HAIFAI: Human-AI Interaction for Mental Face Reconstruction

FLORIAN STROHM, University of Stuttgart, Stuttgart, Germany
MIHAI BÂCE, KU Leuven, Leuven, Belgium
ANDREAS BULLING, University of Stuttgart, Stuttgart, Germany

We present HAIFAI—a novel two-stage system where humans and AI interact to tackle the challenging task of reconstructing a visual representation of a face that exists only in a person's mind. In the first stage, users iteratively rank images our reconstruction system presents based on their resemblance to a mental image. These rankings, in turn, allow the system to extract relevant image features, fuse them into a unified feature vector and use a generative model to produce an initial reconstruction of the mental image. The second stage leverages an existing face editing method, allowing users to manually refine and further improve this reconstruction using an easy-to-use slider interface for face shape manipulation. To avoid the need for tedious human data collection for training the reconstruction system, we introduce a computational user model of human ranking behaviour. For this, we collected a small face ranking dataset through an online crowd-sourcing study containing data from 275 participants. We evaluate HAIFAI and an ablated version in a 12-participant user study and demonstrate that our approach outperforms the previous state of the art regarding reconstruction quality, usability, perceived workload and reconstruction speed. We further validate the reconstructions in a subsequent face ranking study with 18 participants and show that HAIFAI achieves a new state-of-the-art identification rate of 60.6%. These findings represent a significant advancement towards developing new interactive intelligent systems capable of reliably and effortlessly reconstructing a user's mental image.

CCS Concepts: • **Human-centered computing** → **Collaborative interaction**; **User models**; • **Computing methodologies** → **Reconstruction**;

Additional Key Words and Phrases: mental image reconstruction, faces, user modelling, deep learning

ACM Reference format:
Florian Strohm, Mihai Bâce, and Andreas Bulling. 2025. HAIFAI: Human-AI Interaction for Mental Face Reconstruction. *ACM Trans. Interact. Intell. Syst.* 15, 2, Article 10 (May 2025), 26 pages.
https://doi.org/10.1145/3725891

1 Introduction

Many humans are visual thinkers [2], i.e., they heavily rely on mental imagery in their everyday life—visual representations of objects, faces or concepts only available in people's minds. For visual thinkers, the ability to 'see' scenarios, memories or future possibilities is critical for how

Part of this work was conducted by Mihai Bâce while at the University of Stuttgart.
Authors' Contact Information: Florian Strohm (corresponding author), University of Stuttgart, Stuttgart, Germany; e-mail: florian.strohm@vis.uni-stuttgart.de; Mihai Bâce, KU Leuven, Leuven, Belgium; e-mail: mihai.bace@kuleuven.be; Andreas Bulling, University of Stuttgart, Stuttgart, Germany; e-mail: andreas.bulling@vis.uni-stuttgart.de.

ACM 2160-6463/2025/5-ART10
https://doi.org/10.1145/3725891

they process information, solve problems or make decisions [35, 40, 51, 60, 72]. The prospect of visually reconstructing such mental images using computational methods has long fascinated researchers and has led to substantial research efforts in this area. Mental image reconstruction not only promises to improve our understanding of the human visual system by analysing how visual information is stored in the brain. AI systems that can understand human mental processes also hold significant promise for enhancing human-AI interaction.

Mental image reconstruction as a computational task is profoundly challenging given the complex neural encoding of mental images in the human brain [52]. Prior works have used brain sensing techniques, such as **Electroencephalography (EEG)** [5, 14, 68, 101] or **Functional Magnetic Resonance Imaging (fMRI)** [3, 13, 24, 47, 56, 66, 67, 69, 84, 87]. These methods, though promising, are limited by their invasive nature (EEG) or impractical for everyday use due to their cost and technical complexity (fMRI). Consequently, recent works have instead explored passive monitoring of human gaze for mental image reconstruction. Although gaze has been shown to reflect cognitive processes, such as visual memory [6], and is thus often referred to as a 'window into the mind', previous works have achieved only limited success in terms of reconstruction quality [64, 65, 78, 79].

We propose a more practical approach to mental image reconstruction where a human user and an AI system interactively work together to reconstruct mental images using *active* user feedback. While our approach can, by design, be used with all mental images, we specifically focus on human faces given the importance of face perception, e.g., in social interactions, and given highly relevant practical applications, such as reconstructing a suspect's face from a witness's memory in criminology. Existing methods for facial composite generation can be categorised into constructive, holistic and hybrid. Constructive approaches offer extensive catalogues of facial features for users to choose from, such as different eye, nose and mouth shapes and appearances [11, 18, 42, 45]. The main drawback of constructive approaches is that humans struggle to identify individual features accurately in isolation and instead seem to recall faces holistically [19]. Consequently, holistic methods, such as EvoFIT [20], allow users to assemble entire faces iteratively through evolutionary algorithms. However, guiding a holistic generation can be more challenging than selecting specific features from a catalogue. Hybrid methods (e.g., CG-GAN [99]) combine both approaches' advantages by allowing interactive full-face refinement while maintaining control over specific facial features. However, current hybrid approaches are limited by slow exploration of the high-dimensional face appearance space and in terms of user control.

To address these challenges, we introduce HAIFAI (Figure 1)—an interactive mental face reconstruction system designed to maximise the utility of user feedback without depending on random exploration. Our method involves users iteratively ranking sets of face images based on their resemblance to their mental image. This approach significantly simplifies the users' tasks compared to prior methods, such as CG-GAN, which require users to combine various mechanisms to reconstruct their mental images. Our system then extracts facial appearance information from these rankings over multiple iterations. This information is integrated using an end-to-end, data-driven model that predicts a feature vector that encodes likely facial features. Rather than iteratively searching the face space as done in evolutionary-based algorithms, our system holistically integrates user feedback, using the available information's full potential. To visually decode the mental image, HAIFAI uses a state-of-the-art generative model capable of generating realistic face images [38]. Due to the impracticality of collecting large amounts of human ranking feedback for training HAIFAI, we propose a computational user model to simulate this process. This model uses a pre-trained face identification network [15], fine-tuned on a face similarity task with human labels crowd-sourced from 275 participants via **Amazon Mechanical Turk (AMT)**. This enables the generation of synthetic human rankings to train our system. In the second stage, we use UP-FacE [77] to allow

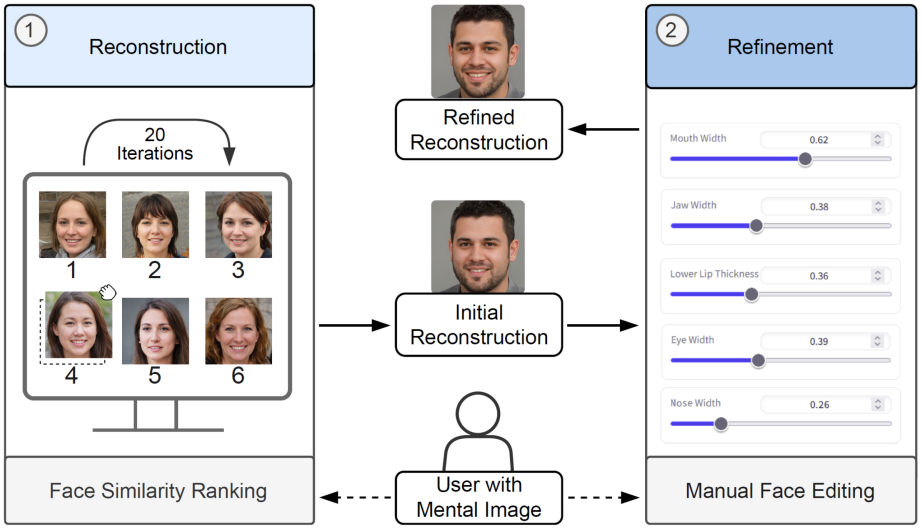


Fig. 1. HAIFAI is an interactive human-AI system designed to reconstruct face images from a user's mental image visually. In this system, users iteratively rank auxiliary face images—random face images of the same age group and sex—based on their perceived visual similarity to the mental image they hold in mind. This feedback is utilised to extract relevant image features, which are then combined by the AI system to reconstruct the user's mental image. Following the initial reconstruction, users can adjust facial features further using a slider interface to refine the visual reconstruction.

users to further refine the initial reconstruction with an easy-to-use slider interface. Providing an initial reconstruction based on user rankings that closely align with their mental image facilitates easier and more efficient subsequent manual editing of the face.

Our work makes the following contributions:

(1) We propose HAIFAI, a two-stage interactive human-AI mental face image reconstruction system. First, our novel method integrates explicit user ranking and reconstructs an initial image. Second, UP-FacE enables fine-tuned control over the facial features of the face reconstruction.
(2) We introduce a computational user model for ranking, trained on a novel crowd-sourced dataset of human face ranking information. This allows us to generate the data required for training the first stage of HAIFAI.
(3) We demonstrate the effectiveness of HAIFAI through evaluations in two user studies, highlighting improvements in usability, reconstruction speed and quality over state-of-the-art methods. Additionally, our system achieves a new state-of-the-art performance for deep learning-based techniques in human identification rates, which is particularly significant for critical applications such as forensics.

2 Related Work

Our work is related to previous works on mental face reconstruction using (1) implicit and (2) explicit user feedback, as well as (3) face editing.

2.1 Mental Face Reconstruction with Implicit User Feedback

Prior research on reconstructing faces using implicit user feedback mechanisms has predominantly relied on EEG, fMRI or gaze data. Pioneering work in fMRI-based face reconstruction by Cowen

et al. [12] involved mapping fMRI signals to principal components of facial structures, also known as Eigenfaces, and reconstructing faces through linear combinations of these components. Nestor et al. [54] have developed a method to infer a facial feature space directly from fMRI data. They employed an SVM to distinguish face identities based on fMRI responses and created template faces for each axis within the derived feature space. These templates were then used to reconstruct faces from fMRI signals through interpolation. This methodology was later adapted by Nemrodov et al. [53] for EEG-based mental face reconstruction. A different approach was introduced by VanRullen and Reddy [87], who used a VAE-GAN architecture to encode facial images into a low-dimensional latent space. They then mapped fMRI responses to this latent space using a simple regression model, allowing for face reconstruction via the VAE-GAN decoder. Remarkably, they could reconstruct faces from fMRI responses even when participants merely thought of a face without visual stimuli. Building on this, Dado et al. [13] have recorded fMRI responses to synthetic faces produced by a pre-trained PGGAN [36]. Similar to VanRullen and Reddy [87], they used linear regression to map fMRI data to the GAN's latent space for face reconstruction. Recent works have used diffusion models as powerful image generation backbones, improving image reconstruction quality. This is accomplished either by mapping the fMRI response onto the latent space of the model [66, 84] or by refining an initial reconstruction from a variational autoencoder [56]. While all of these fMRI and EEG-based methods have demonstrated potential, they are also all limited by their invasiveness, high costs and the challenge of generalising models to unseen users as the relevant brain regions differ in size and neural activity.

To address the limitations of these invasive techniques, Strohm et al. [78] have introduced a gaze-based method for mental face reconstruction. In that study, participants were asked to look at various faces and their gaze patterns were analysed to predict relevant facial features. These features were then combined using a pre-trained decoder to generate mental images. Although less invasive and costly, this first method required prior knowledge of the target face, which is prohibitive for most practical applications. They later developed an improved approach that eliminated this requirement but still relied on a controlled environment of human-like faces and accurate gaze data [79]. Given the challenges of mental face reconstruction using implicit feedback, we propose methods that use explicit feedback instead. Our methods operate in a less constrained domain of real faces and significantly improve reconstruction quality and, thus, practical usefulness.

2.2 Mental Face Reconstruction with Explicit User Feedback

A large body of work has explored the reconstruction of mental images through the use of explicit feedback mechanisms, such as user selection, ranking and manual editing. Early systems for face creation were based on the constructive paradigm, allowing users to select individual facial features from extensive template catalogues [11, 18, 42, 45]. However, such methods were constrained by the holistic nature of human face perception, as individuals struggle to identify isolated features accurately [19]. To address this limitation, Frowd et al. [20] have introduced the popular EvoFIT system that allows for holistic face interpolation within the Eigenface space: Users iteratively select faces that resemble their mental image, and the system generates new faces based on the selected Eigenfaces using an evolutionary algorithm. A similar approach has been proposed by Gibson et al. [22] with added functionality for adjusting for age and facial details, such as wrinkles. Further advances in this field included the deep interactive evolution method by Bontrager et al. [4], which applied an evolutionary algorithm in the latent space of a pre-trained GAN to enhance image quality, and Xu et al. [95] used a GAN conditioned on facial landmarks and iteratively refined the landmarks through user feedback to improve reconstruction quality. Zaltron et al. [99] introduced CG-GAN, which integrates a holistic evolutionary algorithm with constructive functionalities. This approach allows users to modify faces along identified axes within the GAN's latent space using

binary face labels (e.g., *glasses*, *beard*). Lastly, Chiu et al. [9] have proposed a method to explore 1D subspaces of a pre-trained GAN and modify faces using sliders until the desired face was obtained. Our method fundamentally differs in that it learns to interpret user feedback through an end-to-end training process, bypassing the need for traditional evolutionary algorithms or random exploration strategies. This approach enables our system to holistically integrate user feedback, leveraging the contained information more effectively. As a result, we achieved improved reconstructed image quality and reduced reconstruction times.

2.3 Digital Face Editing

Utilising prior work in digital face editing, the second stage of HAIFAI allows users to perform additional manual face edits after the initial reconstruction was obtained. Digital face editing is a challenging but well-established task in computer vision that has received increasing attention, particularly in recent years. Related methods have typically used generative models that can produce high-quality facial images from a latent vector encoding facial features. Face editing is performed by manipulating these corresponding latent vectors. There are several distinct classes of methods, including unsupervised methods, mask-based methods, text-based methods, 3D-based methods and attribute-based approaches. Unsupervised face editing methods decompose generative models' latent space or weights to uncover semantic editing dimensions. GANSpace [27] uses PCA on the latent space, while Shen et al. [71] have decomposed generator weights, and Niu et al. [55] have refined semantic directions. Although these methods do not require labelled data, they are limited in their ability to uncover meaningful editing dimensions in the latent space and often exhibit higher entanglement with other features. This entanglement can lead to unwanted modifications in other facial features. DragGAN [57] allows manual adjustment of facial landmarks, demanding significant user effort and unsuitability for automation. Mask-based techniques condition generative models on face masks for control, such as sketch-based inpainting [7, 8, 61] or mask manipulation edits [23, 46, 48, 74, 81, 82]. These methods, however, require skill and effort to modify the segmentation masks properly. Text-based face editing methods combine mask techniques with text-guided edits [33, 93], such as by integrating generators with the CLIP encoder [59] or by allowing localised edits and broader text prompts [31, 80]. While text-based methods offer a more user-friendly interface and high-level control over many aspects of a face, they lack fine-grained editing precision. 3D-based methods translate a 3D face model to a real image [16, 43, 50, 85, 86] and are great for novel-view synthesis, lighting manipulation or transferring expressions. However, semantic face shape editing necessitates significant 3D modelling efforts. Finally, early attribute-based methods have offered binary control over pre-defined face attributes, but this approach often resulted in unintended changes [10, 25, 49, 94, 96, 100]. Later methods have used attention mechanisms and classifiers for localised edits [21, 29, 30, 44, 83, 91, 100] or SVMs for smooth manipulation [26, 70]. Other research has focused on discovering directions in a generator's latent space for plausible edits [1, 32, 39, 92, 97, 98]. Strohm et al. [77] have recently proposed UP-FacE—a method that uses landmark annotations to find such editing directions in latent space. UP-FacE offers user-predictable and fine-grained control over various shape-based facial features, including the eyebrows, eyes, nose and mouth. These features have been demonstrated to be crucial for face recognition [73].

3 Interactive Mental Face Reconstruction

The objective of mental face reconstruction is to create a visual representation, f_{rec}, of a mental face image f_{m} only present in a person's mind. The goal is to generate f_{rec} such that it resembles the same person identity (id) as best as possible: $f_{\text{rec}} \overset{id}{=} f_{\text{m}}$. Our approach for mental face reconstruction involves a two-step process: First, a human user and our system collaborate to iteratively generate

a face reconstruction based on the user's explicit ranking feedback. In a second step, if desired, the user can easily fine-tune the visual reconstruction using UP-FacE [77] by adjusting various semantic facial features, such as nose width and lip size.

Figure 2 presents an overview of the first step in our approach. The high-level concept involves utilising a generative model G for faces. This model accepts a latent vector w as input, which encodes facial features and subsequently generates a corresponding face image f. The primary objective of our system is to predict a latent vector w_{rec} that encodes the relevant features of a user's mental image. This allows the generative model G to accurately produce a visual representation f_{rec} of the mental image. To generate the image f_{rec} from the latent vector w_{rec} we use a pre-trained StyleGAN2 [38] model—a generative model trained on the FFHQ [37] faces dataset. Importantly, StyleGAN2 first maps the input to a disentangled latent space $\mathcal{W}$, which has a clear advantage over the original input latent space $\mathcal{Z}$: The $\mathcal{W}$ space was shown to exhibit less entanglement [1, 38, 39, 92], resulting in better controllability over the generation process. This disentanglement means that changes to specific latent dimensions in $\mathcal{W}$ tend to correspond to semantically meaningful edits to the generated image, such as modifying expressions, hairstyles or other facial attributes, without inadvertently affecting unrelated features. To generate the initial reconstruction, our system presents the user a selection of n pre-defined auxiliary face images, denoted as $\mathcal{F}_{\text{aux}} = f_1, ..., f_n$, which were generated randomly. The user is asked to rank these auxiliary images based on their resemblance to the mental face image f_{m} they have in mind. HAIFAI uses these rankings to determine the optimal latent vector w_{rec} that, when passed through StyleGAN2's generator G, results in an image f_{rec} that matches the mental face image f_{m}:

$$G(w_{\text{rec}}) = f_{\text{rec}} \overset{id}{=} f_{\text{m}}.$$

The user's ability to accurately rank n images diminishes as the number of auxiliary images increases due to the factorial increase in potential rankings. To address this, we adopt an iterative approach, limiting the number of images per iteration to six. This simplifies the user's task and results in less noisy rankings. In each iteration i, the system presents a different set of six auxiliary images $\mathcal{F}^i_{\text{aux}} = f^i_1, ..., f^i_6$, generated using StyleGAN2 by randomly sampling latent vectors $W^i_{\text{aux}} = w^i_1, ..., w^i_6$.

Given the challenge of comparing and ranking faces across different age and sex groups, our system categorises faces based on sex (female or male) and age (above or below 40 years). This approach is commonly used in mental face reconstruction systems [99]. We employed the InsightFace[1] toolbox to automatically label the sex and age of faces randomly generated using StyleGAN2. This labelling process allowed us to create distinct sets of auxiliary images for each sex and age group. HAIFAI receives the sex and age information as input and uses the corresponding sets of auxiliary images to proceed with the reconstruction process. In the following, we describe each of these components in detail.

3.1 Reconstruction Network

The objective of the reconstruction network is to predict a latent vector w_{rec} such that this vector, when processed by the generator G, produces a face image closely resembling the person identity of the mental image f_m. The architecture of the reconstruction network is shown in Figure 3. The reconstruction network takes as input multiple tuples W^i_{aux}, each containing six auxiliary latent vectors. These vectors within each tuple are ordered according to the user's rankings. For each tuple of latent vectors, we append an additional learnable embedding token of size 512, commonly referred to as the *class token (cls)* [17] and add positional encodings such that the model can extract features based on the ranking of the latent vectors. Each tuple is then processed by a Siamese

[1] https://insightface.ai/projects.

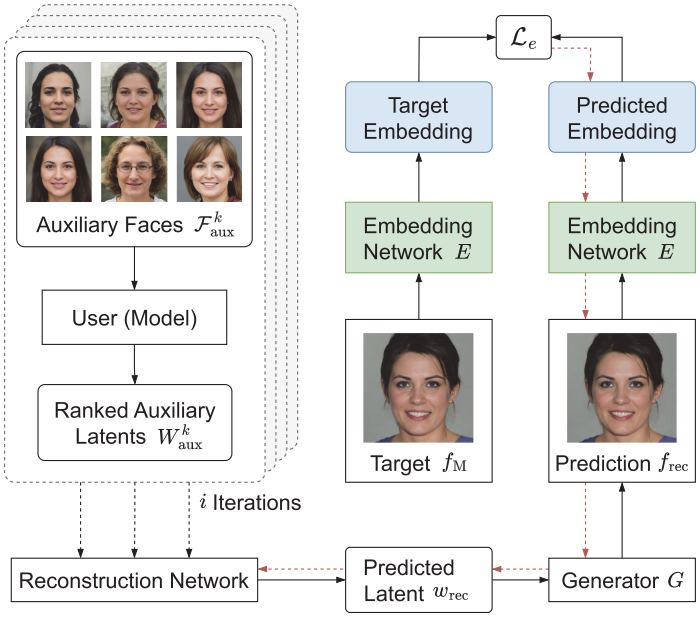


Fig. 2. Our system HAIFAI first shows users sets of auxiliary faces over multiple rounds, asking them to rank these faces based on their resemblance to their mental image. A reconstruction network then predicts the latent vector corresponding to this mental image using the ranked auxiliary latents. A pre-trained generator decodes this vector to recreate the image. The target and reconstructed face images are then passed through a pre-trained embedding network to optimise our reconstruction network based on the similarity of their embeddings. The dotted red arrows indicate the gradient path to train our reconstruction network.

transformer encoder [88], the class token is read out at the end and finally passed through a linear layer, resulting in a 512-dimensional vector containing relevant features for the respective iteration. The feature vectors extracted from each iteration are stacked with another learnable token and passed through a second transformer encoder model. The class token is again extracted at the output of this transformer and passed through a final linear layer, resulting in the reconstructed latent vector w_{rec}.

Solely training the reconstruction network to optimise w_{rec} does not guarantee that $w_{\text{rec}} \overset{id}{=} w_{\text{m}}$ because similar latent vectors do not always result in faces that humans perceive as similar. For instance, Figure 4 shows three faces generated by StyleGAN2. Faces A and B appear more visually similar to each other than faces B and C, despite the mean absolute difference between the latent vectors w_A and w_B being larger than between w_B and w_C. This discrepancy arises because the latent space encodes image features irrelevant to face similarity, such as background, pose and lighting, independent of what makes faces appear similar to a human observer.

To tackle this issue, we aim to optimise our network based on face embedding vectors instead of the latent vectors directly. Models like ArcFace [15] embed faces into a space relevant to identity recognition, as they are specifically trained for this purpose. Unlike the latent space w, using such face embeddings $e_{A,B,C}$ for the faces in Figure 4 results in a mean absolute difference between e_A and e_B that is smaller compared to e_B and e_C. Consequently, we define the loss function for training the reconstruction network as follows:

$$\mathcal{L} = \|\mathbf{w}_{\text{rec}} - \mathbf{w}_{\text{M}}\|^2 - \lambda_{\text{e}} \frac{E(G(w_{\text{rec}})) \cdot E(G(w_{\text{M}}))}{|E(G(w_{\text{rec}}))|\,|E(G(w_{\text{M}}))|}, \tag{1}$$

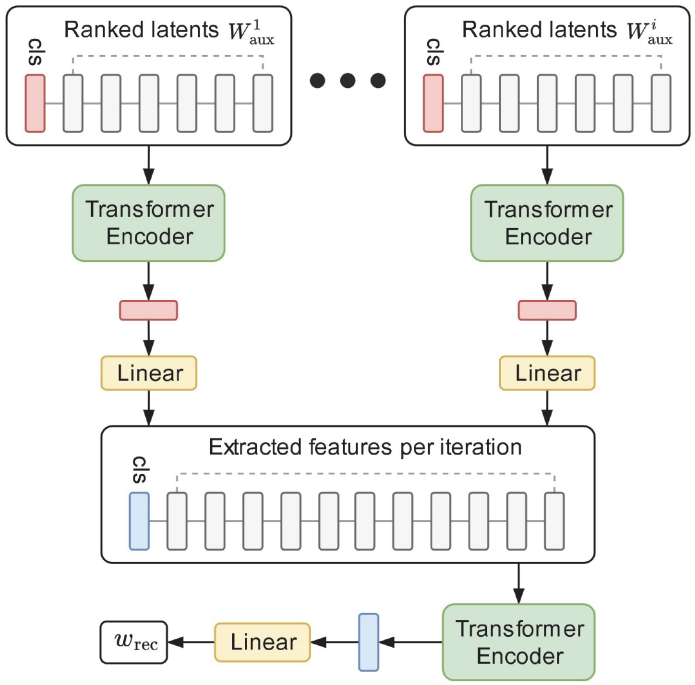


Fig. 3. Architecture of our reconstruction network. The reconstruction network processes a tuple of six auxiliary latent vectors ordered by user rankings for each iteration. For each tuple, a learnable 512-dimensional token (cls) is appended. These tuples are then fed into a Siamese transformer encoder, and the cls tokens are extracted at the end and passed through a linear layer. The resulting feature vectors are combined with another learnable token and processed by another transformer encoder. The class token is extracted again at the output and passed through a final linear layer to produce the reconstructed latent vector of the mental image.

Fig. 4. Three example images generated with a state-of-the-art StyleGAN2 [38] generator. The mean absolute difference between the corresponding latent vectors w_A and w_B is higher compared to the difference between w_B and w_C, although images A and B are visually more similar. However, when using face embeddings extracted with ArcFace [15], the difference between the face embeddings e_A and e_B is smaller compared to e_B and e_C.

where G is a pre-trained generator network mapping latent space to image space, E is a pre-trained face embedding network mapping image space to embedding space, and λ_e is the embedding similarity loss weighting. Thus, the reconstruction network aims to predict a latent vector w_{rec}

similar to the target latent w_{M} that also maximises the cosine similarity between the target and predicted face embedding.

During the training of the reconstruction network, we input a variable number of tuples from which it has to infer w_{rec}. This allows us to support reconstructing the mental image after any number of iterations without changing the network architecture or loss function as in [76], which can negatively impact the reconstruction quality. Moreover, this allows us to automatically stop the reconstruction process once no significant changes to the reconstruction can be observed anymore, thus improving usability and reconstruction times:

$$||w^i_{\mathrm{rec}} - w^{i+1}_{\mathrm{rec}}||_1 < \alpha, \tag{2}$$

where α defines the early termination threshold. Once the mean absolute difference between two reconstructed latent vectors of consecutive iterations is below this threshold, the iterative process stops, and the latest reconstruction is shown.

3.2 Computational User Model for Face Similarity Ranking

Training our method end-to-end requires collecting and annotating a costly large-scale dataset. For each training sample, a user must memorise a generated face and then rank six auxiliary images over multiple iterations based on their similarity to the memorised face. To avoid costly data collection, we introduce a user model that simulates human ranking behaviour, enabling the quick generation of extensive and realistic training data. The central component of the user model is a face-embedding network that extracts embedding vectors to compute the similarity between two faces. We use this network to define the user model described in Algorithm 1. Given a set of auxiliary faces $\mathcal{F}^i_{\mathrm{aux}}$, their latent vectors W^i_{aux}, a target face f_m, an embedding network E and a noise level σ, the cosine similarities between the target and auxiliary face embeddings are calculated. As can be seen in line 6 of Algorithm 1, we add some noise to the calculated cosine similarities uniformly sampled between $[-\sigma, \sigma]$. Auxiliary faces that look alike have a comparable cosine similarity with the target face. Consequently, adding noise can cause random changes in the ranking of these faces. The motivation for this arises from the observed noise in user rankings, characterised by significant variability in the rankings assigned by different humans to the same faces. This observation is discussed in greater detail in Section 4. Introducing noise into the user model creates a non-deterministic version, which helps to prevent severe overfitting during the optimisation of the reconstruction network. The auxiliary latent vectors are then sorted according to these noisy similarities, ranking the most similar face first, followed by the others in decreasing order of similarity.

Existing face embedding models are primarily trained for face identification tasks [15, 58, 89, 90]. Although these models extract meaningful embeddings, they do not explicitly learn to compare and rank faces. Sadovnik et al. [62] demonstrated that measuring identity does not necessarily equate to measuring similarity, leading to rankings that diverge from human judgement. To better align the user model with human behaviour, we found that it is beneficial to fine-tune a pre-existing face embedding model on a small face similarity dataset derived from human feedback. For fine-tuning, we used a novel dataset we collected (see Section 4) comprising triplets (f_a, f_p, f_n), where f_a is a reference face (anchor), f_p is a face more similar to f_a (positive pair) compared to f_n, which is less similar to f_a (negative pair), as determined by human judgement. Using this dataset, the embedding network is fine-tuned with a triplet margin loss objective defined as

$$\mathcal{L}_{\mathrm{c}} = \max((f_a - f_p)^2 - (f_a - f_n)^2 + m, 0), \tag{3}$$

where m defines the required margin between the positive and negative pairs to achieve a zero loss. This fine-tuning process ensures that the network adjusts the embeddings so that faces perceived as similar by humans are also close in the embedding space.

Algorithm 1: User Model of Face Image Ranking Behaviour: Auxiliary and Target Faces Are Projected into an Embedding Space. The Cosine Similarity between Each Auxiliary Face Embedding and the Target Face Embedding Is Computed, and Random Noise Is Added. These Auxiliary Face Latents Are Then Ordered Based on Their Similarity Scores, with the Most Similar Face Ranked Highest, and the Least Similar Face Ranked Lowest.

Input: Auxiliary faces for the current iteration $\mathcal{F}^i_{\text{aux}}$, corresponding latents W^i_{aux}, target face f_{m}, a face embedding network E, and noise level σ.
Output: Ranked auxiliary latents $(w_{R1}, w_{R2}, \ldots, w_{Rn}), w_R \in W^i_{\text{aux}}$
similarities $\leftarrow$ []
for f in $\mathcal{F}_{\text{aux}}$ **do**
    similarity $\leftarrow$ `cosineSimilarity`$(E(f), E(f_{\text{m}}))$
    noisySim $\leftarrow$ `similarity` $+ \mathcal{U}(-\sigma, \sigma)$
    similarities.append(noisySim)
end for
rankedIndices $\leftarrow$ `argSort`$(-\text{similarities})$
rankedLatents $\leftarrow W^i_{\text{aux}}$[rankedIndices]
return rankedLatents

3.3 User-Based Face Refinement

Upon ranking all tuples of auxiliary images or terminating early based on the criterion defined in Equation (2), users are presented with the initial reconstruction generated by HAIFAI. While a holistic approach to face reconstruction is important, prior works like CG-GAN [99] have shown that a hybrid approach combining holistic and constructive methods can lead to improved reconstruction results. Therefore, users can further refine the face manually using UP-FacE [77], a tool designed to manipulate face images produced by generative adversarial networks. Using the same StyleGAN2 model, we can load the initial reconstructed face obtained after the first stage of HAIFAI into UP-FacE. This tool offers a user-friendly interface, shown at the bottom of Figure 10(a), that displays the current image alongside a set of 24 sliders. Each slider corresponds to a distinct semantic face feature, such as *nose width* or *chin length*, which users can adjust simply by moving the sliders. These 24 semantic face features are defined through 2D facial landmarks, and a pre-trained model has learned to modify the latent vector to reflect the desired changes in the face image. We opted to integrate UP-FacE into HAIFAI due to its capability to facilitate easy, fine-grained and precise control over facial features that are vital for face identification, including the eyes, eyebrows, mouth and nose. Integrating UP-FacE into HAIFAI allows for a more interactive and precise customisation process, enhancing the overall quality of the final reconstructed face image. Users can iteratively adjust the facial features, receiving immediate visual feedback on how each slider manipulation alters the face. This interactive refinement ensures that the final output aligns more closely with the user's expectations.

4 Data Collection

To fine-tune the embedding network as described in Section 3.2 and to evaluate HAIFAI on real human data, we conducted a data collection user study.

4.1 Procedure

The data collection study was conducted online with 408 participants recruited through AMT. Each participant completed 23 trials without time limit, each comprising memorisation and ranking steps. During the memorisation phase, we asked participants to look at a random target face generated by StyleGAN2 [38] until they had memorised it. The target face remained consistent throughout the 23

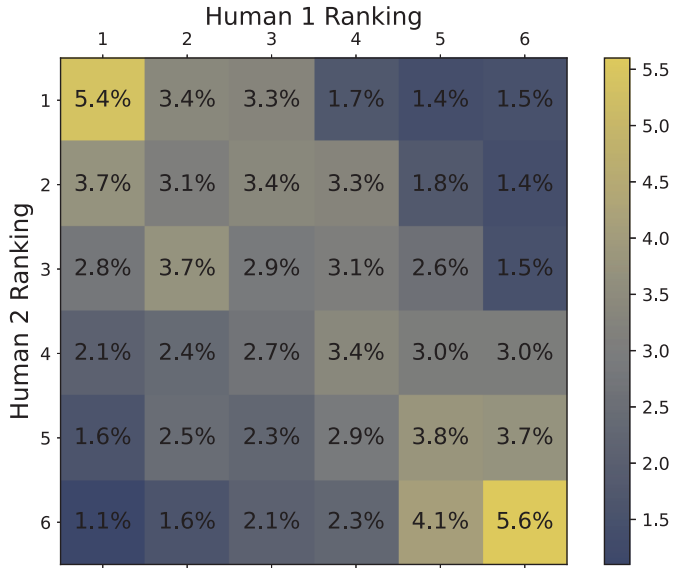


Fig. 5. Agreement in face rankings among humans is shown in this matrix. Each cell (i, j) represents the probability that two separate human raters will assign ranks i and j to the same face. A positive correlation is noted in human rankings, with an average Kendall's Tau value of 0.267. Additionally, humans tend to show higher consensus on the most and least similar faces, whereas the rankings in the middle range exhibit greater variability.

trials, allowing participants to refresh their memory as needed by observing the target face again between trials. After memorisation, participants were shown each set of auxiliary faces, which matched the sex and age category of the target face, and were instructed to rank the six images based on their perceived similarity to the memorised face. Out of the 23 trials, 20 were actual trials, and 3 were attention checks that we added without telling the participants to ensure data quality. For these checks, one auxiliary face was replaced with the actual, ground truth target face. Proper engagement in the data collection was assumed if participants ranked the target face as the most similar in all three attention checks.

4.2 Dataset Statistics

We cleaned the dataset by excluding data from participants who failed at least one attention check. Out of 408 participants, 76 failed all attention checks, 40 failed two and 19 failed one, leaving a total of 275 participants in our dataset. Rejecting approximately one-third of participants based on standard attention checks is typical for AMT data collections [63]. Our dataset included 15 target images for which data from two different participants were collected, enabling the calculation of a ranking agreement between participants. Figure 5 shows the agreement between two participants for each of the six possible ranks. Each cell (i, j) shows the probability that two independent participants assign ranks i and j to the same face. The probability of assigning the same ranks is highest, and significant disagreements are rare. The average Kendall rank correlation coefficient between participants was 0.267 ($p < 0.05$), indicating that humans tend to rank faces similarly, albeit with considerable variability. This variability in the rankings led to our design choice of a noisy, non-deterministic user model as described in Section 3.2.

We randomly selected 75 out of the remaining 275 participants as a validation set to evaluate the performance of HAIFAI on real human data during training. The data from the remaining 200 participants were utilised to generate triplets for fine-tuning the ArcFace [15] embedding network. Out of these, 180 were randomly selected to form the training set, while the data from the remaining 20 participants were used to create the validation set. For each iteration completed by a participant, we generated $\binom{6}{2} = 15$ different (f_a, f_p, f_n) triplets, where $f_a = f_m$ and (f_p, f_n) are all 15 possible pairs of the six ranked auxiliary images. The higher-ranked image in a pair was defined as the positive example f_p, while the lower-ranked image was the negative example f_n. This resulted in 300 triplets per participant (15 pairs for each of the 20 iterations), yielding a total of 54,000 triplets for training and 6,000 for validation.

5 Experiments

5.1 Implementation Details

Embedding Network. For the embedding network in our computational user model and during loss calculation when training the reconstruction network, we used the state-of-the-art face recognition network ArcFace [15]. ArcFace was trained on the IBUG-500K dataset that comprises 11.96 million images and 493K identities. It employs a ResNet50 [28] neural network architecture as its feature extractor, producing 2,048 4×4 feature maps. These feature maps are flattened and fed into an output model that includes batch normalisation [34], dropout (with a 40% drop rate) [75], a fully connected layer with 512 neurons, followed by another batch normalisation layer. The weights of this model were initialised using the pre-trained ArcFace model. We kept the ResNet50 feature extractor frozen while the output model underwent fine-tuning using the collected triplets described in Section 4. The network is trained for 100 epochs with a batch size of 32 using the contrastive loss with a margin of 0.1 defined in Equation (3) using an Adam [41] optimiser with a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Reconstruction Network. Both transformer encoder modules of the reconstruction network consist of four encoder blocks with hidden dimensions of 512 and 8 attention heads each. We add sinusoidal positional encodings [88] to the input of both transformer modules. The intermediate Siamese linear and output linear layers consist of 512 neurons without activation function; non-linearities are only present within the transformer modules. We generated 100K target images for training by randomly sampling latent vectors from a normal distribution and decoding them with StyleGAN2 [38]. Using our sets of auxiliary images and our user model defined in Algorithm 1 with noise $\sigma = 0.22$, we simulated the human ranking of the auxiliary images for each generated target image. The reconstruction network can process a variable number of tuples $i \leq 20$ containing six ranked latent vectors for each of i iterations. We set the maximum number of iterations to 20 and terminated early during test time based on Equation (2) with $\alpha = 0.1$. We used the loss function defined in Equation (1) for training with $\lambda_e = 1$. The model was trained for 100K steps with a batch size of 32 using the Adam optimiser [41] with a learning rate of 0.0001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We used the data from the 75 participants left out of the data collection study for validation and selected the model that achieved the lowest validation loss.

5.2 Reconstruction Study

To evaluate our system, we conducted an additional user study with 12 participants (6 females) between 24 and 54 years old (Mean = 32; SD = 12). Participants were recruited locally among colleagues and friends, and the study design was approved by the university's ethics committee. After giving informed consent, the study started. The study began with explaining the system, allowing participants to familiarise themselves with the process. Once participants were comfortable, a

target face was presented for memorisation. There was no time limit for the memorisation step, but participants could not view the target image again once the experiment had commenced. We used the same pool of target faces as in Strohm et al. [76], allowing us to compare our results to theirs.

The reconstructed mental image after the first stage was displayed after completing all 20 ranking iterations or after HAIFAI stopped early. We then collected data based on the following six evaluation metrics:

- *Mental Rating*: Participants rated the similarity between the memorised image and the reconstruction on a seven-point Likert scale, relying solely on their memory of the target image without viewing it during rating.
- *Visual Rating*: Participants rated the similarity between the target and reconstructed images on a seven-point Likert scale, with both images displayed side by side for comparison.
- *Embedding Similarity*: We calculated the cosine similarity between the embeddings of the target and reconstructed images to assess the reconstruction quality. As the embedding model is fine-tuned to extract similar embeddings for similar looking faces, this metric allows us to quantitatively compare the reconstruction quality of different methods.
- *Task Completion Time*: The time participants took to complete the reconstruction.
- ***System Usability Scale (SUS)***: Participants completed a SUS questionnaire to assess the system's usability, yielding a score from 0 to 100, with higher scores indicating better usability.
- ***NASA Task Load Index (NASA-TLX)***: Participants completed the NASA-TLX questionnaire to evaluate perceived workload across six sub-scales: mental demand, physical demand, temporal demand, performance, effort and frustration. The overall workload score ranges from 0 to 100, with lower scores indicating lower perceived workload.

Following the initial reconstruction, participants were asked what changes would make the reconstructed face more similar to their mental image. In the second stage, participants used UP-FacE [77] to refine the initial reconstruction as much as possible. After this, we again collected data based on the aforementioned six metrics and feedback on possible improvements and tools they would need to enhance the reconstruction further. The total duration of the study was approximately 30 minutes, depending on how long participants took to complete each step.

Figure 6 presents example reconstructions from our conducted user study alongside those from two state-of-the-art deep learning-based methods: CG-GAN [99] and MFRS [76]. CG-GAN is a hybrid reconstruction method that allows users to select and merge faces based on an interactive evolution paradigm, as well as manually edit specific facial attributes. MFRS is our previous holistic face reconstruction system, which requires users to rank sets of faces similarly to HAIFAI. The top row shows the target image that participants had to memorise. The second and third rows display reconstructions produced using the MFRS and CG-GAN methods. The last two rows present results from our method HAIFAI and an ablated version without UP-FacE [77]. Quantitative results from the user study are presented in Table 1. We evaluated the significance of the differences between our methods against the baselines for each metric using either a paired t-test or a Wilcoxon signed-rank test, depending on the normality of the data, as determined by a Shapiro-Wilk test. Differences were considered significant if the Bonferroni–Holm corrected p-value was <0.05. An asterisk (*) indicates a significant difference to the strongest baseline, either CG-GAN or MFRS.

Regarding the *mental rating*, participants rated CG-GAN's reconstructions higher than our method's, with scores of 4.8 compared to 4.2 for HAIFAI without UP-FacE and 4.6 with it. However, our method exceeds the performance of [76], and there is no statistically significant difference between HAIFAI and CG-GAN. Regarding the visual rating, HAIFAI outperforms all baselines. Similar to CG-GAN, we observe a drop from mental to visual rating for HAIFAI; however, this drop

Fig. 6. Example reconstructions from our user study compared with the results from Strohm et al. [76]. Each column shows the results for one participant. The first row shows the target faces participants had to memorise, and the following two rows show the reconstructions of the baselines. The last two columns show the initial reconstruction from HAIFAI after the first stage as well as the results from HAIFAI, where faces were further edited with UP-FacE [77].

Table 1. Comparison of Our Proposed Method HAIFAI with CG-GAN [99] and MFRS [76] Based on Our User Study Results

Method	Mental Rating ↑	Visual Rating ↑	Embedding Sim. ↑	SUS ↑	NASA-TLX ↓	Time (Mins) ↓
CG-GAN [99]	**4.8 ± 0.8**	3.9 ± 0.9	0.36 ± 0.2	59 ± 13	43 ± 11	17.8 ± 5.6
MFRS [76]	4.0 ± 0.8	4.1 ± 0.9	0.38 ± 0.1	85 ± 13	27 ± 18	10.2 ± 4.1
HAIFAI w/o UP-FacE	4.2* ± 0.9	4.3 ± 0.8	**0.43* ± 0.1**	**87 ± 11**	**25 ± 12**	**8.3* ± 3.7**
HAIFAI	4.6 ± 0.8	**4.4* ± 0.8**	**0.43* ± 0.1**	77* ± 13	31* ± 12	11.3 ± 3.3

The best result in each column is highlighted in bold, the second best is underlined. An asterisk (*) indicates a significant difference to the strongest baseline.

is smaller and not statistically significant, unlike for CG-GAN. In our newly introduced embedding similarity metric, HAIFAI significantly outperforms the baselines with scores of 0.43, compared to 0.36 for CG-GAN and 0.38 for MFRS. Further metrics in Table 1 include the average SUS score, NASA-TLX and task completion time. While HAIFAI without UP-FacE outperformed both baselines across these metrics, we can observe a significant performance degradation in these three usability metrics when subsequently using UP-FacE. The SUS decreases from 87 to 77, NASA-TLX increases

from 25 to 31 and the average reconstruction times also significantly increase from 8.3 to 11.3 minutes.

5.3 Lineup Study

Beyond the metrics detailed in Table 1, we performed an additional evaluation to determine the identification rate of our reconstructions through a lineup study, i.e., one of the practical use-cases of our system. This is particularly relevant in fields like forensics, where the key objective may not be a flawless reconstruction of the mental image but rather sufficiently good that it leads to correct identification of the individual. Lineups comprising the true target and similarly appearing faces were created to compute the identification rate. Participants were then tasked with ranking these faces based on their resemblance to their reconstruction. Following prior work [76, 99], the identification rate IR is defined as

$$IR = \frac{\#\text{Rank 1}}{\#\text{Votes}} \times 100. \tag{4}$$

The lineup's composition is critical, as an improper selection can skew the results. If the faces in the lineup are distinctly different from the target, the identification becomes too easy, artificially boosting the identification rate. Zaltron et al. [99] addressed this by introducing noise to the latent vectors of generated target faces to create similar-looking faces. However, generating such variations is more complex since our targets are real faces. Instead, we identified the three nearest neighbours within the FFHQ or CelebA-HQ datasets from which the target faces were chosen. Using the ArcFace [15] embedding space, we selected faces with the closest embedding vectors, excluding different images of the same individual. This approach yielded 24 lineups, each containing four candidate faces paired with reconstructions from both our methods. Example lineups are shown in Figure 7. While this allows us to compare the identification rates of our system with the results from [76], it's important to note that each lineup always includes the target face. This condition, which may not hold in real-world situations, could artificially enhance the identification rates for both methods. We recruited 18 independent raters for an online study. Participants were randomly divided into two groups and completed 12 trials in random order. Each trial involved ranking a lineup based on similarity to reconstructions generated by HAIFAI, either with or without the second stage using UP-FacE, depending on the assigned group. This design ensured that each participant evaluated each lineup only once, mitigating potential biases and enabling cross-group comparison of all reconstructions. The study results showed an identification rate IR of 60.6% for HAIFAI and 59.3% without the second stage. Both of these results mark a statistically significant improvement to the previous leading performance of 56.1% by CG-GAN, as determined by a Wilcoxon signed-rank test with $p < 0.05$. Moreover, participants could rank the target faces within the top three based on our systems reconstructions in 98.5% of the cases, improving over the 95.0% previously achieved by CG-GAN.

5.4 Ablation Experiments

We conducted a series of ablation experiments to evaluate the effectiveness of specific model design decisions. Figure 8 shows Gaussian-smoothed curves for five different models, with the number of batch update steps during training on the x-axis and test-set embedding similarity on the y-axis (higher is better). The baseline model, represented by the brown curve, corresponds to the first stage of HAIFAI without incorporating **Fine-Tuned Embeddings (TE)**, **Variable Iterations (VI)**, or the **Noisy User Model (NUM)**. This model achieves a peak test-set embedding similarity of 0.397 but begins to severely overfit thereafter, as indicated by the decline in test-set embedding similarity with additional updates. The pink line illustrates the performance of the baseline model

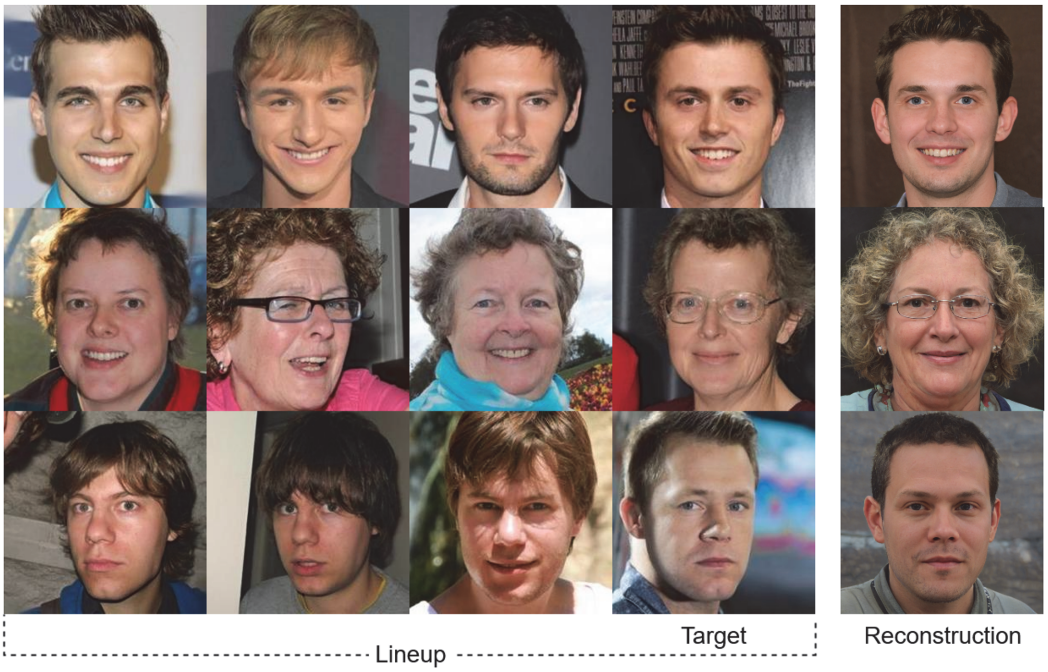


Fig. 7. Example lineups used in our lineup study. Participants had to rank the lineup according to the similarity with the reconstruction generated by HAIFAI.

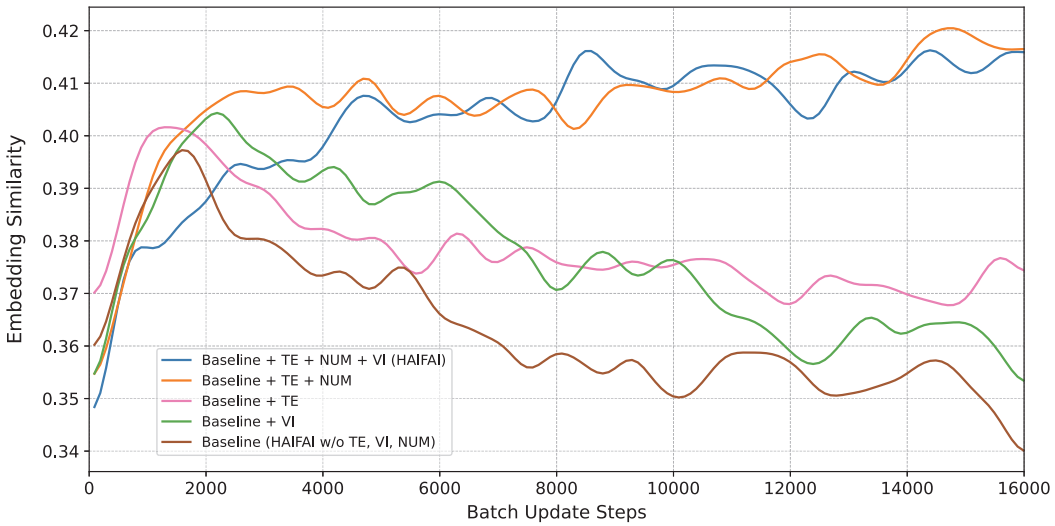


Fig. 8. The figure presents a plot illustrating the embedding similarity (where higher values indicate better performance) as a function of the number of batch update steps during training for various ablated models. The brown baseline curve represents the performance of the first stage of HAIFAI, but with a deterministic user model, without TE, and without VI. It is evident that incorporating VI, TE and a NUM each contributes to improving the overall model performance.

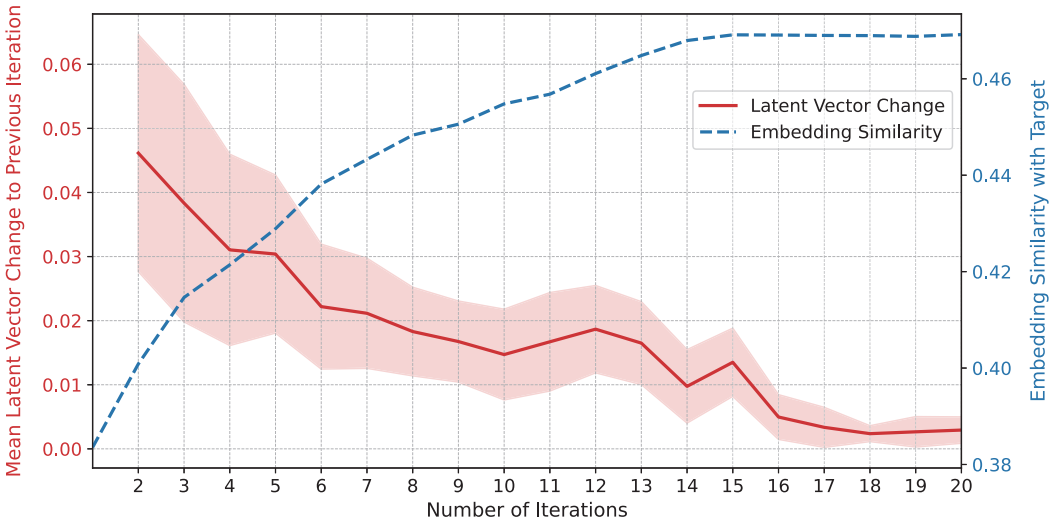
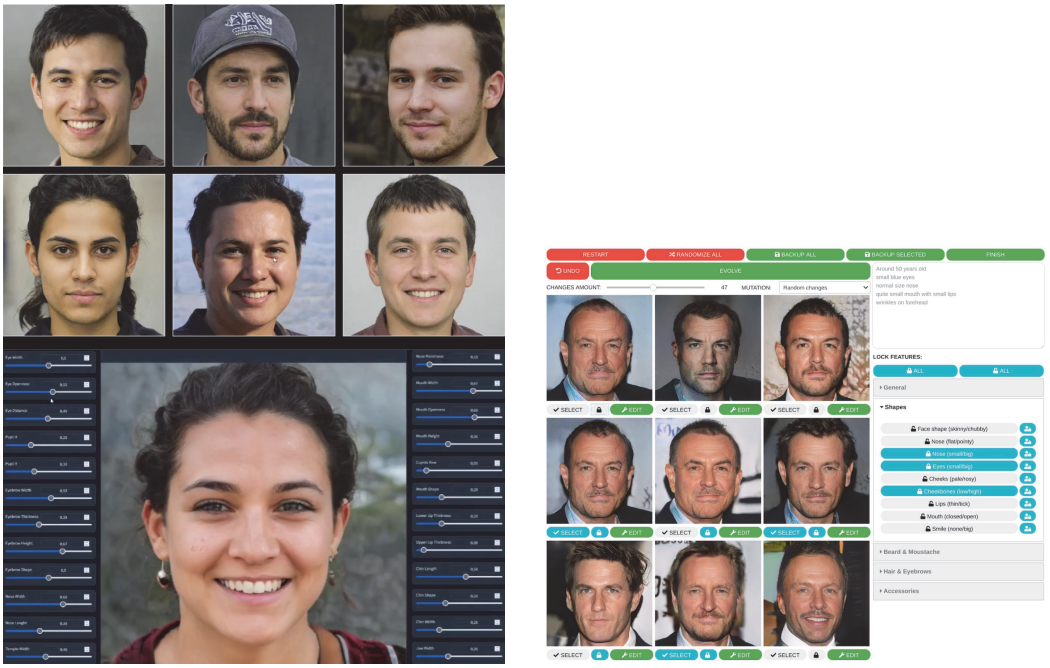


Fig. 9. This figure shows the change in latent vector (red line) as well as the embedding similarity (blue line) over the number of iterations used for reconstructing the image. We observe that the model converges after 15 iterations on average.

when using the embedding network fine-tuned on our collected data, as described in Section 4. We observe a slight improvement in peak embedding similarity to 0.402 and reduced overfitting when utilising TE. The green line represents the results for the baseline model when training with a variable input sequence length instead of a fixed 20 iterations as in [76]. This approach not only permits early termination of the reconstruction process, as described in Section 3, but also appears to regularise the network, resulting in an improved embedding similarity of 0.405 and diminished overfitting. The most significant performance enhancement is achieved by introducing noise into the user model, as described in Algorithm 1. This is evidenced by the orange and blue lines, which no longer overfit the training data and reach a peak embedding similarity of 0.421. The blue line, representing the first stage of HAIFAI used in our evaluation studies, demonstrates that training with a variable number of iterations as input does not negatively impact reconstruction quality. However, it requires a longer training duration to reach a comparable performance. Unlike [76] this allows us to stop the reconstruction at any time without reconstruction quality degradation when using all iterations.

Number of Iterations. The maximum number of iterations for HAIFAI is set to 20, as additional iterations do not further enhance the reconstructions. Figure 9 illustrates the number of iterations on the x-axis, with the mean absolute change of the latent vector compared to the previous iteration on the left y-axis, and the validation set embedding similarity on the right y-axis. As anticipated, the embedding similarity increases monotonically with the number of iterations until it begins to converge around 15 iterations. This trend is consistent with the observed change in the latent vector, which approaches zero after 15 iterations. These observations suggest that, on average, the reconstruction process can be effectively terminated after 15 iterations without compromising reconstruction quality. Based on the observations made in Figure 9, we set the early termination threshold α for Equation (2) to 0.1. The reduction in required iterations through automatic early termination is evident in the decreased reconstruction time presented in Table 1, compared to [76], which utilised a fixed 20 iterations. During our user study, our system terminated after as few as 10



(a) (b)

Fig. 10. Part (a) shows the user interface for the first stage of HAIFAI at the top and UP-FacE at the bottom. For our method, users are iteratively presented with six faces that they can rank via drag and drop (top). Afterwards, users are subsequently fine-tuning the reconstruction with UP-FacE (bottom). Part (b) shows the main interface of CG-GAN. Users are presented with nine faces that they can manipulate via randomising, mutating or manual editing. The interface contains many buttons to lock specific attribute axes and sub-interfaces for different functionalities like manual editing.

iterations and never exceeded 19 iterations. The stopping criterion might be too aggressive, as it immediately halts after no change is observed for a single iteration.

6 Discussion

6.1 Comparison with Baselines

Our user study results (Table 1) revealed substantial improvements in all six metrics for HAIFAI without the second stage compared to the previous state-of-the-art from Strohm et al. [76]. While HAIFAI with the second stage achieves the best reconstruction quality based on the user ratings, it comes at the cost of slightly higher workload, reconstruction times and lower usability. To evaluate reconstruction quality, we used two user-based metrics: *mental rating* and *visual rating*. The *mental rating* gauges users' perceived similarity to the target image based on their memory at the end of the experiment, whereas the *visual rating* involves a side-by-side comparison of the target image and the reconstruction, allowing users to more accurately assess reconstruction quality. Although CG-GAN achieved a higher mental rating (4.8) than our method (4.2/4.6), HAIFAI surpassed both CG-GAN and Strohm et al. [76] in the visual rating. Notably, CG-GAN's score dropped more substantially from mental (4.8) to visual (3.9) rating, while HAIFAI's drop was smaller (4.6 to 4.4) and not statistically significant. This discrepancy between mental and visual ratings may stem from participants' mental image shifting towards the reconstructed face during editing,

a phenomenon we discuss in detail in Section 6.2. In many applications, particularly in forensics, the primary objective is not necessarily to maximise perceived similarity but rather to enhance the probability of correctly identifying the target from the reconstruction. Our additional user study found that HAIFAI achieves an identification rate of 60.6%, which is notably higher than CG-GAN's 56.1%. Given the large sizes of the CelebA-HQ [36] and FFHQ [37] datasets (30K and 70K images, respectively), and our approach of selecting nearest neighbours for challenging lineups, this level of performance is encouraging. Furthermore, in 98.5% of cases, the target face was not ranked last, showing our reconstructions were closer to the true target than at least one of the three nearest neighbours. Interestingly, despite CG-GAN's higher mental rating, our method's higher identification rate suggests the mental rating metric may be skewed by factors unrelated to true similarity. Beyond reconstruction quality, our methods significantly outperform CG-GAN on three key usability metrics: Users gave higher SUS scores, lower NASA-TLX ratings and reported faster task completion times. We attribute these improvements to a more streamlined, ranking-based approach that shifts the complexity of high-dimensional face search from the user to the system.

6.2 Human Factors and Interaction Design

Although the core contribution of HAIFAI is computational, our user studies and participant feedback highlight important human factors and interaction design insights for future mental image reconstruction methods.

Minimising Cognitive Load. Compared to more feature-intensive systems like CG-GAN [99], our ranking-based approach simplifies interaction and reduces cognitive load. In CG-GAN, participants must choose a suitable base face, explore multiple parallel feature spaces and refine attributes individually. This complexity was reflected in CG-GAN's significantly higher NASA-TLX score of 43.4 compared to HAIFAI's 27.2 (Table 1). By contrast, HAIFAI decomposes the reconstruction task into straightforward ranking steps, after which the system takes on the challenge of high-dimensional exploration. This approach aligns with human cognitive patterns, where broad judgments (e.g., 'this face is closer to what I remember') are more easily handled than incremental tuning of many separate attributes.

Structured Task Allocation. A key insight of our work is the clear benefit of a structured, two-stage interaction approach. In the first stage, users only rank sets of candidate faces on overall similarity, offloading the heavy lifting to the AI system, which integrates this feedback holistically. Our participants noted that these small, discrete tasks were less overwhelming and helped maintain a stable mental image. This design contrasts with purely manual editing, where the user must recall many details simultaneously (e.g., nose shape, eye spacing, brow curvature), often leading to frustration. From an interaction-design perspective, our results suggest that incremental user input, such as short ranking tasks, can outperform extensive manual editing for tasks requiring high-quality reconstructions with minimal user burden. Moreover, our SUS results indicate that users appreciate a balanced *division of labour* between themselves and the AI. With HAIFAI, the AI quickly identifies a roughly correct latent space region. The user only steps in at a fine-grained level when they feel confident in adjusting specific details. Interviews and open-text feedback further underscored that participants valued having a manageable number of actions per stage. They felt a sense of agency in the second stage, which is still essential for personal satisfaction, yet welcomed the system's autonomy to handle global exploration. In this sense, HAIFAI underscores a broader principle: Tasks that AI can do quickly and more systematically, like searching for a rough match in a high-dimensional space, are best left to the AI, whereas humans excel at noticing subtle details and making targeted refinements once a decent approximation is available.

Manual Editing. Although the ranking-based reconstruction alone already achieved strong performance, many participants felt it was '*incomplete*' when it came to final subtle adjustments,

particularly for facial features that are highly important and define user identity (eyes, nose, mouth). By integrating UP-FacE [77] as a second stage, HAIFAI becomes a hybrid system that fuses holistic and constructive methods. This inclusion led to improvements in the user-rated similarity, embedding similarity and line-up identification rates, albeit with some tradeoffs: In our study, time to completion increased from 8.3 to 11.3 minutes, and usability ratings dropped from 87 to 77 on the SUS (Table 1). These results indicate a tradeoff: the potential for higher-quality outcomes versus additional user burden and time. In domains like forensics, where accuracy might outweigh ease and speed, participants generally found the second stage worthwhile. When polled on what they wished to modify after the system's initial reconstruction, most participants focused on the most important features: eyes, nose, mouth and general head shape. Secondary interests included hairstyle and colour, beard, age, skin tone and head orientation. Although these secondary features can be changed in real life (e.g., hairstyle), many users still perceived them as relevant for perceived facial similarity. Some participants suggested that future systems might benefit from lightweight text-based editing (e.g., 'make the hair shorter' or 'add a beard') to handle these changeable properties more intuitively. However, participants also expressed concerns about interface complexity, noting that while having more means to modify the face is beneficial, these options should be hidden or optional to avoid causing confusion. Our study results also suggest that some participants would value built-in 'smart suggestions' based on their ranking patterns, e.g., automatic age adjustments or demographic-based style changes if consistently selected. Such adaptivity and deeper mutual understanding between the user and AI could further streamline the process.

Interestingly, the extent of *active* manual editing appears linked to a phenomenon referred to as 'mental shift'. Previous work by Strohm et al. [76] identified a significant drop from mental to visual ratings for CG-GAN [99], where manual editing is central. They attributed it partly to participants unintentionally internalising the edited image as their mental reference. Using their own method, however, this effect was not present, likely because their method does not require active editing. This suggests that methods requiring fewer manual interactions help maintain a more stable memory of the target. Likewise, our current results show that after the first stage, which involves no active face editing, there is even a slight (albeit not statistically significant) increase from mental to visual rating. However, after the second stage, where participants can manually edit face features, we again observe a drop in ratings. Still, it is notably smaller than in CG-GAN, likely because the necessary adjustments in HAIFAI are far less extensive. These findings imply that reducing manual editing demands is beneficial for tasks relying on an intact mental image. A higher degree of active editing can inadvertently shift a person's memory, highlighting a human-factors challenge that goes beyond purely computational concerns. In high-stakes use cases like forensic reconstructions, where accurate memory is paramount, systems should aim to minimise the user's manual intervention or at least offer a structured, short-burst approach to preserve the fidelity of the mental image.

6.3 NUM

Since we are training HAIFAI with simulated human data, achieving high similarity between real human rankings and simulated rankings is essential. Figure 11(a) displays a 6×6 matrix indicating the agreement between our user model and human rankings: Each cell $c_{i,j}$ shows the probability that a human assigns rank i to an image. In contrast, the model assigns rank j. Compared to the human ranking agreement in Figure 5, our computational user model in Figure 11(a) demonstrates rankings that closely align with human judgments. While the human-human Kendall rank correlation coefficient is 0.267, the model-human coefficient is 0.284 ($p < 0.05$), indicating similar average ranking behaviour. We also investigated why injecting noise improves model performance (Figure 8).

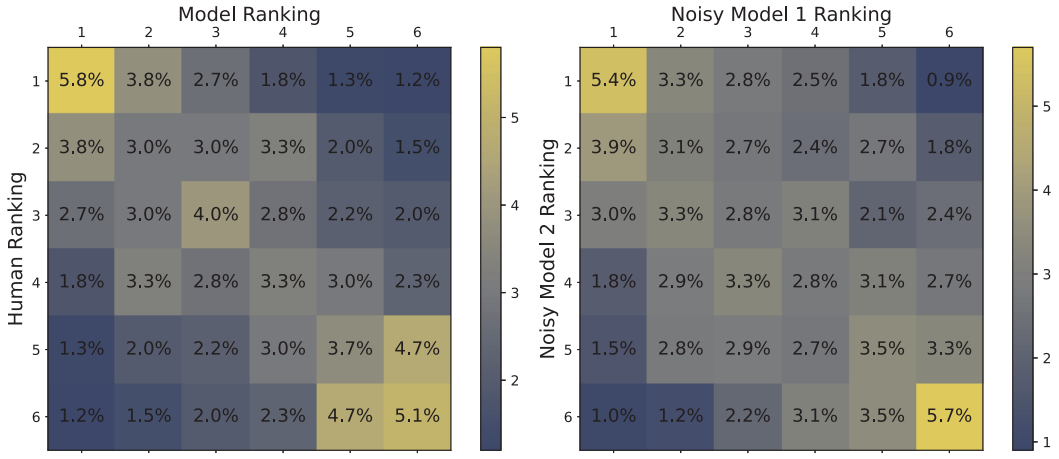


Fig. 11. Face ranking agreement between humans and our user model (a) and agreement between the NUM itself (b). For the latter, we generated two rankings by sampling from the NUM and compared these rankings to assess the internal agreement. Each cell (i, j) shows the probability that human raters/user model assign rank i and the user model rank j to the same face.

A deterministic user model yields a model-model rank correlation of 1.0 by definition and leads to severe overfitting: The network essentially learns to perfectly match an artificially clean distribution, which poorly generalises to real-world data. By injecting noise into the embedding similarities as described in Algorithm 1, we obtain noisy rankings more akin to human variation. We determined the noise level $\sigma = 0.22$ via grid search, minimising the Wasserstein distance between the human-human and model-model ranking distributions. This results in the model-model ranking distribution shown in Figure 11(b), with a Kendall rank correlation of 0.258, which is close to the human-human correlation, thereby making the simulated rankings more realistic.

An important takeaway from our work is that an improved user model improves reconstruction performance. While our evaluation suggests that the current user model already approximates average human ranking behaviour well, as evidenced by the similar human-human and model-human correlation coefficients, there remains potential for further refinement. Human rankings exhibit a degree of noisiness on average, but it might be possible to reduce this variance by better modelling individual user preferences. For instance, instead of using a static user model that represents the average human, a dynamic user model could adapt to the specific ranking behaviour of individual users. This personalised approach may help reduce noise in the simulated rankings, offering the reconstruction network richer and more meaningful signals during training. As a result, it could further enhance reconstruction performance while ensuring robust generalisation to unseen users.

6.4 Limitations and Future Work

A practical limitation of HAIFAI arises not from our method itself but from the biases present in the state-of-the-art generative models on which it builds. Specifically, the StyleGAN backbone used in our approach is trained on well-known facial image datasets (e.g., FFHQ [37], CelebA-HQ [36]) that do not fully capture the global diversity of facial features. Consequently, these datasets introduce biases in the generated faces, particularly if certain racial or ethnic attributes are under-represented.

Due to these biases, our current work has focused primarily on Caucasian faces. However, with a more balanced generative model, our method could easily be extended to include an initial option for selecting a preferred ethnicity, enabling the system to generate and display matching auxiliary outputs accordingly. Moreover, while HAIFAI offers strong baseline performance, certain applications may require even more sophisticated editing of external and stylised attributes (e.g., hairstyle, lighting, accessories). Although UP-FacE already allows refined control of facial features, further integration of attribute-based or language-based editing tools could improve realism at the cost of more interface complexity and user effort. Finally, HAIFAI was tested under controlled lab conditions with *new* faces that participants had only briefly studied. In real-world settings, e.g., generating a facial composite of someone a witness saw weeks before, the user's memory could be incomplete or less accurate. Investigating how different levels of familiarity and recall difficulty impact ranking consistency and final reconstructions is an open area for future work.

7 Conclusion

In this work we introduced HAIFAI—a novel system where human and AI interact to reconstruct the mental image of the user. Unlike previous methods that required users to reconstruct mental images using cumbersome and time-intensive tools, our approach only requires users to iteratively rank images of faces based on their similarity to their mental image. Our system integrates image features across all iterations to visually decode the mental image using a state-of-the-art generative model. In a further step users can optionally further improve the reconstruction manually with an easy-to-use slider interface. We validated our system through extensive quantitative evaluations involving two user studies with a total of 30 participants. The results of these studies showed our system's superior performance in terms of reconstruction quality and time, identification rate, usability and cognitive load. These findings underscore the potential of human-AI interaction in enhancing cognitive tasks, paving the way for future advancements in personalised AI applications. Further research could explore extending this approach to other domains where mental imagery and subjective experience play a crucial role.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. Styleflow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics* 40, 3 (2021), 1–21.

[2] Rudolf Arnheim. 1969. *Visual Thinking*. University of California Press.

[3] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. 2019. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. In *Advances in Neural Information Processing Systems*, 6517–6527.

[4] Philip Bontrager, Wending Lin, Julian Togelius, and Sebastian Risi. 2018. Deep interactive evolution. In *International Conference on Computational Intelligence in Music, Sound, Art and Design*, 267–282.

[5] Andrea Bruera and Massimo Poesio. 2022. Exploring the representations of individual entities in the brain combining EEG and distributional semantics. *Frontiers in Artificial Intelligence* 5 (2022).

[6] Andreas Bulling and Daniel Roggen. 2011. Recognition of visual memory recall processes using eye movement analysis. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 455–464. DOI: https://doi.org/10.1145/2030112.2030172

[7] Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L. Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. 2021. DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control. *ACM Transactions on Graphics* 40, 4 (2021), 1–15.

[8] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. 2020. DeepFaceDrawing: Deep generation of face images from sketches. *ACM Transactions on Graphics* 39, 4 (2020), 72–1.

[9] Chia-Hsing Chiu, Yuki Koyama, Yu-Chi Lai, Takeo Igarashi, and Yonghao Yue. 2020. Human-in-the-loop differential subspace search in high-dimensional latent space. *ACM Transactions on Graphics* 39, 4 (2020), 85–1.

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8789–8797.
[11] Donald F. Christie and Hadyn D. Ellis. 1981. Photofit constructions versus verbal descriptions of faces. *Journal of Applied Psychology* 66, 3 (1981), 358.
[12] Alan S. Cowen, Marvin M. Chun, and Brice A. Kuhl. 2014. Neural portraits of perception: Reconstructing face images from evoked brain activity. *Neuroimage* 94 (2014), 12–22.
[13] Thirza Dado, Yağmur Güçlütürk, Luca Ambrogioni, Gabriëlle Ras, Sander Bosch, Marcel van Gerven, and Umut Güçlü. 2022. Hyperrealistic neural decoding for reconstructing faces from fMRI activations via the GAN latent space. *Scientific Reports* 12, 1 (2022), 1–9.
[14] Hiroto Date, Keisuke Kawasaki, Isao Hasegawa, and Takayuki Okatani. 2019. Deep learning for natural image reconstruction from electrocorticography signals. In *2019 IEEE International Conference on Bioinformatics and Biomedicine*, 2331–2336.
[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.
[16] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. 2020. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5154–5163.
[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805
[18] Hadyn D. Ellis, Graham M. Davies, and John W. Shepherd. 1978. A critical examination of the Photofit system for recalling faces. *Ergonomics* 21, 4 (1978), 297–307.
[19] Martha J. Farah, Kevin D. Wilson, Maxwell Drain, and James N. Tanaka. 1998. What is "special" *about face perception? Psychological Review* 105, 3 (1998), 482.
[20] Charlie D. Frowd, Peter J. B. Hancock, and Derek Carson. 2004. EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on Applied Perception* 1, 1 (2004), 19–39.
[21] Yue Gao, Fangyun Wei, Jianmin Bao, Shuyang Gu, Dong Chen, Fang Wen, and Zhouhui Lian. 2021. High-fidelity and arbitrary face editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16115–16124.
[22] Stuart J. Gibson, Chris J. Solomon, Matthew I. S. Maylin, and Clifford Clark. 2009. New methodology in facial composite construction: From theory to practice. *International Journal of Electronic Security and Digital Forensics* 2, 2 (2009), 156–168.
[23] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. 2019. Mask-guided portrait editing with conditional GANs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3436–3445.
[24] Yağmur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, and Marcel A. van Gerven. 2017. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in Neural Information Processing Systems*, Vol. 30.
[25] Jingtao Guo, Zhenzhen Qian, Zuowei Zhou, and Yi Liu. 2019. MulGAN: Facial attribute editing by exemplar. arXiv:1912.12396. Retrieved from https://arxiv.org/abs/1912.12396
[26] Yuxuan Han, Jiaolong Yang, and Ying Fu. 2021. Disentangled face attribute editing via instance-aware latent space search. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
[27] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANspace: Discovering interpretable GAN controls. In *Advances in Neural Information Processing Systems*, Vol. 33, 9841–9850.
[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
[29] Zhenliang He, Meina Kan, Jichao Zhang, and Shiguang Shan. 2020. PA-GAN: Progressive attention generative adversarial network for facial attribute editing. arXiv:2007.05892. Retrieved from https://arxiv.org/abs/2007.05892
[30] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* 28, 11 (2019), 5464–5478.
[31] Xianxu Hou, Linlin Shen, Or Patashnik, Daniel Cohen-Or, and Hui Huang. 2022. Feat: Face editing with attention. arXiv:2202.02713. Retrieved from https://arxiv.org/abs/2202.02713
[32] Xianxu Hou, Xiaokang Zhang, Hanbang Liang, Linlin Shen, Zhihui Lai, and Jun Wan. 2022. Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. *Neural Networks* 145 (2022), 209–220.
[33] Ziqi Huang, Kelvin C. K. Chan, Yuming Jiang, and Ziwei Liu. 2023. Collaborative diffusion for multi-modal face generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6080–6090.
[34] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456.
[35] Marc Jeannerod. 1995. Mental imagery in the motor context. *Neuropsychologia* 33, 11 (1995), 1419–1432.

[36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196. Retrieved from https://arxiv.org/abs/1710.10196

[37] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

[38] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

[39] Siavash Khodadadeh, Shabnam Ghadar, Saeid Motiian, Wei-An Lin, Ladislau Bölöni, and Ratheesh Kalarot. 2022. Latent to latent: A learned mapper for identity preserving editing of multiple face attributes in StyleGAN-generated images. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 3184–3192.

[40] Minjeong Kim, Jung-Hwan Kim, Minjung Park, and Jungmin Yoo. 2021. The roles of sensory perceptions and mental imagery in consumer decision-making. *Journal of Retailing and Consumer Services* 61 (2021), 102517.

[41] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980

[42] Christine E. Koehn and Ronald P. Fisher. 1997. Constructing facial composites with the Mac-a-Mug Pro system. *Psychology, Crime and Law* 3, 3 (1997), 209–218.

[43] Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. 2020. Config: Controllable neural face image generation. In *16th European Conference on Computer Vision (ECCV '20)*. Springer, 299–315.

[44] Jeong-gi Kwak, David K. Han, and Hanseok Ko. 2020. CAFE-GAN: Arbitrary face attribute editing with complementary attention feature. In *16th European Conference on Computer Vision (ECCV '20)*. Springer, 524–540.

[45] Kenneth R. Laughery and Richard H. Fowler. 1980. Sketch artist and Identi-kit procedures for recalling faces. *Journal of Applied Psychology* 65, 3 (June 1980), 307–316.

[46] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards diverse and interactive facial image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5549–5558.

[47] Yunfeng Lin, Jiangbei Li, and Hanjing Wang. 2019. DCNN-GAN: Reconstructing realistic image from fMRI. In *2019 16th International Conference on Machine Vision Applications*, 1–6.

[48] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. 2021. EditGAN: High-precision semantic image editing. In *Advances in Neural Information Processing Systems*, Vol. 34, 16331–16345.

[49] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. 2018. Attribute-guided face generation using conditional CycleGAN. In *European Conference on Computer Vision (ECCV)*, 282–297.

[50] Safa C. Medin, Bernhard Egger, Anoop Cherian, Ye Wang, Joshua B. Tenenbaum, Xiaoming Liu, and Tim K. Marks. 2022. MOST-GAN: 3D morphable StyleGAN for disentangled face image manipulation. In *AAAI Conference on Artificial Intelligence*, Vol. 36, 1962–1971.

[51] Samuel T. Moulton and Stephen M. Kosslyn. 2009. Imagining predictions: Mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1521 (2009), 1273–1280.

[52] Thomas Naselaris, Cheryl A. Olman, Dustin E. Stansbury, Kamil Ugurbil, and Jack L. Gallant. 2015. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage* 105 (2015), 215–228.

[53] Dan Nemrodov, Matthias Niemeier, Ashutosh Patel, and Adrian Nestor. 2018. The neural dynamics of facial identity processing: Insights from EEG-based pattern analysis and image reconstruction. *eNeuro* 5, 1 (2018), 1–17.

[54] Adrian Nestor, David C. Plaut, and Marlene Behrmann. 2016. Feature-based face representations and image reconstruction from behavioral and neural data. *Proceedings of the National Academy of Sciences* 113, 2 (2016), 416–421.

[55] Yongjie Niu, Mingquan Zhou, and Zhan Li. 2023. Disentangling the latent space of GANs for semantic face editing. *PLoS One* 18, 10 (2023), e0293496.

[56] Furkan Ozcelik and Rufin VanRullen. 2023. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports* 13, 1 (2023), 15666.

[57] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. 2023. Drag your GAN: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

[58] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *the British Machine Vision Conference*.

[59] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of StyleGAN imagery. In *IEEE/CVF International Conference on Computer Vision*, 2085–2094.

[60] Joel Pearson, Thomas Naselaris, Emily A. Holmes, and Stephen M. Kosslyn. 2015. Mental imagery: Functional mechanisms and clinical applications. *Trends in Cognitive Sciences* 19, 10 (2015), 590–602.

[61] Tiziano Portenier, Qiyang Hu, Attila Szabó, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. 2018. Faceshop: Deep sketch-based face image editing. *ACM Transactions on Graphics* 37, 4 (2018), 1–13.

[62] Amir Sadovnik, Wassim Gharbi, Thanh Vu, and Andrew Gallagher. 2018. Finding your lookalike: Measuring face similarity rather than face identity. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2345–2353.

[63] Antonios Saravanos, Stavros Zervoudakis, Dongnanzi Zheng, Neil Stott, Bohdan Hawryluk, and Donatella Delfino. 2021. The hidden cost of using Amazon Mechanical Turk for research. In *HCI International 2021-Late Breaking Papers: Design and User Experience: 23rd HCI International Conference (HCII '21)*. Springer, 147–164.

[64] Hosnieh Sattar, Andreas Bulling, and Mario Fritz. 2017. Predicting the category and attributes of visual search targets using deep gaze pooling. In *IEEE International Conference on Computer Vision Workshops*, 2740–2748.

[65] Hosnieh Sattar, Mario Fritz, and Andreas Bulling. 2020. Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. *Neurocomputing* 387 (2020), 369–382.

[66] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. 2024. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems*, Vol. 36.

[67] Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel A. J. van Gerven. 2018. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage* 181 (2018), 775–785.

[68] Sophia M. Shatek, Tijl Grootswagers, Amanda K. Robinson, and Thomas A. Carlson. 2019. *Decoding images in the mind's eye: The temporal dynamics of visual imagery. Vision* 3, 4 (2019), 53.

[69] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. 2019. Deep image reconstruction from human brain activity. *PLoS Computational Biology* 15, 1 (2019), e1006633.

[70] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. *InterfaceGAN: Interpreting the disentangled face representation learned by GANs. IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 2004–2018.

[71] Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in GANs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1532–1540.

[72] Andrew Sims and Marcus Missal. 2019. Perceptual decision-making and beyond: Intention as mental imagery. In *Free Will, Causality, and Neuroscience*. Bernard Feltz , Marcus Missal, and Andrew Cameron Sims (Eds.), Brill, 13–34. DOI: https://doi.org/10.1163/9789004409965

[73] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. 2006. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE* 94, 11 (2006), 1948–1962.

[74] Linsen Song, Jie Cao, Lingxiao Song, Yibo Hu, and Ran He. 2019. Geometry-aware face completion and editing. In *AAAI Conference on Artificial Intelligence*, Vol. 33, 2506–2513.

[75] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[76] Florian Strohm, Mihai Bâce, and Andreas Bulling. 2023. Usable and fast interactive mental face reconstruction. In *36th Annual ACM Symposium on User Interface Software and Technology*, 1–15.

[77] Florian Strohm, Mihai Bâce, Markus Kaltenecker, and Andreas Bulling. 2024. SeFFeC: Semantic facial feature control for fine-grained face editing. arXiv:2403.13972. Retrieved from https://arxiv.org/abs/2403.13972

[78] Florian Strohm, Ekta Sood, Sven Mayer, Philipp Müller, Mihai Bâce, and Andreas Bulling. 2021. Neural photofit: Gaze-based mental image reconstruction. In *IEEE/CVF International Conference on Computer Vision*, 245–254.

[79] Florian Strohm, Ekta Sood, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2022. Facial composite generation with iterative human feedback. In *the Machine Learning Research*.

[80] Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. 2022. Anyface: Free-style text-to-face synthesis and manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18687–18696.

[81] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. 2022. Ide-3D: Interactive disentangled editing for high-resolution 3D-aware portrait synthesis. *ACM Transactions on Graphics* 41, 6 (2022), 1–10.

[82] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. 2022. Fenerf: Face editing in neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7672–7682.

[83] Qiushi Sun, Jingtao Guo, and Yi Liu. 2022. PattGAN: Pluralistic facial attribute editing. *IEEE Access* 10 (2022), 68534–68544.

[84] Yu Takagi and Shinji Nishimoto. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14453–14463.

[85] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics* 39, 6 (2020), 1–14.

[86] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020. Stylerig: Rigging StyleGAN for 3D control over portrait images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6142–6151.

[87] Rufin VanRullen and Leila Reddy. 2019. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology* 2, 1 (2019), 1–10.

[88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[89] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters* 25, 7 (2018), 926–930.

[90] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5265–5274.

[91] Yi Wei, Zhe Gan, Wenbo Li, Siwei Lyu, Ming-Ching Chang, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. 2020. MagGAN: High-resolution face attribute editing with mask-guided generative adversarial network. In *Asian Conference on Computer Vision*.

[92] Zongze Wu, Dani Lischinski, and Eli Shechtman. 2021. Stylespace analysis: Disentangled controls for StyleGAN image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12863–12872.

[93] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-guided diverse face image generation and manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2256–2265.

[94] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. 2018. Elegant: Exchanging latent encodings with GAN for transferring multiple face attributes. In *the European Conference on Computer Vision (ECCV)*, 168–184.

[95] Caie Xu, Ying Tang, Masahiro Toyoura, Jiayi Xu, and Xiaoyang Mao. 2019. Generating users' desired face image using the conditional generative adversarial network and relevance feedback. *IEEE Access* 7 (2019), 181458–181468.

[96] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. 2021. L2m-GAN: Learning to manipulate latent space semantics for facial attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2951–2960.

[97] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. 2021. Discovering interpretable latent space directions of GANs beyond binary attributes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12177–12185.

[98] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. 2021. A latent transformer for disentangled face editing in images and videos. In *IEEE/CVF International Conference on Computer Vision*, 13789–13798.

[99] Nicola Zaltron, Luisa Zurlo, and Sebastian Risi. 2020. Cg-GAN: An interactive evolutionary GAN-based approach for facial composite generation. In *AAAI Conference on Artificial Intelligence*, Vol. 34, 2544–2551.

[100] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. 2018. Generative adversarial network with spatial attention for face attribute editing. In *European Conference on Computer Vision (ECCV)*, 417–432.

[101] Xiao Zheng, Wanzhong Chen, Mingyang Li, Tao Zhang, Yang You, and Yun Jiang. 2020. Decoding human brain activity with deep learning. *Biomedical Signal Processing and Control* 56 (2020), 101730.

Received 22 September 2024; revised 31 January 2025; accepted 18 February 2025