MeFEm: Medical Face Embedding model

Yury Borets [1] BORETSYURY@GMAIL.COM
[1] *Sber AI Lab*

Stepan Botman [1] STEPAN.BOTMAN@GMAIL.COM
[1] *Sber AI Lab*

Editors: Under Review for MIDL 2026

Abstract

We present MeFEm, a vision model based on a modified Joint Embedding Predictive Architecture (JEPA) for biometric and medical analysis from facial images. Key modifications include an axial stripe masking strategy to focus learning on semantically relevant regions, a circular loss weighting scheme, and the probabilistic reassignment of the CLS token for high quality linear probing. Trained on a consolidated dataset of curated images, MeFEm outperforms strong baselines like FaRL and Franca on core anthropometric tasks despite using significantly less data. It also shows promising results on Body Mass Index (BMI) estimation, evaluated on a novel, consolidated closed-source dataset that addresses the domain bias prevalent in existing data. Model weights are available at https://huggingface.co/boretsyury/MeFEm, offering a strong baseline for future work in this domain.

Keywords: self-supervised learning, face embedding

1. Introduction

The rise of large-scale foundational models has demonstrated unprecedented capabilities across language (Naveed et al., 2025) and vision (Awais et al., 2025) tasks, typically trained on broad internet-scale data. A compelling, yet underexplored, application of this paradigm lies in medicine (Khan et al., 2025), where significant progress has been made despite a severe constraint: the scarcity of suitable training data. The utility of a foundational model depends fundamentally on the volume of high-quality data, and the poor generalization of general-purpose models on specialized medical tasks is a well-known challenge (e.g., (Ma et al., 2024)).

Facial images, in contrast, represent a uniquely powerful and accessible modality. They are ubiquitous, easily captured, and rich in physiological information that can reflect systemic health status and potential underlying pathologies. Datasets of facial images annotated with medical labels are typically small, proprietary, or both. This bottleneck arises from two fundamental challenges: the stringent privacy and ethical concerns surrounding the sharing of sensitive biometric data like faces, and the significant cost and expertise required for accurate data labeling, which often involves laboratory tests, specialized instrumentation, or access to confidential medical records.

This work addresses this challenge by proposing a bridge between the two domains: we leverage self-supervised learning (SSL) on large-scale, non-medical facial image collections to build a model that captures robust, medically-relevant features. The core hypothesis is that a model trained to understand the fundamental structure and variation of human faces

through SSL will learn representations that are highly effective for downstream medical and biometric prediction tasks, even with limited labeled data. This approach bypasses the need for direct, sensitive medical labels during pretraining while still aiming to capture the subtle phenotypic markers associated with health states.

2. Related works

2.1. Self-supervised learning

SSL provides a powerful framework for learning representations from unlabeled data by defining a pretext task where the target is generated from the data itself. Early context-based methods, such as predicting image rotations or solving jigsaw puzzles, demonstrated promise but often learned features that did not transfer well to high-level semantic tasks. The field was subsequently dominated by contrastive learning paradigms (e.g., SimCLR (Chen et al., 2020), MoCo (He et al., 2020; Chen et al., 2021)), which learn by maximizing agreement between differently augmented views of the same image while pushing apart views of different images, and self-distillation methods (e.g., BYOL (Grill et al., 2020), DINO (Caron et al., 2021; Oquab et al., 2023)), which avoid negative samples by having a student network predict the output of a teacher network derived from its own past states. While highly effective, these methods can be computationally expensive and require careful management of negative samples or sophisticated asymmetric architectures to avoid collapse.

A more recent and influential shift has been towards generative masked modeling, where a model learns by predicting masked portions of the input. These methods are primarily distinguished by their prediction target. Some, like Masked Autoencoders (MAE) (He et al., 2022) and SimMIM (Xie et al., 2022), perform pixel-level reconstruction. In contrast, our work builds upon methods that perform representation-level prediction, aiming to predict the latent features of masked regions rather than raw pixels. This category includes frameworks like data2vec (Baevski et al., 2022) and the Joint Embedding Predictive Architecture (JEPA) (Assran et al., 2023). These representation-based methods inherit a core principle from contrastive and self-distillation learning—the importance of learning in a structured latent space—but implement it through a predictive, masked-modeling objective. This approach is more computationally efficient and encourages the model to capture higher-level semantics. The JEPA framework, in particular, has proven to be a highly scalable and versatile architecture, forming the foundation for state-of-the-art models in video understanding (Assran et al., 2025) and other modalities. Its ability to learn robust representations by predicting in latent space makes it a premier choice for domains where semantic understanding is paramount over pixel-level fidelity.

2.2. Embedding models

The landscape of computer vision has been reshaped by the release of powerful open-weight foundational models. These models, often Vision Transformer (ViT) (Dosovitskiy, 2020) based, are characterized by their scale, training data, and—critically—their public availability, serving as standardized feature extractors and benchmarks. This ecosystem is divided into distinct paradigms based on their learning objective.

One dominant paradigm relies on language supervision, using image-text pairs. This includes contrastive models like Open-CLIP (Ilharco et al., 2021) and generative Vision-Language Models (VLMs) like CoCa (Yu et al., 2022) and Emu (Sun et al., 2023). While revolutionary for general-purpose reasoning, their representational space is optimized for semantic alignment with text, which may not capture the fine-grained, sub-semantic features crucial for specialized domains.

In parallel, a suite of purely visual foundational models has been established through SSL. This includes models like DINOv2 and I-JEPA, which learn powerful representations from images alone, without any text-based guidance. A notable recent advancement in this line is the Franca model (Venkataramanan et al., 2025), which scales this SSL approach to billions of parameters, establishing a new state-of-the-art for general visual recognition.

The success of these foundational encoders has inspired their adaptation to specialized domains. A key example is FaRL (Zheng et al., 2022), which applies a CLIP-like, image-text contrastive objective to a massive dataset of faces to learn a general-purpose facial encoder. Our work builds upon this progression but pursues an alternative path. We hypothesize that for specialized facial analysis tasks where textual alignment offers no inherent advantage, further specializing a purely visual SSL approach can yield superior performance by focusing the model's capacity on fine-grained visual features rather than text-aligned semantics.

3. Materials and methods

This work employs the standard ViT architecture as its encoder backbone, consistent with the original JEPA implementation and other foundational models. While subsequent hierarchical architectures have been developed to address the fixed-scale receptive field of the vanilla ViT, we demonstrate that a targeted preprocessing strategy effectively resolves this issue for structured facial data. By standardizing input images through rigorous scaling and cropping, we enforce a consistent spatial alignment between facial features and the model's patch grid. This ensures the fixed receptive field becomes a predictable and stable basis for feature extraction rather than a limitation, thereby justifying the use of ViT architecture for our domain.

We treat face detection as a solved preliminary step, utilizing established models to form the foundation of our preprocessing pipeline. In this first stage, we employ the Blazeface model (Bazarevsky et al., 2019) from MediaPipe to obtain a bounding box, which is then used to crop image, ensuring the face is centered and scaled to consistent proportions. This directly addresses the ViT's fixed-scale constraint by creating a stable correspondence between anatomical landmarks and the input patches.

Standard face detection models typically produce a tight bounding box, spanning from chin to brow and cheek to cheek. While computationally efficient, this approach excludes potentially informative regions such as the forehead, ears, and neck. To retain this contextual information, we define our final crop as a square with dimensions twice the maximum dimension (height or width) of the original bounding box, centered on the detected face. This strategy ensures the inclusion of peripheral anatomical features without introducing distortion. In practice, this method produces stable and consistent inputs, with only minor shifts due to variations in camera angle or head pose.

3.1. Data

3.1.1. Training dataset

A primary source of open-access facial data is the FaceCaption-15M dataset (Dai et al., 2024), which provides images alongside textual descriptions (unused in utilized self-supervised approach) and face bounding boxes, which we utilized directly, bypassing the need to run our own detection model. This dataset is distributed as a list of hyperlinks and, as its authors noted regarding the parent LAION-Faces-50M, suffers from inherent link rot problem. From the accessible images, we performed a cleaning procedure that removed samples where a face could not be properly cropped (e.g. due to proximity to the image boundary) or where the resulting crop resolution was below 224×224 pixels. The final subset amounts to approximately one-third of the original, totaling under 4.6 M samples, and forms the backbone of our training set, as detailed in Table 1.

To enrich our training data, we incorporated samples from the AVSpeech dataset (Ephrat et al., 2018), which consists of short video clips of people speaking. These clips are also distributed via hyperlinks and are consequently subject to the same link rot issue. Since each clip features a single speaker, we extracted one frame per clip to maximize data diversity and avoid redundancy. We processed the resulting images using the same pipeline applied to FaceCaption-15M, yielding an additional 1.5 M samples that provide more 'in-the-wild' variation.

As illustrated in Figure 1, both the AVSpeech and FaceCaption datasets contain a minor proportion of irrelevant or mislabeled samples, an inevitable consequence of their scale and automated collection pipelines. We estimate these constitute a small single-digit percentage of the total data. Although filtering these samples would likely improve downstream performance, developing a robust detection method is a non-trivial task that lies beyond the scope of this work.

Finally, we incorporated the SFHQ dataset (Beniaguev, 2022), which contains 0.5 M synthetically generated, high-quality face images at a resolution of 1024×1024 pixels. This dataset provides high variability in age, ethnicity, and facial expression. Since the faces are pre-centered and cropped, the only required preprocessing was downscaling the images to our target resolution.

Dataset name	Original Size	Media type	Usable size
FaceCaption-15M	15 078 935	images	4 572 807
AVSpeech	2 621 845	videos	1 446 778
SHFQ	425 258	images	425 258
		Total	6 444 843

Table 1: Composition of training dataset. The 'usable' refers to samples that were both accessible online and met face-cropping and resolution criteria.

3.1.2. Evaluation datasets

A direct assessment of medical screening proficiency would require datasets of facial images labeled with specific diseases, which are scarce and rarely publicly available. Consequently,

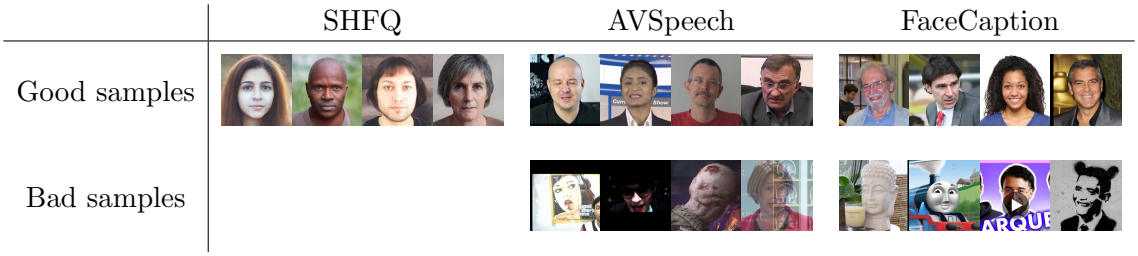


Figure 1: Visual examples from the datasets comprising the training set.

we evaluate a critical prerequisite for such tasks: the accurate prediction of anthropometric and demographic parameters from images. This approach is justified for two reasons. First, several public datasets provide established tasks for age, gender, and ethnicity prediction, enabling a direct comparison of model capabilities. Second, these parameters are themselves well-established biomarkers, strongly correlated with morbidity, mortality, and overall health status. A model that excels at estimating these features demonstrates a core competency that is directly relevant to downstream clinical applications.

For evaluation, we utilize the CelebA dataset (Liu et al., 2015), using its provided attribute labels as prediction targets. For age classification, we employ the FairFace dataset (Karkkainen and Joo, 2021), structuring the task by predicting gender within separate 10-year age bins for white and non-white subject groups.

A core component of our evaluation is the estimation of Body Mass Index (BMI), a physiological parameter of primary clinical importance, alongside age, gender, and ethnicity. Predicting BMI from facial morphology is notoriously difficult, and while state-of-the-art models can achieve a Mean Absolute Error (MAE) within the 2—5 range on specific datasets (Siddiqui et al., 2022), their performance deteriorates significantly under domain shift. This fragmentation across numerous small datasets leads to models that fail to generalize. To address this, we constructed a new, consolidated BMI dataset to enable a more robust and reliable assessment of model performance on this crucial task. The final consolidated BMI dataset comprises three sources: FIW-BMI (Jiang et al., 2019) (7769 images), MCD-rPPG (Egorov et al., 2025) (750 images), and a custom-collected dataset (4784 images). The custom dataset was assembled by curating publicly available images with user-reported height and weight. While this source provides valuable diversity in subject demographics, body types, and imaging conditions, the data cannot be shared publicly due to its nature. Finally, we leverage the instrumentally and laboratory-measured biomarkers provided in the MCD-rPPG dataset to conduct a preliminary assessment of our model's performance on more challenging, clinically-grounded prediction tasks.

3.2. Training process

3.2.1. JEPA basics

The core idea of JEPA approach is predicting representation of some image portion in latent space based on information in the remaining parts of image. That is achieved by splitting set of image patch tokens $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ into two subsets: source $\mathcal{P}_S$ and target $\mathcal{P}_T$ so that $\mathcal{P}_S \cup \mathcal{P}_T = \mathcal{P}$ and $\mathcal{P}_S \cap \mathcal{P}_T = \emptyset$. While split can done in purely random fashion, in

JEPA split is structured so that both subsets contain relatively large continuous segments of original image which allows model to learn not only local features but also long-range dependencies. The source tokens $\mathcal{S}$ and target tokens $\mathcal{T}$ are passed through separate encoders Enc_S and Enc_T, which share same architecture. Final piece of this approach is predictor Pred which reconstruct latent representation of targets tokens from source ones, which can be compared to what was obtained from target encoder:

$$\mathcal{L} = \frac{1}{|\mathcal{P}_T|} \sum_{p_i \in \mathcal{P}_T} \left\| \text{Pred}\left(\text{Enc}_S(\mathcal{P}_S)\right) - \text{Enc}_T(\mathcal{P}_T) \right\|_2^i, \tag{1}$$

where $\mathcal{L}$ is L_2 norm based loss function used in training loop (technically smooth $L1$ is used in practice instead, but this distinction is not essential for the following reasoning). Weights of source encoder and predictor are updated the usual way using error backpropagation, while target encoder weights are updated as exponentially averaged source encoder weights. This stabilizes training dynamics and impede representation collapse.

3.2.2. Probabilistic CLS token assignment

The original implementation utilized vanilla ViT for patch token encoder, with the only notable difference that CLS (class) token being disabled. Omitting CLS token certainly streamlines architecture and training process and while the reason was not explicitly stated in JEPA paper it's a natural choice for fundamental model. While the CLS token is commonly used in transformers for image data and beyond with new methods regularly proposed (Liu et al., 2023; Li et al., 2024) to enrich its expressiveness, we suggest that JEPA training setup provide sufficient pressure for CLS token to learn meaningful representation with only minimal modifications to it.

Let c be CLS token, with complete token set for image denoted as $\mathcal{X} = \mathcal{P} \cup \{c\}$. Since masking strategy operates in term of patch positions — a property CLS token lacks — we propose to append it to either source or target sets in probabilistic manner, independently of image patches tokens:

$$\mathcal{S} = \begin{cases} \mathcal{P}_S \cup \{c\} & \text{if } b = 1 \\ \mathcal{P}_S & \text{if } b = 0 \end{cases}, \quad \mathcal{T} = \begin{cases} \mathcal{P}_T & \text{if } b = 1 \\ \mathcal{P}_T \cup \{c\} & \text{if } b = 0 \end{cases}, \tag{2}$$

where b is obtained by sampling from Bernoulli distribution $b \sim \text{Bernoulli}(P_{c \in \mathcal{S}})$ where $P_{c \in \mathcal{S}} \in [0, 1]$ — probability of assigning CLS token to source set. The loss is then calculated as usual:

$$\mathcal{L} = \frac{1}{|\mathcal{T}|} \sum_{t_i \in \mathcal{T}} \left\| \text{Pred}\left(\text{Enc}_S(\mathcal{S})\right) - \text{Enc}_T(\mathcal{T}) \right\|_2^i \tag{3}$$

For most of our experiments we use $P_{c \in \mathcal{S}} = 0.5$ as a balanced choice, which enables the CLS token to learn a meaningful global representation without compromising the detailed, localized learning of the patch embeddings.

3.2.3. Axial stripe masking

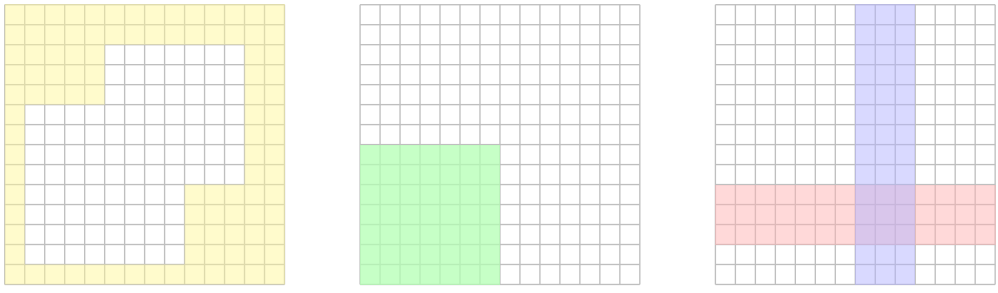
One of the core alterations to the original JEPA approach we introduce stems from the fact that we operate on images structured in a certain way, particularly a face centered

within frame boundaries. This leads to the necessity of introducing domain-specific masking strategy, namely axial stripe masking.

In multiblock masking strategy utilized in JEPA, the final target mask is formed by merging set of rectangular masks with sizes, aspect ration and positions sampled from predetermined intervals (see Figure 2 a). Particularly, two modes are used simultaneously during training: one with small number of large masks and another with large number of small masks. A direct consequence of this strategy is that image patches have non-uniform probability of being assigned to the target subset which becomes function of their spatial positions. Patches near image edges are more likely to be included into source mask as they are less likely to be covered by a randomly placed mask. This is self evident for mask larger than half image in size but remains true even for the case of smaller masks. This positional bias creates a fundamental problem for our domain as the most semantically relevant image portions are concentrated in the center of the image, while background portions that are irrelevant for our purposes are located near edges. While the original JEPA avoids such issues through the use of random crop augmentations, we strive to preserve the consistent spatial structure of face images. This necessitates an alternative, more appropriate masking solution.

The core requirement for the new masking strategy is that the source mask must, by design, contain patches relevant for reconstructing facial representation, with background patches being optional. While quadrant masking can easily fulfill this requirement (see Figure 2 b), it lacks flexibility to control source-target mask ratio during training. This negatively impacts metrics, rendering it suboptimal choice for the problem at hand.

Thus, we introduce a new masking strategy — axial stripe masking — which defines source mask as either horizontal or vertical stripe spanning full image (see Figure 2 c). In contrast to a quadrant mask, its area can be controlled via width parameter and, unlike multiblock masking, these full-height or full-width are designed to consistently intersect semantically relevant regions, ensuring source context always contain meaningful information.



(a) (b) (c)

Figure 2: Schematic representation of different masking strategies: multiblock (a), quadrant (b) and, axial stripes (c). Colored regions represent source the mask; remaining part is the target.

We enhance this approach even further by sampling stripe positions i_c from Gaussian distribution centered on the image midpoint:

$$i_c = \text{round}\left(\mathcal{N}_{[0,L-1]}\left(\frac{L-1}{2}, L^2\,k^2\right)\right) \tag{4}$$

where $\mathcal{N}_{[a,b]}\left(\mu, \sigma^2\right)$ — is truncated normal distribution, $L = 14$ is number of patches per axis, and coefficient $k = 0.175$ was chosen empirically.

During training, the stripe's direction and position are chosen randomly for each sample, while width is fixed at 3 patches, a value found to be optimal. When CLS token is appended to the source set, the last image patch is dropped from the mask to maintain consistent shape across the batch, which provides a slight training speedup. Furthermore, the dropped patch is intentionally located at the image border making it safe to remove.

3.2.4. Circular loss weighs

The final improvement we propose aims to reduce the contribution of irrelevant patches to the total loss. While we lack per-patch annotations for our unlabeled data, and full image segmentation, while possible, would be computationally expensive, we argue for a more efficient alternative.

The argument is that significant advantage can be achieved by down-weighting the loss from irrelevant patches, even via a crude approximation. We propose to do this by weighting each patch's loss based on its distance from image center:

$$\mathcal{L} = \frac{1}{|\mathcal{T}|}\sum_{t_i \in \mathcal{T}} w_i \cdot \left\| \text{Pred}\left(\text{Enc}_S(\mathcal{S})\right) - \text{Enc}_T(\mathcal{T}) \right\|_2^i \tag{5}$$

where w_i — element of weight matrix and $w_i = 1$ for $t_i \equiv c$. For the weight matrix construction, we selected sigmoid function because it provides a flat region for center patches and smooth falloff towards the edges:

$$w(r) = \frac{1}{1 + \exp\left(\sigma(r_0 - r)\right)}, \tag{6}$$

where we implied that $w_i = w(x, y) = w(r)$ and $r = \|x - x_0, y - y_0\|_2$ is distance from image center (x_0, y_0).

3.3. Evaluation process

In order to evaluate the advantages of our proposed model MeFEm, we compare it against three strong baselines: FaRL (ViT base) and Franca (ViT large) and Buffalo (ResNet50). FaRL is the natural choice as it is the only existing foundational model specifically designed and trained for target domain. Franca represents the state-of-the-art general-purpose image foundational models; to compensate for inherent domain shift, we use its large variant instead of the base one. Buffalo refers to the encoder from the Buffalo-L model pack from InsightFace and represents a classic baseline. It is based on a ResNet50 architecture and has been trained using ArcFace (Deng et al., 2019) on the WebFace600K dataset. In terms of

parameter count, it is closer to the small variants of ViT. All ViT-based models, including our own, use an identical input resolution, which further simplifies the comparison.

To evaluate the quality of representation on downstream tasks, we employ two probing methods: an attentive pooler (following the JEPA implementation) for per-patch embeddings, and a perceptron with a single hidden layer for CLS token only input case.

4. Results

The results for all models on the three downstream evaluation tasks are presented in Table 2. MeFEm model demonstrates a consistent advantage over the baselines across all tasks. Notably, the performance of CLS token based probing is only a few percentage points lower that the full attentive pooler, while significantly outperforming baseline models. In this table, as in all subsequent ones, both the 'Patches' and 'Patches + CLS' configurations use the same attentive poller. The sole difference is whether CLS token is included in the input.

Model	Size	Features	BMI dataset		CelebA	FairFace (age)	
			r^2	MAE	Acc.	Acc. (W)	Acc. (NW)
Franca	Large	Patches + CLS	0.425	3.744	0.904	0.538	0.529
FaRL	Base	Patches	0.380	3.924	0.891	0.488	0.478
FaRL	Base	CLS	0.445	3.604	0.904	0.563	**0.574**
Buffalo	Small*	CLS	0.477	3.578	0.888	0.498	0.522
MeFEm	Base	Patches	**0.506**	**3.453**	**0.910**	**0.572**	**0.575**
MeFEm	Base	CLS	0.488	3.479	0.907	0.550	0.560
MeFEm	Base	Patches + CLS	0.496	3.495	**0.910**	**0.570**	**0.576**

Table 2: Metrics for core downstream tasks compared to baseline models. Note: * indicates the approximate parameter count for a comparable non-ViT model.

In next step we tested MeFEm model by predicting essential medical parameters from single frame of MCD-rPPG dataset and compared results against baseline provided in (Egorov et al., 2025). As expected, the model failed to predict cholesterol levels and diastolic pressure. For glycated hemoglobin and oxygen saturation, the results indicate some meaningful predictive capability, while metrics for total hemoglobin and systolic pressure warrant cautious optimism. We interpret these findings as follows: while a single image frame should be insufficient to directly predict parameters acquired via laboratory analysis or specialized measuring instrument, model appears to learn some associations between visual appearance and underlying pathological states, trying to discern subtle changes linked to different types of physiological abnormalities. The baseline in Table 3 is based on remote photopletismography (rPPG), which reconstructs the pulse wave from the entire video sequence. The relatively low metrics for both baseline and MeFEm emphasize the inherent difficulty and potential limits of solvability of this problem from visual data alone.

5. Discussion

Our evaluation demonstrates that MeFEm model surpasses conventional baselines in estimating basic biometric parameters, indicating its potential sustainability for medical-related

Model	Size	Features	Hemoglobin (mmol/l)		Glycated hemoglobin (%)		Cholesterol (mmol/l)	
			r^2	MAE	r^2	MAE	r^2	MAE
MeFEm	Base	Patches	0.233	0.948	0.089	0.47	-0.060	0.621
MeFEm	Base	CLS	0.263	0.952	0.076	0.49	-0.084	0.640
MeFEm	Base	Patches + CLS	0.234	0.959	0.055	0.48	-0.015	0.623
Baseline	—	Video superpixels	—	—	—	0.41	—	0.60

Model	Size	Features	Systolic pressure (mmHg)		Diastolic pressure (mmHg)		Saturation (%)	
			r^2	MAE	r^2	MAE	r^2	MAE
MeFEm	Base	Patches	0.136	10.8	-0.060	7.15	0.073	0.8
MeFEm	Base	CLS	0.058	11.8	-0.090	7.20	-0.557	1.3
MeFEm	Base	Patches + CLS	0.115	11.2	-0.054	7.14	0.081	0.8
Baseline	—	Video superpixels	—	12.8	—	8.39	—	0.9

Table 3: Metrics for medical parameter prediction from a single frame. Note: the video-based rPPG baseline results are reproduced as-reported from (Egorov et al., 2025) (train/test split unknown).

applications. While this potential requires validation through further experimentation, we argue that the proposed model can serve as new baseline for relevant studies, either as a frozen feature extractor or as a foundation for finetuning. Consequently, the performance uplift cannot be attributed to the training dataset scale.

The core distinction between approach and it's closest analog, FaRL, lies in the training methodology. FaRL employs a two-stage process: initial contrastive image-text alignment (CLIP-like training) followed by a masked image reconstruction task. While this second stage shares conceptual ground with JEPA ideas, FaRL implements it through a separate autoencoder, calculating loss in that dedicated latent space rather than within the encoder's representation. We contend that for medical applications, where predictive power is paramount, the FaRL framework has an inherent limitation. Its initial contrastive stage is constrained by the quality and specificity of its text captions. While subtle physiological cues can be described in principle, the captions available for large-scale web-scraped datasets like LAION almost never contain this level of medical detail. Consequently, the model is optimized to align images with generic descriptions, which can dilute its focus on subtle, clinically relevant features. A pure self-supervised objective bypasses this constraint by learning directly from the visual data without relying on this intermediate, often irrelevant, textual signal.

Although we incorporate additional data sources, the final dataset for MeFEm remains substantially smaller than the baseline's ones — several time smaller than FaRL's FaceCaption-15M and about two orders of magnitude smaller than LAION-Faces-50M used for Franca. Since all training sets derive from LAION 5B (LAION, 2024), the efficiency gains must stem from our architectural modifications rather than data superiority.

Appendix A. Ablation studies

For ablations studies, we selected BMI prediction as the most indicative task and used the small variant of MeFEm model to reduce computational cost. This work introduces several modifications to the classical JEPA pipeline: an axial stripe masking strategy, a per-patch loss weighting scheme, and a probabilistic assignment of CLS token. We ablate the impact of each component, with results summarized in the upper section of Table 4. The lower section of the table presents results for different values of $P_{c\in\mathcal{S}}$ probability.

Model	Size	Features	Masking	$P_{c\in\mathcal{S}}$	Weighted loss	BMI	
						r^2	MAE
MeFEm	Small	Patches	Stripes 2/14	—	Circular	0.291	4.206
MeFEm	Small	Patches	Stripes 3/14	—	Circular	**0.447**	3.666
MeFEm	Small	Patches	Stripes 4/14	—	Circular	0.368	4.001
MeFEm	Small	Patches	Quadrants	—	Circular	0.305	3.698
MeFEm	Small	Patches	Multiblcock	—	Circular	0.371	3.961
MeFEm	Small	Patches	Stripes 3/14	0.5	Uniform	0.321	4.070
MeFEm	Small	CLS	Stripes 3/14	0.5	Uniform	0.169	4.693
MeFEm	Small	Patches + CLS	Stripes 3/14	0.5	Uniform	0.330	4.066
MeFEm	Small	Patches	Stripes 3/14	0.0	Circular	0.388	3.752
MeFEm	Small	CLS	Stripes 3/14	0.0	Circular	0.353	4.084
MeFEm	Small	Patches + CLS	Stripes 3/14	0.0	Circular	0.378	3.697
MeFEm	Small	Patches	Stripes 3/14	0.5	Circular	0.398	3.631
MeFEm	Small	CLS	Stripes 3/14	0.5	Circular	0.430	3.781
MeFEm	Small	Patches + CLS	Stripes 3/14	0.5	Circular	0.437	**3.625**
MeFEm	Small	Patches	Stripes 3/14	1.0	Circular	**0.444**	3.631
MeFEm	Small	CLS	Stripes 3/14	1.0	Circular	0.389	3.939
MeFEm	Small	Patches + CLS	Stripes 3/14	1.0	Circular	**0.446**	**3.627**

Table 4: Ablation results. Note: models in top/bottom sections were trained for 33/66 epochs respectively.

The ablation results indicate that the proposed axial stripe masking strategy is superior to both multiblock and quadrant masking alternatives. Furthermore, the chosen strip width of 3 patches represents an optimum value, as performance degrades significantly with either decrease or increase from this value. As expected, disabling the circular loss weighting also have negative impact, with CLS-only downstream metrics exhibiting a severe degradation. This collapse in performance occurs because the CLS token is now forced to encode highly variable background information, a task that offers no theoretical or practical advantage for the target objective.

Considering probability $P_{c\in\mathcal{S}}$ values, apart from balanced default value of 0.5 we additional considered two extremes: assigning CLS token exclusively to source set or using it only as prediction target. Although models trained with $P_{c\in\mathcal{S}} = 1$ yield the highest raw metrics, they do so by sacrificing the predictive capability of the CLS token. In contrast, the value of $P_{c\in\mathcal{S}} = 0.5$ produce a more balanced model, with performance for patches embeddings and

the CLS token approximately on par, while the case of $P_{c \in \mathcal{S}} = 0$ produces subpar results. We see value in the balanced approach, as it allow for reduction in both downstream model size and training time while providing competitive results. It may have additional merits, which which we will explore in the discussion.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pages 1298–1312. PMLR, 2022.

Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.

David Beniaguev. Synthetic faces high quality (SFHQ) dataset, 2022. URL https://github.com/SelfishGene/SFHQ-dataset.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

Dawei Dai, YuTang Li, YingGe Liu, Mingming Jia, Zhang YuanHui, and Guoyin Wang. 15M multimodal facial image-text dataset. *arXiv preprint arXiv:2407.08515*, 2024.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Konstantin Egorov, Stepan Botman, Pavel Blinov, Galina Zubkova, Anton Ivaschenko, Alexander Kolsanov, and Andrey Savchenko. Gaze into the heart: A multi-view video dataset for rPPG and health biomarkers estimation. *arXiv preprint arXiv:2508.17924*, 2025.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021.

Min Jiang, Yuanyuan Shang, and Guodong Guo. On visual BMI analysis from facial images. *Image and Vision Computing*, 89:183–196, 2019.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

LAION. Releasing Re-LAION-5B: transparent iteration on LAION-5B with additional safety fixes. https://laion.ai/blog/relaion-5b/, 2024. Accessed: 30 aug, 2024.

Zhong-Yu Li, Yu-Song Hu, Bo-Wen Yin, and Ming-Ming Cheng. Multi-token enhancing for vision representation learning. *arXiv preprint arXiv:2411.15787*, 2024.

Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. A simple romance between multi-exit vision transformer and token reduction. In *The Twelfth International Conference on Learning Representations*, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Hera Siddiqui, Ajita Rattani, Karl Ricanek, and Twyla Hill. An examination of bias of facial analysis based BMI prediction models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2926–2935, 2022.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Spyros Gidaris, Elias Ramzi, Andrei Bursuc, and Yuki M Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. *arXiv preprint arXiv:2507.14137*, 2025.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709, 2022.