 (0.816–0.911)	0.754 (0.667–0.827)
NRS	Image-only	0.990 (0.971–1.000)	0.682 (0.571–0.785)	0.539 (0.465–0.608)
NRS	Video-only	0.990 (0.971–1.000)	0.500 (0.390–0.608)	0.706 (0.649–0.760)
NRS	Video+Audio	0.990 (0.971–1.000)	0.509 (0.393–0.618)	0.600 (0.519–0.681)

Note: In Panel A, patient-balanced performance was calculated by weighting each recording by the inverse of the number of recordings contributed by its patient. In Panel B, paired differences were calculated using predictions from the same recordings. In Panel C, class-specific recall was calculated at the recording level using pooled out-of-fold predictions for all 666 recordings. All 95% CIs were estimated using 10,000 patient-cluster bootstrap iterations.

BA, balanced accuracy; QWK, quadratic weighted kappa; CI, confidence interval.

Supplementary Table S4. Clinical Characteristics of Pain Assessments in the Postanesthesia Care Unit

Characteristic	Result
PACU recordings, n	564
Recordings with valid assessment timestamps, n (%)	563 (99.8%)
Time from PACU admission to assessment, median (IQR), min	22 (6–41)
Time from PACU admission to assessment, range, min	0–69
Analgesic exposure before assessment, n (%)	
Confirmed prior analgesic administration	412 (73.0%)
No documented analgesic administration before assessment	102 (18.1%)
Indeterminate exposure status	50 (8.9%)
Standardized sedation assessment	Not Available

Note: Data are presented at the recording level. Analgesic exposure was determined from documented administration times relative to recording timestamps; ambiguous cases were classified as indeterminate.

Supplementary Table S5. Outer-Fold Performance of PAINAD and NRS-Supervised Models

A. PAINAD-supervised models

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± SD
Image-only	BA	0.807	0.780	0.807	0.796	0.739	0.786 ± 0.029
	QWK	0.817	0.763	0.771	0.776	0.669	0.759 ± 0.055
Video-only	BA	0.813	0.813	0.791	0.833	0.790	0.808 ± 0.018
	QWK	0.825	0.824	0.756	0.824	0.737	0.793 ± 0.043
Video+Audio	BA	0.845	0.816	0.823	0.840	0.761	0.817 ± 0.033
	QWK	0.855	0.806	0.812	0.832	0.697	0.800 ± 0.061

B. NRS-supervised models

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± SD
Image-only	BA	0.769	0.752	0.731	0.725	0.711	0.738 ± 0.023
	QWK	0.713	0.606	0.621	0.654	0.733	0.665 ± 0.056
Video-only	BA	0.754	0.755	0.743	0.703	0.718	0.735 ± 0.023
	QWK	0.663	0.733	0.739	0.812	0.742	0.738 ± 0.053
Video+Audio	BA	0.695	0.744	0.636	0.659	0.765	0.700 ± 0.055
	QWK	0.838	0.732	0.410	0.673	0.781	0.687 ± 0.166

Note: Performance was evaluated separately on each held-out outer fold of the five-fold nested cross-validation. Mean ± SD represents the unweighted mean and sample standard deviation across the five outer folds.

BA, balanced accuracy; QWK, quadratic weighted kappa; SD, standard deviation.

Supplementary Table S6. Distribution of Pain Scores by Recording Setting

Reference	Recording setting	Total (N)	0, n/N (%)	1–3, n/N (%)	4–10, n/N (%)
PAINAD	Preoperative ward	102	102/102 (100.0%)	0/102 (0.0%)	0/102 (0.0%)
PAINAD	Postoperative PACU	564	22/564 (3.9%)	306/564 (54.3%)	236/564 (41.8%)
NRS	Preoperative ward	102	102/102 (100.0%)	0/102 (0.0%)	0/102 (0.0%)
NRS	Postoperative PACU	564	1/564 (0.2%)	110/564 (19.5%)	453/564 (80.3%)

Supplementary Table S7. Distributions of Raw Pain Scores and PAINAD Ratings

A. Raw PAINAD and NRS score distributions

Pain scale	Overall (n = 666)	Preoperative ward (n = 102)	PACU (n = 564)
PAINAD	3 (1–4)	0 (0–0)	3 (2–5)
NRS	5 (3–7)	0 (0–0)	5 (4–7)

Note: Values are presented as median (Q1–Q3)

B. Overall distribution of PAINAD domain-level scores

PAINAD domain	Score 0, n (%)	Score 1, n (%)	Score 2, n (%)
Breathing	446 (67.0%)	220 (33.0%)	0 (0.0%)
Negative vocalization	297 (44.6%)	296 (44.4%)	73 (11.0%)
Facial expression	130 (19.5%)	363 (54.5%)	173 (26.0%)
Body language	236 (35.4%)	394 (59.2%)	36 (5.4%)
Consolability	590 (88.6%)	72 (10.8%)	4 (0.6%)

Note: 666 recordings for each domain. Percentages are calculated within each domain.

C. Distribution of PAINAD domain-level scores stratified by pain category

PAINAD category	PAINAD domain	Score 0, n (%)	Score 1, n (%)	Score 2, n (%)
0 (n = 124)	Breathing	124 (100.0%)	0 (0.0%)	0 (0.0%)
	Negative vocalization	124 (100.0%)	0 (0.0%)	0 (0.0%)
	Facial expression	124 (100.0%)	0 (0.0%)	0 (0.0%)
	Body language	124 (100.0%)	0 (0.0%)	0 (0.0%)
	Consolability	124 (100.0%)	0 (0.0%)	0 (0.0%)
1–3 (n = 306)	Breathing	286 (93.5%)	20 (6.5%)	0 (0.0%)
	Negative vocalization	167 (54.6%)	139 (45.4%)	0 (0.0%)
	Facial expression	6 (2.0%)	297 (97.1%)	3 (1.0%)
	Body language	112 (36.6%)	194 (63.4%)	0 (0.0%)
	Consolability	305 (99.7%)	1 (0.3%)	0 (0.0%)
4–10 (n = 236)	Breathing	36 (15.3%)	200 (84.7%)	0 (0.0%)
	Negative vocalization	6 (2.5%)	157 (66.5%)	73 (30.9%)
	Facial expression	0 (0.0%)	66 (28.0%)	170 (72.0%)
	Body language	0 (0.0%)	200 (84.7%)	36 (15.3%)
	Consolability	161 (68.2%)	71 (30.1%)	4 (1.7%)

Supplementary Table S8. Cross-supervision evaluation of pain classification models

A. Model performance across reference standards

Supervision	Evaluation reference	Balanced accuracy	QWK	Ordinal MAE
PAINAD	PAINAD	0.8078	0.7928	0.1952
NRS	PAINAD	0.6896	0.6126	0.3844
PAINAD	NRS	0.7390	0.6386	0.4084
NRS	NRS	0.7322	0.7360	0.2883

B. Paired differences between PAINAD-supervised and NRS-supervised models

Evaluation reference	Metric	Difference (PAINAD − NRS)	95% CI
PAINAD	Balanced accuracy	+0.1181	0.0798 to 0.1536
PAINAD	QWK	+0.1802	0.1228 to 0.2397
NRS	Balanced accuracy	+0.0067	−0.0442 to 0.0549
NRS	QWK	−0.0975	−0.1441 to −0.0522

Supplementary Table S9. Face-cropping fallback rates and associated prediction errors

A. Face-cropping outcomes

Measure	n/N	Percentage (%)
Successfully cropped frames	5,178/5,328	97.18
Frames requiring fallback to original image	150/5,328	2.82
Missing cropped-image files	0/5,328	0.00
Recordings without fallback	597/666	89.64
Recordings with ≥1 fallback frame	69/666	10.36

B. Classification errors in the video-only models by face-cropping fallback status

Model	Fallback status	Recordings, n	Classification errors, n	Error rate (%)
PAINAD-supervised	No fallback	597	107	17.92
PAINAD-supervised	≥1 fallback frame	69	20	28.99
NRS-supervised	No fallback	597	179	29.98
NRS-supervised	≥1 fallback frame	69	10	14.49

Supplementary Table S10. Exploratory audio perturbation analysis

Model	Audio condition	N	Balanced Accuracy	QWK	ΔBA	ΔQWK
PAINAD	Original	666	0.8170	0.8008	Reference	Reference
	Zeroed	666	0.7849	0.7855	−0.0321	−0.0153
	Shuffled	666	0.7571	0.7426	−0.0599	−0.0582
NRS	Original	666	0.6999	0.6856	Reference	Reference
	Zeroed	666	0.7302	0.7426	+0.0303	+0.0570
	Shuffled	666	0.6814	0.6300	−0.0185	−0.0556

Note: *Original audio* refers to the unmodified audio paired with each video recording. *Zeroed audio* refers to the condition in which the preprocessed audio features were replaced with zeros while retaining the original video input. *Shuffled audio* refers to the condition in which the audio features were randomly exchanged between recordings from different patients, while the original video input was preserved. No model retraining was performed.

Supplementary Table S11. Sensitivity analysis of prior-adjusted predictions from the video-only models

Model	Prediction	Balanced accuracy	QWK	Recall 0	Recall 1	Recall 2	Brier score	ECE
PAINAD	Original	0.8078	0.7928	0.8145	0.8333	0.7754	0.2895	0.0605
PAINAD	Prior-adjusted	0.8159	0.8017	0.8145	0.9085	0.7246	0.2878	0.0534
NRS	Original	0.7322	0.7360	0.9903	0.5000	0.7064	0.3802	0.0541
NRS	Prior-adjusted	0.6658	0.8376	0.9903	0.0182	0.9890	0.2659	0.0333

Note: Prior adjustment was applied to out-of-fold class probabilities using class proportions from the corresponding outer training folds and the expected uniform class distribution under inverse-frequency weighted sampling. Performance was assessed without model retraining. The Brier score summarizes overall probability prediction error, and ECE summarizes estimated miscalibration.

Higher values indicate better balanced accuracy, QWK, and recall; lower values indicate better Brier score and ECE.

Table S12. Performance of NRS-supervised models by input configuration.

Modality	**Model**	**QWK**	**Balanced Accuracy**	**Recall** (No Pain)	**F1** (No Pain)	**Recall** (Mild)	**F1** (Mild)	**Recall** (Moderate to Severe)	**F1** (Moderate to Severe)
Image-Only (Single Frame)	NRS	0.666	0.737	0.990	0.971	0.682	0.383	0.539	0.668
Video-Only (8 Frames)	NRS	0.736	0.732	0.990	0.986	0.500	0.372	0.706	0.772
Video + Audio	NRS	0.686	0.700	0.990	0.990	0.509	0.322	0.600	0.699

Performance metrics were calculated at the recording level using pooled out-of-fold predictions from the five outer folds of patient-level nested cross-validation. Reference NRS scores were categorized as no pain (score 0), mild pain (scores 1–3), and moderate-to-severe pain (scores 4–10). Balanced accuracy was calculated as the unweighted mean of class-specific recall. QWK measured chance-corrected ordinal agreement between predicted and reference categories, with progressively greater penalties for disagreements between more distant categories. NRS-supervised models were evaluated as a secondary exploratory clinical benchmark. Differences among input configurations were interpreted descriptively; no formal hypothesis tests were performed. NRS, Numeric Rating Scale; QWK, quadratic weighted kappa.

Table S13. Descriptive reference-specific classification errors of the PAINAD- and NRS-supervised video-only models.

Classification error	Definition	PAINAD n/N (%)	NRS n/N (%)
Higher-category false positive	Lower reference category; higher predicted category	48/430 (11.2%)	56/213 (26.3%)
Higher-category false negative	Higher reference category; lower predicted category	53/236 (22.5%)	133/453 (29.4%)

Analyses used pooled out-of-fold predictions from the video-only models, with each model evaluated against its corresponding reference. Scores of 0–3 defined the lower category and scores of 4–10 the higher category. False-positive denominators comprised lower-reference recordings, whereas false-negative denominators comprised higher-reference recordings. These target-specific proportions were reported descriptively without between-reference risk differences or hypothesis tests. PAINAD, Pain Assessment in Advanced Dementia; NRS, Numeric Rating Scale.

Table S14. Performance of PAINAD- and NRS-supervised video-only models evaluated using PACU-restricted out-of-fold predictions.

A. Distribution of reference pain categories in PACU recordings

Reference	Total recordings	No pain	Mild pain	Moderate-to-severe pain
PAINAD	564	22 (3.9%)	306 (54.3%)	236 (41.8%)
NRS	564	1 (0.2%)	110 (19.5%)	453 (80.3%)

B. Overall performance in the full dataset and PACU-only recordings

Supervision	Full-dataset BA	PACU-only BA (95% CI)	Full-dataset QWK	PACU-only QWK (95% CI)
PAINAD	0.808	0.536 (0.510–0.560)	0.793	0.584 (0.506–0.654)
NRS	0.732	0.402 (NE*)	0.736	0.149 (0.063–0.234)

C. Class-specific recall in PACU recordings

Supervision	No pain recall (95% CI)	Mild pain recall (95% CI)	Moderate-to-severe recall (95% CI)
PAINAD	0.000 (NE†)	0.833 (0.777–0.884)	0.775 (0.698–0.841)
NRS	0.000 (NE*)	0.500 (0.392–0.611)	0.706 (0.649–0.760)

The pooled analysis included 666 recordings from the ward and PACU. The PACU analysis used the subset of 564 PACU recordings. Both analyses used the original out-of-fold predictions from the PAINAD- and NRS-supervised video-only models, retaining the original patient-level fold assignments. Each model was evaluated against its corresponding reference scale. Percentages in Panel A are based on the 564 PACU recordings for each scale. The 95% CIs were estimated using 10,000 patient-cluster bootstrap replicates. † The PAINAD-supervised model correctly classified none of the 22 PACU recordings categorized as no pain by PAINAD (0/22). The empirical bootstrap interval collapsed to 0.000–0.000 and is marked NE. * The NRS-supervised model misclassified the only PACU recording categorized as no pain by NRS (0/1). No-pain recall and three-class BA are presented as descriptive point estimates without CIs. BA, balanced accuracy. CI, confidence interval. NE, confidence interval not meaningfully estimable. NRS, Numeric Rating Scale. PACU, postanesthesia care unit. PAINAD, Pain Assessment in Advanced Dementia. QWK, quadratic weighted kappa.

Method S1. CORN probability conversion

Ordinal classification was performed using the Conditional Ordinal Regression for Neural Networks (CORN) framework with three ordered pain categories (0, no pain; 1, mild pain; and 2, moderate-to-severe pain). The model generated two logits, which were transformed into conditional probabilities using the sigmoid function.

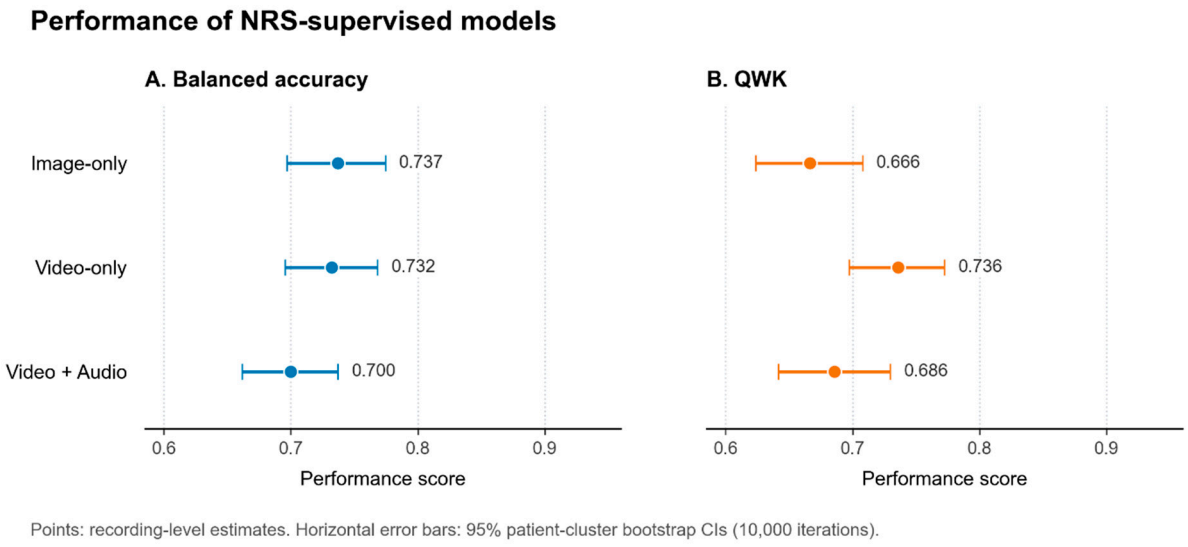
Let $q_1 = \sigma(z_1)$ and $q_2 = \sigma(z_2)$, where σ denotes the sigmoid function. The probabilities of the three pain categories were calculated as follows:

$$P(Y = 0) = 1 - q_0$$
$$P(Y = 1) = q_0(1 - q_1)$$
$$P(Y = 2) = q_0 q_1$$

For each outer test fold, the class probabilities obtained from the five inner-fold models were averaged, and the final predicted category was determined by selecting the class with the highest averaged probability.

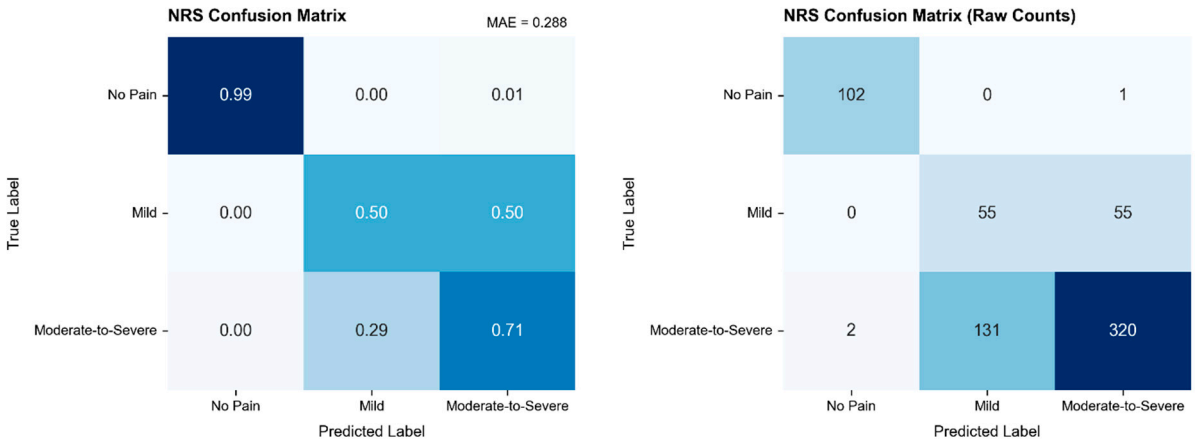
Figure Legends

Figure S1. Performance of NRS-supervised models across different input configurations.



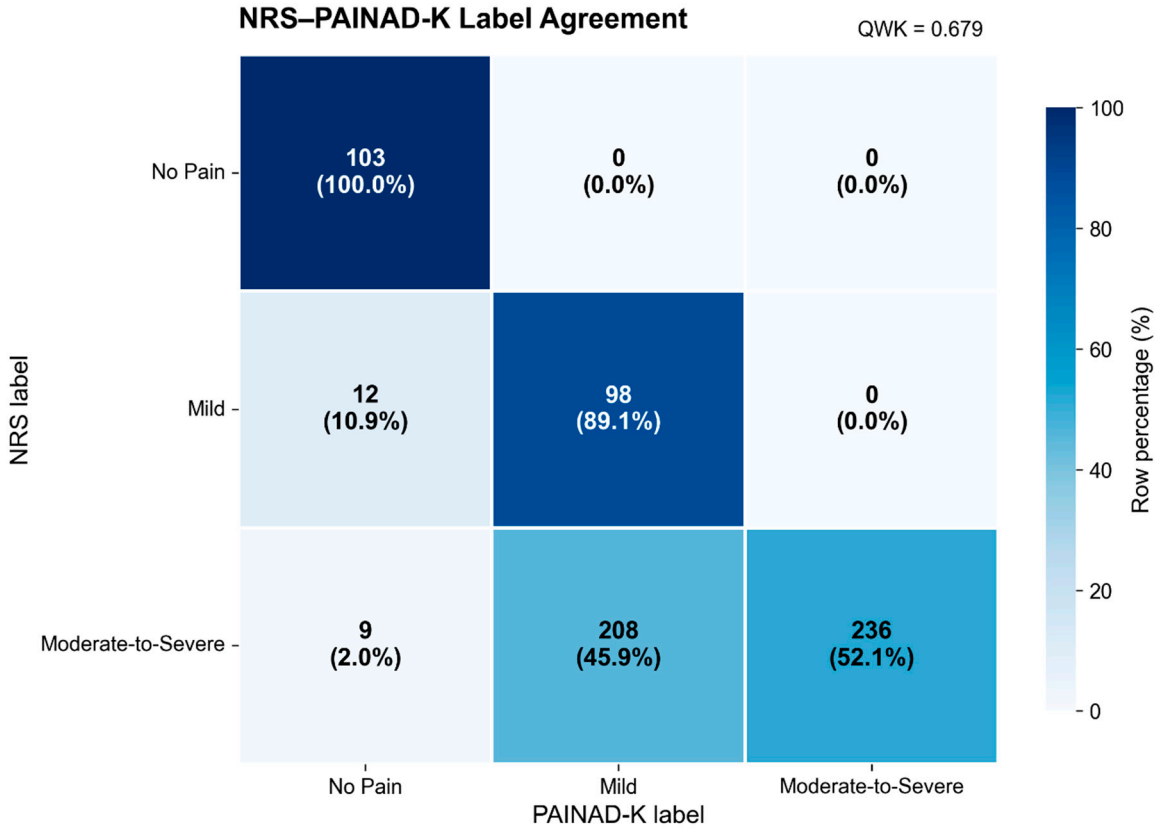
Balanced accuracy and quadratic weighted kappa are presented for the image-only, video-only, and video-plus-audio models. These values were calculated from pooled out-of-fold predictions derived from the five outer folds of patient-level nested cross-validation. Numbers adjacent to the markers indicate the corresponding point estimates. Horizontal error bars indicate 95% confidence intervals estimated from 10,000 patient-cluster bootstrap replicates. Differences among input configurations were evaluated descriptively without formal hypothesis testing. NRS, Numeric Rating Scale.

Figure S2. Confusion matrices for the NRS-supervised video-only model based on pooled out-of-fold predictions for 666 recordings.



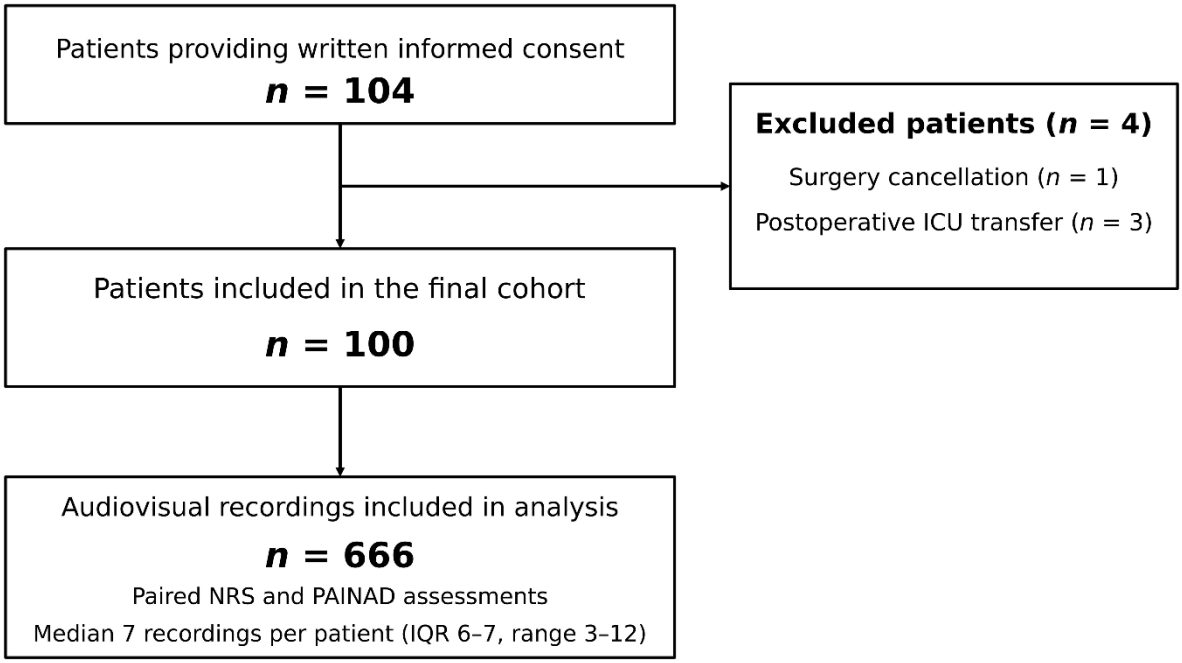
Rows represent reference NRS categories, and columns represent predicted categories. The left panel shows proportions calculated within each reference category, and the right panel shows raw recording counts. Ordinal mean absolute error was 0.288 category steps, with no pain, mild pain, and moderate-to-severe pain coded as 0, 1, and 2, respectively. NRS, Numeric Rating Scale.

Figure S3. Cross-tabulation of paired NRS and PAINAD-K categories for 666 recordings.



Rows represent NRS categories, and columns represent PAINAD-K categories. Cells show raw recording counts and percentages calculated within each NRS category. Scores were grouped as no pain (0), mild pain (1–3), and moderate-to-severe pain (4–10). Exact agreement was 65.6% (437/666), and quadratic weighted kappa was 0.679. NRS, Numeric Rating Scale. PAINAD-K, Korean version of the Pain Assessment in Advanced Dementia scale.

Figure S4. Flow of participant enrollment and audiovisual recordings included in the analysis.



ICU, intensive care unit; IQR, interquartile range; NRS, Numeric Rating Scale; PAINAD, Pain Assessment in Advanced Dementia.