When the Codec Hallucinates: User Perceptions of Miscompressed Images

Nora Hofer
University of Innsbruck
Innsbruck, Austria
nora.hofer@uibk.ac.at

Rainer Böhme
University of Innsbruck
Innsbruck, Austria
rainer.boehme@uibk.ac.at

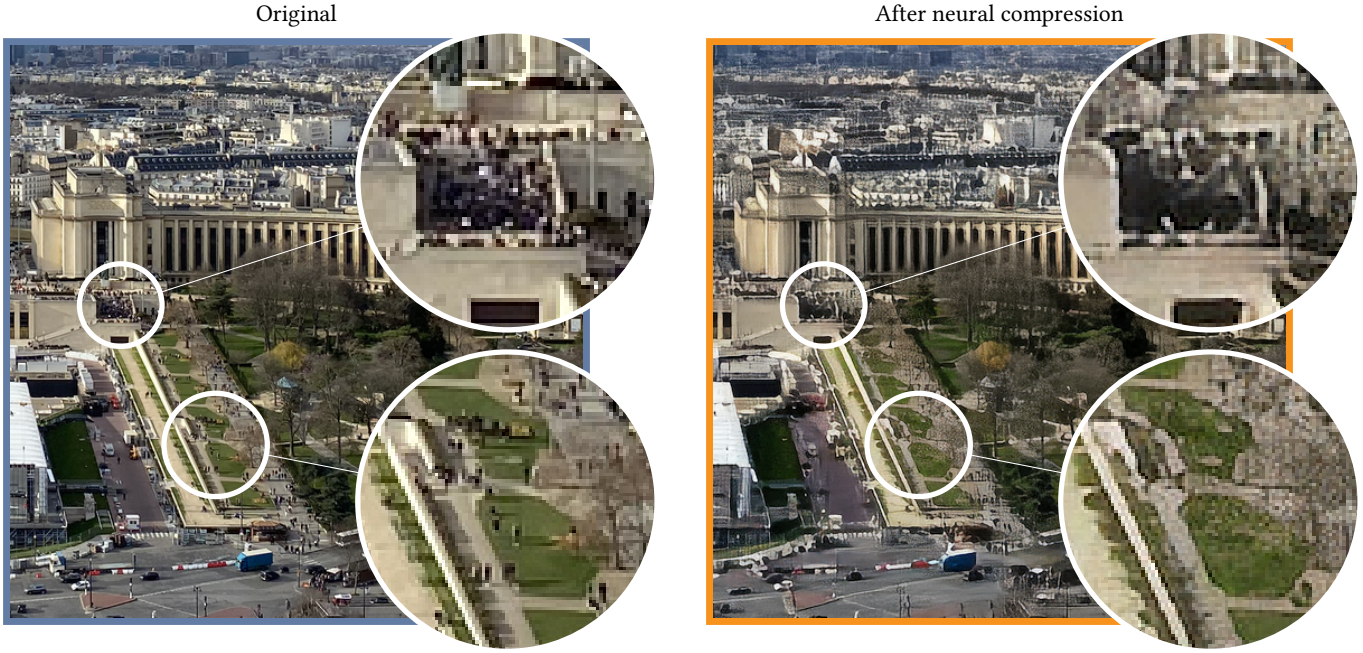


Figure 1: M-Paris. Neural image compression can introduce subtle but sometimes semantically relevant changes to image details. Here, the crowd on the stairs is no longer recognizable and people on the grass disappear (after compression with HiFiC [62] at 0.17 bits per pixel). In our fictional introduction story, such *miscompressions* lead to misunderstandings. In this paper, we study how people perceive these novel compression artifacts. Rather than recognizing them as distortions, participants often interpret them as intentional edits and report an elevated risk of misunderstandings. All figures in this paper are best viewed on screen and in color.

Abstract

People exchange images every day. New methods for image compression leverage neural networks to save bandwidth, but they can undermine the semantic integrity. The term *miscompression* refers to unintended semantic changes of image details, introduced by generative AI during neural (de)compression. Although prior work has speculated about the resulting risks, no empirical evidence exists on how people perceive these novel compression artifacts. In this study, 115 human subjects compared original images with conventionally compressed, neurally compressed, and miscompressed images. Participants perceive that miscompressions elevate the risk of misunderstandings when communicating with images. They also frequently attribute miscompressions to intentional editing, whereas conventional JPEG artifacts are more often recognized as distortions. This paper proposes a method to study this new phenomenon, provides the first empirical evidence of user perceptions of miscompressions, and derives implications for trust in images, as well as interface designs that mitigate the risk.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Reconstruction**.

ACM Reference Format:
Nora Hofer and Rainer Böhme. 2026. When the Codec Hallucinates: User Perceptions of Miscompressed Images. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3772318.3790293

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790293

1 Introduction

Imagine Cassandre, a photojournalist, who covers a breaking news story about a protest in the Trocadéro district of Paris. She takes photos of the barriers, the police presence, and the crowds heading towards the site. She uses a messenger app to send the photos to her colleague Tiresias, who is writing the story in the London newsroom. Cassandre later reads the published article and is shocked. Tiresias describes a small crowd and does not mention the police. The public is outraged, and activists accuse them of misrepresentation. Confused, Tiresias insists that he reported exactly what he saw in the photos. When they compare them with Cassandre's originals, they realize that the versions received are different from the ones sent. What was supposed to be an objective report has become a controversy, caused by *neural image compression*. Our example in Figure 1 suggests that the next generation of image compression methods may make this fictional story a reality.

Compression is essential for the efficient transmission and storage of digital images. The most common methods are lossy. This means they remove information imperceptible to the human eye in order to reduce the file size. While JPEG, a standard from the 1990s, still dominates the web [22], researchers have turned their attention to *neural* image compression. The idea is to replace conventional signal processing operators in the image compression and decompression pipeline with trained neural networks. Early proposals leverage variational autoencoders [6, 7]. State-of-the-art methods use image transformers [5, 91] for the encoder and generative adversarial networks [62] (GANs) or diffusion models [88] for the decoder. These methods achieve unprecedented compression rates at comparable or superior perceptual quality. Wide deployment in consumer devices may be just a matter of time,[1] thanks to new standards like JPEG AI [5].

This development raises concern. Such low bitrates can inconspicuously undermine the integrity of image details. Recently, the signal processing literature has proposed the term *miscompressions* [38] to describe semantic changes of image details caused by neural compression. Conventional methods, such as JPEG, often introduce visible indicators of compression, like blocking [68], blurring [60], or ringing [31, 36] that allow viewers to judge the reliability of an image. Neurally compressed images, however, lack such indicators and tend to appear visually flawless. Thereby they can create a false sense of trust.

While the cause of miscompressions is technical, the consequences are social and may pose risks to humans in various ways. First, miscompressions can lead to the uncontrolled and unintended spread of misinformation, especially when the reconstruction is realistic [46]. Second, miscompressions can change the semantics of images in a way that resembles intentional editing. Therefore, comparisons between the original and the reconstructed images, *e.g.*, in court, by the media, or by insurers, could result in false accusations. A recent CHI paper discovered that viewing images enhanced by artificial intelligence (AI) can alter one's memory of a scene [70]. Eyewitnesses to a scene may fall victim to this effect as miscompressions can resemble semantic changes of AI enhancements. Third, the intended applications of neural compression include downstream computer vision tasks in critical domains, such as public surveillance and autonomous driving [5, p. 103]. Clearly, potential classification errors can cause severe and irreversible harm.

Researching miscompressions is difficult because the definition of a semantic change and its severity are subjective. The existing examples are based on the subjective view of a few researchers [38, 72]. However, semantic understanding differs between individuals, based on different experiences and cultural backgrounds. To illustrate this point, not all authors are in full agreement on how unexpected and severe the missing people in Figure 1 are. This calls for the involvement of a broader set of users to reduce the reliance on individual subjective opinions.

The research question of our empirical study is to find out if a wider population perceives miscompressions of state-of-the-art neural compression methods as concerning. We also want to understand whether the aforementioned risk of miscompressions being mistaken for intentional image editing or manipulation is supported by real users. Finally, we want to measure whether the users are familiar with the visible distortions produced by conventional lossy compression and can attribute the differences to JPEG artifacts. We operationalize the research question in three hypotheses that can be tested with quantitative methods:

- **H1** Miscompressed images are *more likely* to cause misunderstandings than other similar images.
- **H2** The differences between a miscompression and its original are *more likely* attributed to intentional editing than for other similar images.
- **H3** The differences between a miscompression and its original are *less likely* attributed to uncontrollable distortion than for other similar images.

The term "other similar images" refers to images representing the same scene that have been compressed using conventional JPEG or other neural compression codecs which do not result in miscompression for this input. Note that H2 and H3 are not two ends of the same continuum. The hypotheses are conceptually different as they relate to different causes (intention vs. accident), operations (editing vs. compression), and actors (human vs. machine). Moreover, they are neither mutually exclusive, as images can be edited and compressed, nor complete, as images can differ for other reasons, *e.g.*, slightly different camera angle or acquisition time.

To test the hypotheses, we collect standardized responses in a lab study on a sample of 115 users, presenting them a curated set of stimulus images compressed with state-of-the-art neural compression codecs. Using a combined between-subjects and within-subjects design, we get a total of 1 380 image views and 1 131 ratings on image differences. Figure 2 shows the results in a nutshell, supporting all three hypotheses.

In summary, we make the following contributions. i) We design and test an empirical method for a user study on a new phenomenon which can serve as a baseline for future studies on miscompressions. ii) We present evidence confirming that miscompressions are perceived differently from conventional compression artifacts and that

[1] "[T]he first ever implementation [...] of JPEG AI encoder and decoder on their mobile phone" https://www.linkedin.com/posts/touradjebrahimi_wearejpeg-activity-7346065622880976896-Izoh (posted: July 2025; accessed: August 2025)

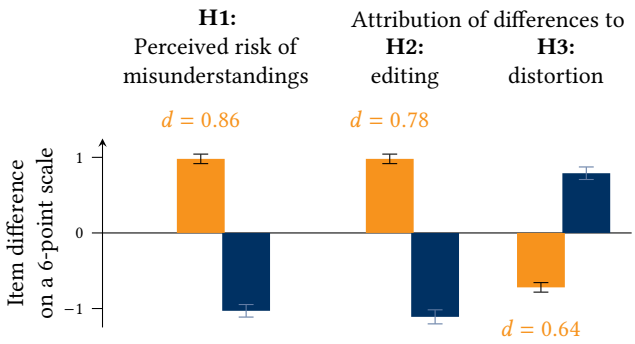


Figure 2: Users perceive that miscompressions (orange bars, left) carry an increased risk of causing misunderstandings. They attribute differences between originals and reconstructions more to intentional editing and less to uncontrollable distortion. The same metrics for JPEG images (blue bars, right) have the opposite direction. Standard errors and Cohen's d for effect size are shown. All results are statistically significant at the $p < 0.001$ level after accounting for the panel data structure with subject and image fixed effects.

they increase the risk of misunderstandings. iii) We discuss implications for the design of future interfaces that enable users to assess and react to the risks posed by miscompressions. The collected primary data is available on Zenodo (DOI:10.5281/zenodo.18311405).

The remainder of this paper is structured as follows. Section 2 reviews research on user perception of digital images. Section 3 documents our method and instrument, Section 4 presents the results, both aggregated and broken down by images. Section 5 discusses the findings and derives implications for interface designs. Lastly, Section 6 concludes.

2 Related Work

The perception of digital images has long been of interest to the research communities of human-computer interaction (HCI), signal processing, and imaging. However, no comprehensive theoretical framework exists: "people's perception of image distortion is complex" [66, p. 3235]. Intervening factors include the image content and the viewer's cultural background. Here, we highlight selected human-factor works that have influenced our design or will help to interpret our results. A summary of the technology behind neural image compression is provided in Appendix A.1.

2.1 User Involvement in Image Compression Research

The role of user studies in the lossy image compression literature is to measure how humans perceive the quality of images after compression. It distinguishes between subjective and objective image quality metrics.

Subjective metrics are collected by asking human observers. They are extensively researched and standards exist [41–43]. In *reference studies*, human observers compare the original (reference) to the compressed image. In *no-reference studies*, the observers are asked to judge the image quality of the compressed image alone. A popular method is called two-alternative forced-choice [33] (2AFC). Participants are shown two images and asked to choose their preferred one. In no-reference studies, 2AFC is applied using different quality settings or codecs.

Objective metrics are mathematically defined. The spectrum ranges from simple distance metrics, such as the signal-to-noise ratio, to perceptually weighted reference metrics, such as SSIM [84], MS-SSIM [85], and FSIM [89]. Recently, learning-based metrics have gained popularity as both reference and, predominantly, no-reference quality metrics [35, 90]. They are part of the loss functions used to train generative AI and neural compression models. While technically "objective", their learned parameters depend on human input, trying to mimic "subjective" human perception. This alignment remains challenging and is an active field of research [14, 34, 71]. One possible cause of miscompressions is the overemphasis on such no-reference quality metrics that improve realism at the expense of fidelity.

In the neural compression literature, user studies typically accompany proposals for new codecs [62, 73]. Their goal is to demonstrate the codec's performance rather than understanding people's perception of neurally compressed images. They are often based on a a small number of viewers and images, and do not include control variables. Semantic differences are acknowledged occasionally, *e.g.*, [73, p. 316], [62, p. 10], but we are not aware of any user study investigating human perception of miscompressions. Tsereh et al. [82] present a dataset of machine-detected JPEG AI compression artifacts with crowdsourced human verification. Their focus is to improve compression performance rather than safeguarding semantic integrity.

2.2 Users' Ability to Detect Authentic Images

An image is considered *not* authentic if the original has been edited to alter its semantics, or if it has been entirely generated by AI [15, Fig. 2]. Therefore, we review the literature on human performance in detecting *image editing*, which relates to our Hypothesis 1, and on detecting *AI-generated images*, since neural compression relies on the same technology.

Image editing. Ostrovsky et al. [67] explore how different lighting configurations in the image influence participants' ability to detect irregularities. Farid and Bravo [24] study the ability to detect forgeries based on irregularities in geometric shades and reflections. Carvalho et al. [21] involve human subjects to validate a forensic forgery detection method based on color classification of scene illuminants. Sun et al. [77] propose a dataset to benchmark the detectability of AI editing operators and involve 35 human subjects to assess the task difficulty. Schetinger et al. [75] crowdsource by asking 400 non-experts to localize suspected forgeries in images. Their participants often relied on contextual cues. Relevant for the selection of our stimuli, they find that images with high structural complexity are harder to evaluate, whereas the size of the edited area does not matter. Experience with digital imaging improved detection capability. Across these studies, one consistent finding emerges: humans' ability to detect image editing is poor, with detection accuracies between 40 and 60 %.

AI image generation. Early image generation technologies were still detectable by humans, as demonstrated repeatedly by Farid and colleagues [23, 25, 40, 58]. A turning point came with advances in deep learning, especially GANs [30]. Lago et al. [55] crowdsource a comparison of GAN-based face generators. The participants were not only unable to detect generated images, but they also misclassified generated images as real more often than the real control images. Follow-up studies have confirmed this negative result and investigated influencing factors. Nightingale and Farid [65] vary the gender and ethnicity of the stimuli and find that white male faces are the least distinguishable, presumably due to biases in the GAN's training data. Frank et al. [27] test a subset of the same face images on a representative sample, confirming the results obtained from convenience samples. Mink et al. [63] explore the effect of identity-based biases. They find that viewers are better at detecting artificial faces if they share the gender or racial identity of the portrayed subject. Wöhler et al. [87] use eye tracking to study the cues participants use to detect face swapping in videos. They find that participants decide based on artifacts, such as blur or unnatural expressions and eye movements. Lu et al. [56] generalize the experiment from faces to all kinds of AI-generated images and ask participants to select from eight cues that may have influenced their judgement. "Detail" and "smoothing", two artifacts common in neural compression, are mentioned most frequently. They also control for participants' experience with generative AI, but the effect is not significant. Finally, Kamali et al. [51] present the probably most comprehensive study. Their collection of 750k data points from 50k subjects, including 35k qualitative comments, allows them to dig deep into potential cues people use to make a decision. General image quality is mentioned most often as a cue, supporting the concern that high realism may create a false sense of trust [46]. Some of their findings may transfer to the perception of miscompressions. Note that neither Lu et al. [56] nor Kamali et al. [51] consider conventional JPEG compression artifacts as cues, which highlights the novelty of our Hypothesis 3.

2.3 Risks of Semantically Distorted Images

The authenticity of images and their effect on the credibility of information has been studied for decades, *e.g.*, for news articles [32, 64], web pages [52], and user-generated content [29, 46, 63]. In the 1990s, researchers even suspected that a mere demonstration of Photoshop, an image editor, could erode viewers' trust in images [54]. While this hypothesis was rejected, it reinforces the idea that experience with editing technologies should be a control variable.

All this research focuses on intentional and controlled edits to the semantics of an image. Miscompressions are a new phenomenon. They compromise the semantic integrity in so far unpredictable and undetectable ways. Research on miscompressions is sparse. Hofer and Böhme define them as discrepancies "between the semantic meaning of an original image (detail) and its reconstructed version after neural compression." [38, p. 3] and collect a human-annotated dataset of miscompressions from different compression codecs [39]. Agustsson et al. [2] propose a technical remedy. Their decoder has a parameter to select between blurry outputs that are closer to the input, and outputs with synthesized image details that are more realistic. Qiu et al. [72] take a closer look at one specific type of miscompression, revealing that the semantic distortions are subject to bias: African–American faces are commonly reconstructed to appear more Caucasian, while Caucasian faces largely retain their original features. The authors demonstrate the bias using a few examples and measure it by passing reconstructed test images to a learned phenotype classifier.

The machine learning community has studied potential challenges for neural compression, regardless of their impact on visible semantics. Chen and Ma [18] warn that malicious actors could scramble image content by exploiting a vulnerability to adversarial perturbations. Madden et al. [57] show that it is possible to gain control over the output by triggering bitstream collisions. Both attacks require control over the input image and detailed information on the codec and its implementation. In non-malicious use cases, the performance of downstream computer vision tasks may degrade [10, 45, 59], especially for iris recognition [10]. Learning-based image forensics can also be affected, including detectors of image manipulation [11, 16, 17], provenance [12], and deepfakes [9, 16]. By contrast, Cardenuto et al. [17] report that the evidential value of medical images in science does not deteriorate at an equal bitrate compared to conventional compression. All of these works process images with machines and do not involve any human viewers. Because the risks of semantically distorted images may affect human perception more than machines, a study on the users' perspective appears overdue.

3 Method

To our knowledge, this is the first user study on miscompressions. Our aim is to validate whether the concept of miscompressions used by researchers matches the understanding of users. To this end, we collected ratings of multiple human subjects on images depicting multiple scenes. This allows us to generalize from the subjective understanding of individuals as well as from the distinctive characteristics of a scene. As miscompressions are defined by differences between images, we opted for a **full-reference study** with one test image displayed next to the reference image at a time. To ensure external validity while not requiring to introduce our subjects to the topic of neural compression, we came up with a **scenario** set in the context of social media. This scenario is handy because it is plausible for an image to undergo unknown processing in transit, including lossy compression, retouching, or manipulation. To eliminate any confounding effect of the display device or light conditions [83], we conducted the study in a **controlled lab environment**. To make the effect of miscompressions measurable, we included test images that were not miscompressed for comparison. We created these **control images** of the same scenes by compressing the source images with a different neural compression codec or JPEG. This allows us to control for the effect of scene content (using scene fixed effects in the analysis). At the same time, we wanted to ensure that each subject saw each scene only once. This requires a **between-subjects** design. To increase the number of ratings and the diversity of scenes, we have combined it with a **within-subject** design (making subject fixed effects necessary to account for repeated measurements).

Our power estimation suggested that we should have at least 15 ratings for each image pair.[2] As we could not expect every subject to spot all the differences, we added a margin and aimed for 25 views of each image pair. Our pretests revealed that twelve ratings per subject would fit into the time frame of 15 minutes. With a conservative estimate of 100 participants, we decided to have four groups. Section 3.2 provides details on the stimulus selection and placement within the instrument.

3.1 Instrument

Our instrument had four parts: introduction, demographics, image comparisons, and control variables.

Part 1: Introduction. After receiving a briefing and signing a consent form, participants entered the instrument. The start page informed them that the study aimed to measure their perception of distortions in digital images introduced during the transmission over the internet, and that they would compare pairs of images and answer questions about the differences. They were not told about neural image compression or miscompressions.

Part 2: Demographics. We asked for the participants' age, gender identity, and whether they had any visual impairment, and used visual aids during the study.

Part 3: Image comparisons. In the main part of the instrument, participants were asked to imagine a scenario where they had taken an image and uploaded it to a social media platform. The image had gone viral, and another person discovered it in their feed on another social media platform. Two images were displayed next to each other on the same screen, the original (reference image) on the left and the received image (test image) on the right. Figure 3 shows a screenshot of the stimulus presentation in the instrument. The same page stated that the two images were *not* identical and asked whether participants could see at least one difference. If the answer was "no", the study proceeded to the next image pair.

Participants who responded with "yes" were asked three follow-up questions. To investigate Hypothesis 1, we asked whether the differences could lead to misunderstandings between them and the person receiving the image. We deliberately left the terms "difference" and "misunderstanding" vague because we feared that providing a definition with examples could bias the interpretation. Participants could express their certainty on a rating scale annotated with the labels "certainly", "very likely", "likely", "unlikely", "very unlikely", and "certainly not". We chose six points to prevent undecided respondents from choosing a neutral midpoint [13]. To investigate Hypothesis 2, we asked participants whether they would attribute the differences to intentional editing, mentioning retouching, filters, or manipulation as examples. We recorded their answer on the same scale as before. To test Hypothesis 3, we asked whether they would attribute the differences to uncontrolled distortions, using transmission errors or compression as examples, again on the same scale. The description of the scenario and the two images remained visible for all three questions.

[2]Pre-data, our rationale was to not miss a "large" ($f = 0.36$) effect with more than 50 % probability at the $p \leq 0.05$ significance level when analysed individually (Section 4.3). Since we had multiple image pairs, we could tolerate missing effects in half of them. We used the R package `pwr` to calculate the sample size for a balanced one-way analysis in two groups (miscompressed vs other similar image).

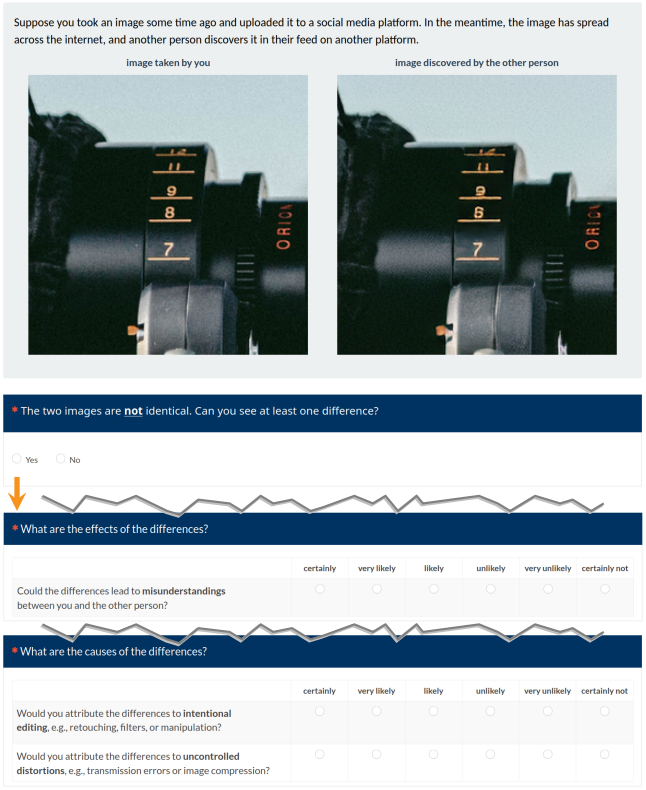


Figure 3: Screenshot of the standardized stimulus presentation (translated from German). The images remained displayed above of all three question blocks. The last two blocks were skipped if the first question was answered "no." The stimulus here is Camera: the original on the left, and the miscompression on the right, where the engraved number 8 has been altered to a 6.

After the first image pair, we informed the participants that the same scenario will be repeated for multiple images. We also reminded them to stay focussed and examine each pair carefully. We did not mention the number of image pairs (twelve), but a progress bar gave indications.

Part 4: Control variables. The literature documents various factors that may influence perception, such as previous experience with digital processing [47], photography [75], generative AI [46], and media literacy [46, 49]. We therefore included control questions for those variables, asking about participants' previous exposure to generative and conventional image processing technologies and how often they verify the authenticity of images online.

To assess the external validity of the data collected in our hypothetical scenario, we also asked participants about their experience with image sharing and how realistic they thought the scenario was. The study concluded with an open-text field for feedback on image selection, question clarity, and any difficulties encountered.

The instrument was implemented using an online survey tool[3] hosted by the university. We disabled navigation and recorded responses as well as response times. All questions required responses, except for the final open-text feedback question. As display time restrictions can impair participants' ability to detect anomalies in generated images [51], we did not impose any time limits. An English translation of the original German questionnaire can be found in Appendix B.2.

3.2 Stimuli

Because miscompressions are rare, and no method exists to deliberately craft targeted examples, we used four state-of-the-art neural compression codecs [5, 7, 62, 88], including the reference model for the upcoming JPEG AI standard, to compress and then manually inspect images that are suitable as stimuli. We use the definition of miscompressions in the literature, which requires that a human observer would use a different verbal description for the compressed image than for the original [38, p. 3]. We report all details of the compression, including codecs, settings, and bit rates, in Appendix B.3 and Table 2.

Selection. From a shortlist of manually collected miscompressions, we selected 19 scenes using criteria designed to ensure diversity and maintain participant engagement. Specifically, we varied image properties and content known to influence image perception [24, 51, 67, 81], including texture, scene complexity, detail, global brightness, contrast, structural complexity, and the presence of dominant edges or flat areas. We also varied indoor and outdoor environments of differing perspectives (near objects, wider scenes). The selected scenes cover objects, buildings, vehicles, identifiable and non-identifiable persons, and body parts. We further varied the *types* [38] and *severity* of miscompressions, with severity determined by the authors' subjective perception. For instance, we include a miscompression where a crescent moon tattoo is turned into a full moon tattoo (M-TATTOO) as a severe example, and a miscompression where open window shades appear closed (M-WINDOW SHADES) as a less severe example. We also included miscompressions of objects with strong semantic meaning. Examples are the miscompression where the number 8 is turned into the number 6 on a camera objective (M-CAMERA), and the arrow-shaped traffic light that is turned into a regular round traffic light (M-TRAFFIC LIGHT (ARROW)). For contrast, one scene was intentionally chosen for its semantic irrelevance (CARDIGAN). All test images can be viewed in Appendix B.5 and are provided in the supplemental material.

Placement in the instrument. Figure 4 documents how and in which order the stimuli were assigned to the four groups, thereby balancing content, properties, and control type. Boxes with letters indicate which test image the respective group compared to the original reference image. We included six control images and six miscompressions (**M**) per group. The control images consisted of three JPEGs (**J**), two neurally compressed images (**C**) that were not miscompressed, and one uncompressed test image (**U**) that did not differ from the reference, as an attention check.

While selecting the stimuli, we found two instances of multiple miscompressions from different neural compression codecs for the same scene. The first scene is TRAFFIC LIGHT, with the modified shape (M-TRAFFIC LIGHT (ARROW)) and a "No Cars" sign, that resembles a "No Photography" sign (M-TRAFFIC LIGHT (SIGN)). The second scene is WATCH. In one version the smartwatch display appears to be turned off (M-WATCH (OFF)) and in the other version, the watch appears physically broken (M-WATCH (BROKEN)). We decided to include both versions to gain insights into different forms of miscompressions for the same scene. The semantically irrelevant stimuli, CARDIGAN, is the only scene without a corresponding miscompressed variant. This resulted in a total of 36 image pairs.

To better compare responses between groups, we selected four *anchor images* that were identical for all groups (*cf.* anchor symbols in Fig. 4). Two anchor images were placed at the beginning to give all groups an equal opportunity to familiarize themselves with the task. The first anchor (M-CAMERA) was a motivating miscompression and the second (J-SANTORINI) was a JPEG control image containing visible compression artifacts. The two last images were again anchor images with a miscompression (M-BAG) and the uncompressed control image (U-ROAD). We decided to place U-ROAD at the very end, as we were concerned that participants might get demotivated when they could not find a single difference. Between the anchors, all participants in the same group viewed the same images in random order, to reduce potential bias caused by order or fatigue effects.

Presentation. All participants used identical desktop computers with 23-inch monitors (1920×1080 resolution) and accessed the instrument via Firefox. The images were displayed at 512^2 pixels (13.5 cm side length). As some miscompressions were very small (*e.g.*, ROCK HOUSE, BRAKE LIGHTS), we used smaller crops (128^2 or 256^2 pixels) and scaled them up to 512^2 pixels with nearest neighbor upsampling to ensure constant size and avoid uncontrolled upsampling by the browser. The increasing visibility of individual pixels also signals a low resolution to the participant.

3.3 Statistical Analysis

We test our main hypotheses by fitting linear regressions with the ordinary least squares method. The specifications we consider take the form,

$$y_{i,k} = b_0 + b_1 \cdot x_{\mathrm{mc},k} + b_2 \cdot x_{\mathrm{jpeg},k} + \cdots + d_i + s_j + \varepsilon_{i,k} \,,$$

where $y_{i,k}$ is the rating of the i-th subject on the k-th image pair in the range $\{1, \ldots, 6\}$, $x_{\mathrm{mc},k} \in \{0, 1\}$ is an indicator for miscompressions, $x_{\mathrm{jpeg},k} \in \{0, 1\}$ is an indicator for control images with conventional JPEG compression, "…" are placeholders for additional control variables, and $\varepsilon_{i,k}$ is the residual. The coefficients d_i and s_j are the estimated subject and scene fixed effects, respectively, and b_l are the estimated coefficients we report and interpret. Control images that were compressed with neural compression have $x_{\mathrm{mc},k} = x_{\mathrm{jpeg},k} = 0$. The fixed effects aim to capture the panel structure in our mixed within and between-subject design, reducing the likelihood that the residuals are unduly correlated (as confirmed by regression diagnostics). All anchor images share one scene fixed effect to prevent collinearity. We report four specifications of three dependent variables, one for each hypothesis: the perceived risk of misunderstanding (H1), the perceived likelihood of intentional editing (H2), and the perceived likelihood of uncontrollable distortion (H3). The specifications differ in which coefficients are forced

[3] https://www.limesurvey.org/

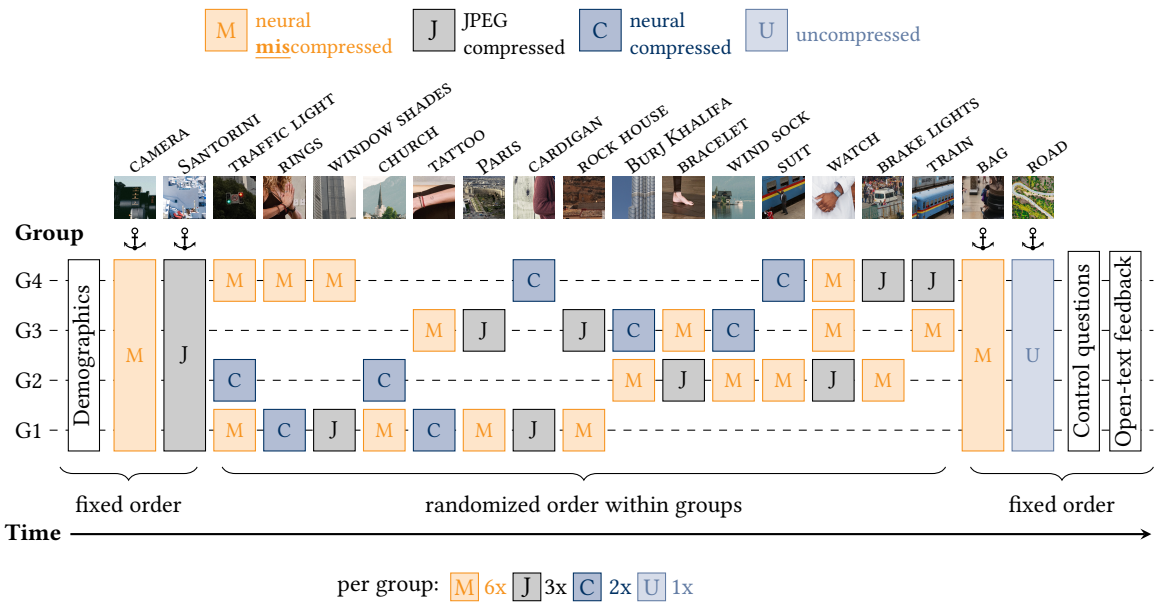


Figure 4: Flowchart of our instrument. Participants were split into four groups, G1–G4, depicted in rows. Image pairs (in columns) were assigned so that each group viewed six control images and six miscompressed images. The color of a box indicates in which version a test image was shown to all members of the respective group. For example, WINDOW SHADES was shown to Group 1 as JPEG and to Group 4 as miscompression. Groups 2 and 3 have not rated this stimulus. The order of the presentation was randomized for each participant in the indicated range. All groups viewed the remaining four anchor image pairs (positionally fixed as the first and last two pairs). The uncompressed anchor image pair at the end was an attention check.

to zero. We exclude image pairs where the subject did not report seeing any differences.

We use Cohen's d [19] to report the effect size. The numerator is the estimated mean difference, b_1, and the pooled standard deviation in the denominator is calculated from the residuals.

3.4 Ethics and Data Protection

The study was IRB approved. All participants consented to their participation and to the collection and processing of their personal data for the purpose of scientific research. They also agreed to the publication of the data in a way that would not allow them to be identified. Participants were informed that participation was voluntary and that they could withdraw from the study at any time. After completing the study, we offered them a printed debriefing sheet to inform them about the research project. They also had the option of leaving a contact address to be informed of the study results. Participants received no financial compensation. When selecting stimuli, we took into account potential triggers of negative emotions in order to avoid any psychological or social harm.

3.5 Limitations

Our method has limitations. First, our sample of participants consists of German-speaking undergraduate computer science students that agreed to participate without compensation. Although their perception of the risk of misunderstanding and attribution of image differences may not be representative of the general population, this sample is arguably similar to early adopters of new technology. Moreover, previous research found that the cultural background can influence the perception of image distortion [66]. Future research should consider a more heterogeneous sample. Second, our stimuli are hand-selected. While there is no commonly agreed way of sampling representative images, our selection strategy was geared towards showcasing a spectrum of miscompressions of varying type, visibility, and severity (according to the authors' subjective perception). The main results are averages over all scenes, and could have been much stronger if we had included more scenes like ROCK HOUSE or WATCH, or much weaker if all our scenes were like WINDOW SHADES. We report breakdowns and interpret individual scenes to increase confidence in our findings. Third, we did not collect which specific differences participants noticed and referred to. We considered asking for a written description of the detected differences [51, 75], but decided against it for fear of fatigue effects. We also considered visually highlighting predefined differences in the images [70] or providing hints, but also decided against this, as it has been shown to influence decisions [75]. There may also be hidden limitations that will only become apparent as the body of work in this area grows. As with any first empirical study of a new phenomenon, this one relies on some decisions that are essentially educated guesses.

4 Results

We will now present our results. Section 4.1 describes the sample, Section 4.2 presents the main results on the aggregate level, and Section 4.3 offers a breakdown by stimulus.

4.1 Descriptive Statistics

Our sample has 115 participants (31 female, 80 male, 4 non-binary or other) in the age range 19 to 40 (median 21). The median response time for the whole instrument was 12 *m*15 *s* (quartiles 10 *m*29 *s* and 14 *m*16 *s*). 45% of the participants report a visual impairment and 75% of them used optical aids to compensate. The participants are balanced across groups (25, 28, 31, 31) with the expected random variation. Each image pair in a group has an average of 25 ratings, with a minimum of 11 for J-Paris. Recall that participants only rated images in which they noticed differences. The anchor images have 97 or more ratings, except for U-Road, which is the uncompressed check image and does not have any difference (6 participants report having seen one).

Figure 5 shows the proportion of participants who noticed a difference (rightmost bars) for all images along with the median response time to answer the yes/no question. For most image pairs, it took participants less than 50 seconds to notice a difference. The anchor images M-Camera and J-Santorini stand out on the slow end as they were the first stimuli of the instrument and participants had to familiarize themselves with the scenario and questions. It also took some time to check that there are no differences in U-Road. Differences were most often overlooked for J-Cardigan and for both M-Paris and J-Paris.

Tables reporting the descriptive statistics of our control questions are provided in Appendix C.2. We emphasize that the self-reported attitudes or behaviors in our control questions are intended to contextualize the sample. With regard to conventional image processing (retouching, montage, and digital photography), a majority report having tried them and being able to explain how they work. This contrasts with AI-supported techniques (generation, generative inpainting). Most participants have only heard of them and the practical experience is limited to having tried image generation, but not inpainting. While most participants are experienced and report a good understanding of conventional compression, they have no practical experience with neural compression. Only 32% have heard of it. A non-existing technique, "virtual image compression," was included as an attention check [61]. Most participants are honest and some think they have heard about it (Tab. 3). We also asked how often our participants try to verify images across different platforms and contexts (Tab. 4). A majority verifies images from unknown social network profiles at least occasionally, but tends to trust images from private contacts. Images on reputable news sites are verified slightly more often, at around the level of social network profiles of public persons or organizations. Our final control question aimed to measure the external validity by asking how realistic the scenario of a photo going viral is (Tab. 5). 18% state that they have already experienced a situation like this, and 68% find it realistic that an image could get modified while sharing. We interpret this to mean that our scenario is not too artificial.

From the answers to the open feedback question we extracted three categories of repeating comments: 11% of participants report uncertainties about the term *misunderstanding* and would like more context, 4% are uncertain about what constitutes a *difference*, and 2% inform us that they have heard about the research project before, which may introduce bias. We use this for a robustness check.

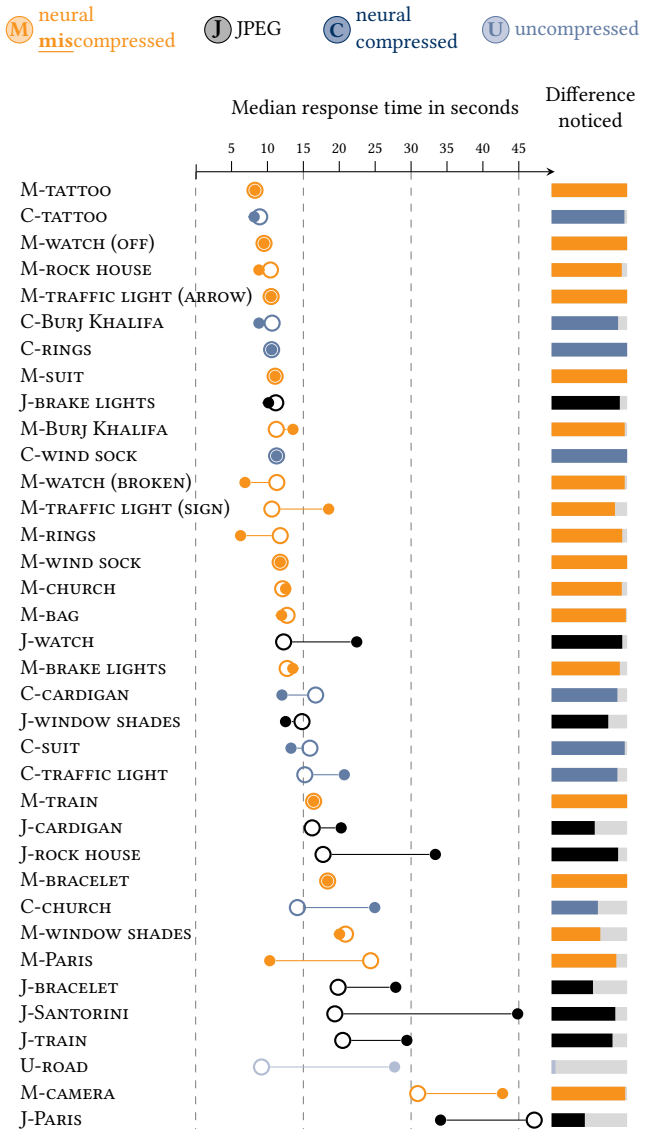


Figure 5: Time needed to spot differences. Rings ○ for respondents who noticed a difference, bullets • for respondents who did not. Share of respondents who noticed a difference is indicated on the right. Sorted by increasing median response time over all respondents and color coded by image type.

4.2 Main Results

Table 1 shows the regression results analyzing how different predictors influence users' perceived risk of how certain image differences can lead to misunderstandings (H1), are attributed to intentional editing (H2), and uncontrollable distortion (H3). The coefficients show the effect of the predictors on the user ratings measured in units of a 6-point scale.

Table 1: Regression results supporting our hypotheses in the panel data

	Dependent variable											
Predictor	Misunderstanding (H1)				Intentional editing (H2)				Uncontrollable distortion (H3)			
Specification	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
Miscompression	0.90^{***}	0.98^{***}	0.98^{***}		0.81^{***}	0.99^{***}	1.00^{***}		-0.57^{***}	-0.71^{***}	-0.73^{***}	
JPEG				-1.03^{***}				-1.11^{***}				0.79^{***}
Size 128^2			0.20				-1.34^{*}				1.47^{**}	
Size 256^2			−0.33				-1.09^{*}				1.14^{*}	
Visually impaired			0.43				0.22				−0.57	
Scene fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Subject fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Within R^2 (adjusted)		0.35	0.35	0.35		0.30	0.30	0.30		0.24	0.24	0.24
Total R^2 (adjusted)	0.09	0.46	0.46	0.43	0.06	0.40	0.40	0.38	0.04	0.30	0.31	0.29

Coefficients normalized to one unit of the 6-point rating scale from "certainly not" to "certainly."
$N = 1131$. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

For each dependent variable, we present four specifications: (i) a null model without fixed effects (which should not be interpreted), (ii) a baseline model with scene and subject fixed effects, (iii) an extended model including selected binary control variables, and (iv) a variant of the baseline model where the predictor is replaced with the indicator for JPEG compression. The latter is not directly related to our hypotheses and included as a contrast when the familiar JPEG artifacts are present. The estimated coefficients from specifications (ii) and (iv) are visualized as bars in the headline results (Fig. 2 in Sect. 1).

In the baseline model, the main effect for miscompressions on misunderstanding is positive, close to 1.0, and statistically significant at the $p < 0.001$ level. This means that, after controlling for all scene and subject-specific variation, the average participant sees the risk of a misunderstanding one step closer to "certainly" if the image is a miscompression. **This supports Hypothesis 1.** The effect does not change when control variables for the crop size and a visual impairment are included. We also included control variables derived (by approximate median splitting) from the questions on theoretical knowledge and practical experience with both conventional and AI-based image processing techniques. We observed no significant effect and refrained from reporting these specifications. We also get null results for the control variables on image verification behavior and for all dependent variables. Interpreting the R^2 measures, we observe that about 11% of the variance is explained by the presence of a miscompression, compared to 35% explained by differences between scenes and subject-specific level shifts.

We see almost the same effect for miscompressions on intentional editing. After controlling for heterogeneous subjects and scenes, the average participant attributes the difference to intentional editing one step closer to "certainly." **This supports Hypothesis 2.** The variance explained by the predictor is about the same, but the fixed effects explain slightly less for intentional editing (30%) than for misunderstandings. Unlike before, small patches with visible pixelation due to upscaling tend to offset the attribution to intentional editing. The statistical significance level is lower due to the small number of images created from small crops.

The main effect for miscompressions on the attribution to uncontrollable distortion is negative, at around −0.7, and statistically significant at the $p < 0.001$ level. This means that after controlling for heterogeneous subjects and scenes, the average participant responds 70% of a step closer to "certainly not" when asked whether they attribute the differences to uncontrollable distortion, thus **supporting Hypothesis 3**. This compares to more than one step towards "certainly" if the image is a 128^2 crop, indicating that some participants attribute pixelation to distortion. They apparently find it hard to distinguish what distortion is already present in the reference and what is added in the test image.

Regression diagnostics do not reveal anything of concern. The stability of the coefficients across the specifications indicates the absence of excessive collinearity or suppression effects. We also explored whether the number of images a subject had rated or the position of the image in the instrument has an effect. The former has no effect and the latter has a very small learning effect, which we do not interpret because the order is confounded with the scene. As additional robustness checks, we re-estimated all regressions on two subsets of the data, first excluding the 19 subjects who mentioned one of the three concerns in the open feedback question, resulting in $N_1 = 954$ ratings, and second excluding the six subjects who saw a difference in the U-Road image, failing the attention check ($N_2 = 1064$). The signs, magnitudes, and significance levels of the main effects are unchanged.

To better interpret what a difference of one step on a 6-point scale from "certainly" to "certainly not" means for the noise level in our data, we calculate Cohen's d as a measure of effect size [19]. We obtain a "large" effect, $d = 0.86$, for misunderstanding (H1), and "moderate" effects for the other dependent variables, $d = 0.78$ (H2) and $d = 0.64$ (H3), respectively. An exploratory analysis of potential gender effects revealed that noticing differences in image pairs does not depend on gender. We observe a tendency for male participants to have slightly stronger effects in the hypothesized direction than females, especially for H1. As the interaction terms are statistically less significant ($0.01 < p < 0.1$) than our main results and the sample is imbalanced, we and refrain from reporting and interpreting quantitative results.

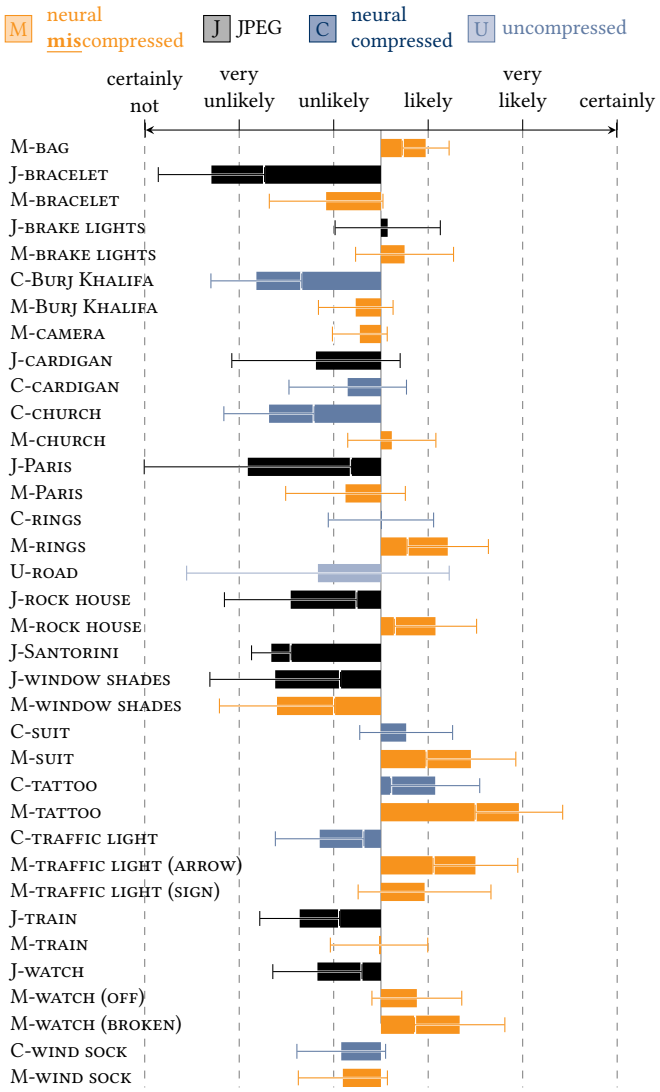


Figure 6: Perceived risk of misunderstanding after image transmission. Means of a 6-level scale with 95 % confidence intervals, broken down by stimulus image and color coded by image type. Miscompressions (M, orange) dominate on the side of elevated risk. See text for important exceptions.

4.3 Breakdown by Image

While the main results give a consistent message, it is important to note that this is the average effect across a diverse set of images curated by us. Figure 6 unpacks this by showing participants' perceived risk of misunderstanding for each image. Although most of the miscompressed images (orange bars) point towards "certainly" on the right, and the control images (all other colors) tend to point to the left, there are a few exceptions. We will discuss these in an exploratory way to learn more about users' perceptions and potential reasoning.

Paris. This scene surprised us by the relatively small proportion of participants who noticed the difference: 85% for the M-paris and only 44% for J-paris. Unlike the reader in Figure 1, the participants did not have a magnifying glass to zoom into this 512^2 image. While there is a difference in the reported susceptibility to misunderstandings in the hypothesized direction, the level is shifted to the "unlikely" half of the scale and the confidence intervals overlap, indicating that there is no statistically significant difference for this scene alone. However, this scene allows us to interpret the differences in the response time between participants who did and did not notice the differences (Fig. 5). Observe that participants who did not notice any differences were much faster, suggesting that they overlooked the missing people. Prior work has shown that recognizing the high-frequency image content, like the people in this image, requires more effort [75].

Wind sock. The differences were almost unanimously found for both M-Wind sock and C-Wind sock, and participants perceived the risk of misunderstandings as equally "unlikely" for both versions. An explanation could be that participants primarily notice the global blur, rather than the absence of the red wind sock in the miscompressed image (Fig. 7). While we tried to select miscompressions of universally known objects, it could also be that not all participants recognized the wind sock.

Figure 7: M-Wind sock. Participants did not detect the vanishing color of the wind sock or did not perceive it as a risk of misunderstandings. A reason could be the increased smoothness of the background.

Rings. This scene was the only case where more participants noticed a difference in the neurally compressed control image than in the miscompression (100 vs. 94%). This can be explained by the smoothing, as shown in Figure 8, which may resemble popular beauty filters [4]. Interestingly, while the participants attribute the differences in the control image *"very likely"* to intentionally editing (see Fig. 11, below), they do not perceive an increased risk of misunderstandings.

Tattoo. This is the only scene where the neurally compressed control image is perceived as causing misunderstandings. (The confidence intervals are right of the midpoint.) Miscompression and control image still differ in the hypothesized direction, but the confidence intervals overlap. We conjecture that tattoos are perceived as

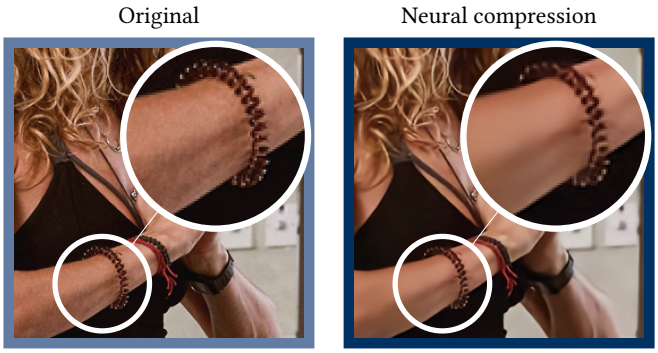


Figure 8: C-Rings: Participants attributed differences in the neurally compressed control image to intentional editing.

sensitive features that allow one to draw inferences about a person's personality or identify individuals. This may incline participants to flag an issue (Fig. 9). While testing this explanation would require a tailored study, we see a similar tendency in the Suit scene, where the person's face is disfigured.

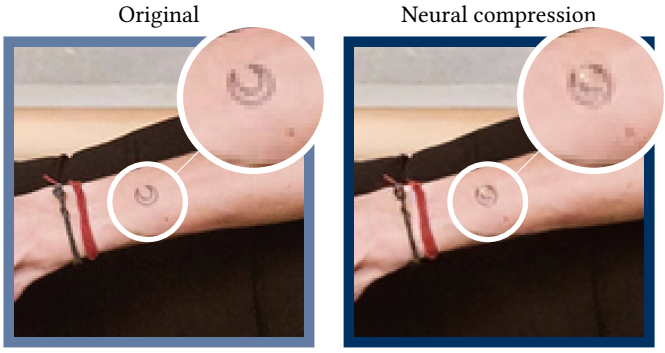


Figure 9: C-Tattoo: Participants perceive the neurally compressed control image as a likely cause of misunderstandings.

Window shades. This stimulus was included as an example for a miscompression of low severity. 64% of the participants notice differences in the miscompression, the lowest share of all miscompressions, and only a few believe that these could lead to misunderstandings. Participants may overlook the closed window shades because they are hard to spot (Fig. 10). Alternatively, they may not perceive window shades as semantically relevant enough to cause misunderstandings. Moreover, the differences were attributed to uncontrollable distortion rather than intentional editing (Fig. 11). Of course, this is context dependent [75]. If the scene was presented to experts who review construction faults, the results could be very different. This alludes to a general challenge for miscompression research. The number of domains of expertise to cover is huge.

We present the breakdown for the two remaining dependent variables in a combined scatter plot. Figure 11 shows the mean and confidence intervals for the attribution of the differences to intentional editing on the horizontal axis and to uncontrollable

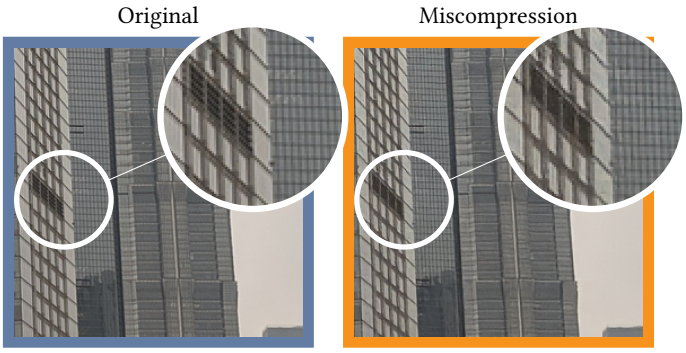


Figure 10: M-Window shades: Participants perceived a low risk of misunderstanding.

distortion on the vertical axis. We have annotated extreme points and interesting scenes discussed above. Observe that while the data span a large range of the intentional editing scale, the responses for uncontrollable distortion are much more concentrated on the side of "certainly." This means that our participants can identify conventional JPEG compression artifacts. The negative correlation between the means on both axes shows how H2 and H3 are related across diverse scenes, and even more so between types. However, at the level of individual responses for any given scene, the two causes are not perceived as mutually exclusive (see Fig. 14 in the Appendix).

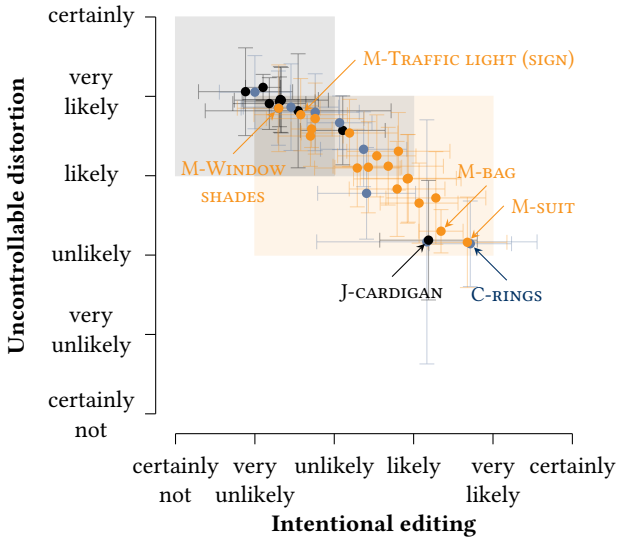


Figure 11: Suspected causes of the image differences: scatter plot of means over the responses for intentional editing (H2) versus uncontrollable distortion (H3) aggregated for each image pair (with 95 % confidence intervals as bars). The Pearson correlation between image means is $\rho = -0.95$ ($p < 0.001$). The shaded areas visualize the empirical interquartile ranges over all responses in each type (color coding as in Fig. 4).

5 Discussion

Image-based communication is part of everyday life and commonly relies on lossy compression. Today's codecs work invisibly for end users, fostering the expectation that mere compression does not alter the semantics of an image. Neural compression, if adopted, will disrupt this expectation for the sake of bandwidth savings. Even if codecs improve further, any information not transmitted will need to be synthesized, leaving room for hallucinations by the generative AI in the decoder. Understanding how such disruptions affect people's perception and use of images for communication is therefore critical.

This study is the first to confirm that concerns previously raised by individual researchers about miscompressions in neural compression codecs are shared by a wider set of people. This has several implications. We begin by discussing specific implications of our results for each hypothesis, before turning to broader implications for the design of future image communication interfaces.

5.1 Implications of Our Results

5.1.1 Miscompressions Elevate the Risk of Misunderstandings (H1). The full-reference design allowed us to probe how users perceive semantic changes at the level of detail. Our results indicate that even subtle differences can have an effect and lead to misunderstandings, a catch-all term for a range of potential social consequences. One can easily envision the resulting risks: a changed arrow in a traffic light could cause accidents (TRAFFIC LIGHT (SIGN)), a missing wedding ring might spark false accusations (RINGS), interpreting miscompressed images in science and engineering can lead to false conclusions (WINDOW SHADES), and law enforcement could fail if biometric identifiers are misrepresented (TATTOO).

However, the current real-world impact of miscompressions is limited because the technology is not yet rolled out. This gives the community time for more human subjects research, which could inform the development of better codecs. It can include direct end-to-end tests for complete codecs, similar to our study. Moreover, data from user studies can be used to train learnable image quality metrics, which in turn can be used to train neural compression codecs.

Future research. Whether a miscompression poses a risk of a misunderstanding might depend on the context. A follow-up study should test other scenarios than social media. Studies involving experts in journalism, law enforcement, and forensics can complement the picture for high-stakes applications. Concerning the instrument design, we believe that our participants had examples of misunderstandings in mind, but we did not ask for them. Future work could ask users for this information for each scene. This information would make it possible to distinguish between the likelihood and the expected severity of misunderstandings, linking these judgments to established categories in risk management [76]. On a broader level, these insights can guide a responsible deployment of neural compression across application areas. More narrowly, they can inform the calibration of future neural compression codecs to enable a risk-adjusted allocation of bits. Details which are prone to be miscompressed and cause potentially severe consequences should be preserved.

5.1.2 Miscompressions Are Confused With Intentional Editing (H2). Until now, the only way semantics of parts of images can change is when they are edited. This explains why people who are unaware of neural compression wrongly attribute the local differences of miscompressions to editing. This confusion can be problematic: mistaking compression artifacts for editing might hinder attempts to resolve misunderstandings (H1) through image comparison, especially if the parties involved (or even an independent third party) are unaware of the true cause of the differences. Moreover, neural compression has been shown to mislead "deep fake" detectors [16]. Accusations of malicious manipulation are quickly raised.

Moreover, previous research has shown that people are better at detecting manipulations than at identifying generated images (see Sect. 2.2). If neural compression becomes more widely adopted, people may lose this ability as they get used to neural compression artifacts and can no longer use them as indicators of manipulation. Likewise, as the decoders of neural compression leverage generative AI, more images will appear "synthetic", diminishing users' (already limited) ability to detect generated images. Both effects further contribute to the erosion of trust in images [26].

Future research. This finding paves the way for a number of follow-up research questions. Future work could contrast miscompressions with actual manipulations, both manual and AI-supported (*e.g.*, inpainting). It should also involve participants who are familiar with neural compression (*e.g.*, after passing a training phase) and experts who deal with image manipulations professionally. Our results may also prompt further research in the legal domain. The EU's AI Act [69] requires providers of AI systems to embed machine-readable marks in generated content. Systems that "do not substantially alter the input data [...] or the semantics thereof" are exempt [69, Art. 50 (2)]. It must be clarified whether potential miscompressions would constitute an alteration to the semantics as defined by this law.

5.1.3 Miscompressions Are Not Recognized as Compression Artifacts (H3). The generative networks in neural compression tend to produce visually appealing, photorealistic images. Our neurally compressed test images do not show typical compression artifacts regardless of whether they are miscompressed or not. The absence of these indicators can lead to misplaced trust in images and overconfidence in their content.

This interpretation is supported by our finding that people are indeed familiar with JPEG artifacts and interpret them as signs of compression. JPEG test images are associated with a significantly lower risk of misunderstanding and differences are less often explained with editing. In contrast, for neurally compressed control images, participants often resorted to uncontrollable distortion as a fallback explanation when no plausible cause for the differences was apparent. We conclude this from the strong negative correlation of the image means in Figure 11.

Future research. To date, little is known about how compression-induced cues interact with people's perception and trust in images, likely because researchers assumed the effect to be marginal. Our findings challenge this assumption. Follow-up studies should examine whether specific personal factors, such as experience and education draw people's attention to such cues. These studies should

vary the type and strength of compression artifacts (*e.g.*, blocking, blurring, and ringing) and control for image content, device, and resolution. More than 30 years of JPEG compression could have affected how people perceive digital images.

5.2 Possible Interventions

Miscompressions challenge long-standing assumptions about the integrity of image communication. Given that research on neural compression is still at an early stage, the first widely adopted codecs will likely surpass those we can currently evaluate. Nevertheless, it remains highly uncertain whether it will be possible to fully avoid miscompressions with technical means. This calls for the development of strategies to mitigate potential risks. In the following, we consider approaches that allow users to recognize and respond to the risks posed by miscompressions. We distinguish passive strategies that notify users and raise awareness (Sect. 5.2.1) and active strategies that involve user interaction and feedback (Sect. 5.2.2).

5.2.1 Notifications and Awareness. Users who are aware of potential semantic changes are, in principle, better positioned to evaluate how much to trust an image. Since miscompressions cannot be detected reliably with current technology, flagging individual instances is not feasible. The risk is present in every neurally compressed image. Therefore, users should always be informed about the use of neural compression. This can be done with labels or visible watermarks in the images or image captions. Technical initiatives, such as C2PA [1, 80] and JPEG Trust [78] may serve as sources of provenance information.

Prior work provides guidance on effective labeling strategies for edited [53], AI generated [26, 28, 86], and identified misinformative content [44, 50]. While these works are excellent starting points, their findings may not translate directly to neural compression. Labels for generated content typically signal that an image is not authentic. By contrast, a label flagging the use of neural compression only indicates that there is a possibility that an otherwise authentic image contains altered details. More research is needed to design and test effective image labeling systems adapted to the neural compression context. This should not be studied in isolation. For labeling systems to be usable in practice, they must consider the entire user experience, spanning all kinds of content sources and level of trustworthiness. The abstract risk of a miscompression must be weighed against the certain presence of completely artificial material, possibly generated with the intent to mislead. A comprehensive image labeling systems must integrate all this information and ensure that users understand it and facilitate a reaction that corresponds to the risk.

5.2.2 Interaction. Tailored user interfaces can support users to detect and react to instances of miscompressions. Drawing inspiration from progressive encoding [37], one approach is to deliver images at a low bitrate by default while retaining a high-quality version that can be requested on demand, for instance, when a user zooms into an image or activates a "view-as-sent" function. A user interface could also allow viewers to select whether the decoder should produce an appealing, realistic looking version of an image, or rather a version that is of lower quality but closer to the original. This is possible by transmitting a single bitstream [2]. Senders concerned about potential miscompressions could similarly benefit from a "view-as-received," option, enabling them to verify what recipients will see. Although such features cannot eliminate risk entirely, maintaining access to the original version for some period of time (by storing it on the server) could be valuable for resolving potential misunderstandings after they arise.

These techniques allow users to detect miscompressions. Once detected, users should be able to report them to the platform operator, similar to existing interfaces to report harmful content or misinformation [20]. Reactions to these reports include notifying all affected users of the specific miscompression, censoring the problematic region, replacing the image with a high-quality version, and collecting a database of miscompressions that can be used to improve future neural compression codecs. Several HCI studies are needed to determine the suitability of these measures and to identify effective ways to design and integrate such interaction mechanisms into real-world image communication platforms.

6 Conclusion

With billions of images exchanged through messaging apps every day, reliable image compression has become central to digital communication. This study takes a new direction of research that investigates how people perceive artifacts of conventional and future image processing technologies, and how they interpret them as cues for reliability. While the controversy involving the journalists in the introduction was fictional, our findings suggest that similar misunderstandings could arise in the real world. As compression shifts toward neural methods that can inadvertently alter the meaning of images, it is important to understand the personal and social consequences of this transition. Addressing these challenges requires not only technical advances but also sustained HCI research into how users perceive, interpret, and respond to the risks introduced by neural compression. Only in this way can we, as a society, be put in a position to decide how much bandwidth is worth saving at the expense of trust in images.

Acknowledgments

We thank all participants for taking part in our user study. We also thank Max Ninow for his help in setting up the instrument; Pascal Knierim, Florian Alt, and the members of Florian's research group for helpful advice on writing a CHI paper; and Simon Koch and Kristina Magnussen for valuable comments on the draft. This work received funding by the state of Tyrol (F.50541/6-2024). Computational results were achieved using the LEO HPC infrastructure at the University of Innsbruck.

References

[1] Rafal Ablamowicz and Bertfried Fauser. 2025. *Content Credentials: C2PA Technical Specification.* Retrieved Aug 29, 2025 from https://spec.c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html

[2] Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. 2023. Multi–realism image compression with a conditional generator. In *Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 22324–22333.

[3] Eirikur Agustsson and Radu Timofte. 2017. NTIRE 2017 challenge on single image super–resolution: Dataset and study. In *Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 126–135.

[4] Markus Appel, Fabian Hutmacher, Theresa Politt, and Jan-Philipp Stein. 2023. Swipe right? Using beauty filters in male Tinder profiles reduces women's evaluations of trustworthiness but increases physical attractiveness and dating intention. *Computers in Human Behavior* 148 (2023), 107871.

[5] Joao Ascenso, Elena Alshina, and Touradj Ebrahimi. 2023. The JPEG AI standard: Providing efficient human and machine visual data consumption. *IEEE Multimedia* 30, 1 (2023), 100–111.

[6] Jona Ballé, Valero Laparra, and Eero P Simoncelli. 2017. End-to-end optimized image compression. In *International Conference on Learning Representations*.

[7] Jona Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. 2018. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*.

[8] Martin Beneš, Nora Hofer, and Rainer Böhme. 2022. Know Your Library: How the libjpeg version influences compression and decompression results. In *Workshop on Information Hiding and Multimedia Security*. ACM, 19–25.

[9] Sandra Bergmann, Denise Moussa, Fabian Brand, André Kaup, and Christian Riess. 2024. Forensic analysis of AI-compression traces in spatial and frequency domain. *Pattern Recognition Letters* (2024), 41–47.

[10] Sandra Bergmann, Denise Moussa, and Christian Riess. 2024. Trustworthy compression? Impact of AI-based codecs on biometrics for law enforcement. *arXiv preprint arXiv:2408.10823* (2024).

[11] Alexandre Berthet and Jean-Luc Dugelay. 2022. AI-based compression: A new unintended counter attack on JPEG-related image forensic detectors. In *International Conference on Image Processing*. IEEE, 3426–3430.

[12] Alexandre Berthet, Chiara Galdi, and Jean-Luc Dugelay. 2023. On the impact of AI-based compression on deep learning-based source social network identification. In *International Workshop on Multimedia Signal Processing*. IEEE, 1–6.

[13] George Bishop. 1987. Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly* 51, 2 (1987), 220–232.

[14] Yochai Blau and Tomer Michaeli. 2019. Rethinking lossy compression: The rate–distortion–perception tradeoff. In *International Conference on Machine Learning*. PMLR, 675–685.

[15] Rainer Böhme and Matthias Kirchner. 2012. Counter-forensics: Attacking image forensics. In *Digital Image Forensics: There is More to a Picture Than Meets the Eye*, Husrev T. Sencar and Nasir D. Memon (Eds.). Springer, 327–366.

[16] Edoardo Daniele Cannas, Sara Mandelli, Natasa Popovic, Ayman Alkhateeb, Alessandro Gnutti, Paolo Bestagini, and Stefano Tubaro. 2025. Is JPEG AI going to change image forensics?. In *IEEE/CVF International Conference on Computer Vision*. 1564–1575.

[17] João Phillipe Cardenuto, Joshua Krinsky, Lucas Nogueira, Aparna Bharati, and Daniel Moreira. 2025. Implications of neural compression to scientific images. In *Workshop on Information Hiding and Multimedia Security*. ACM, 80–85.

[18] Tong Chen and Zhan Ma. 2023. Toward robust neural image compression: Adversarial attack and model finetuning. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 12 (2023), 7842–7856.

[19] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge.

[20] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.

[21] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. 2013. Exposing digital image forgeries by illumination color classification. *Transactions on Information Forensics and Security* 8, 7 (2013), 1182–1194.

[22] Benedikt Dornauer and Michael Felderer. 2023. Web image formats: Assessment of their real-world-usage and performance across popular web browsers. In *Conference on Product-Focused Software Process Improvement*. Springer, 132–147.

[23] Hany Farid and Mary J Bravo. 2007. Photorealistic rendering: How realistic is it? *Journal of Vision* 7, 9 (2007), 766–766.

[24] Hany Farid and Mary J Bravo. 2010. Image forensic analyses that elude the human visual system. In *Media Forensics and Security II*. SPIE, 52–61.

[25] Hany Farid and Mary J Bravo. 2012. Perceptual discrimination of computer generated and photographic faces. *Digital Investigation* 8, 3-4 (2012), 226–235.

[26] KJ Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X Zhang. 2023. Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Human-Computer Interaction* 7, CSCW2 (2023), 1–42.

[27] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. 2024. A representative study on human detection of artificially generated media across countries. In *Symposium on Security and Privacy*. IEEE, 55–73.

[28] Dilrukshi Gamage, Dilki Sewwandi, Min Zhang, and Arosha K Bandara. 2025. Labeling synthetic content: User perceptions of label designs for AI-generated content on social media. In *CHI Conference on Human Factors in Computing Systems*. 1–29.

[29] Gina Gayle. 2020. *The Perceived Credibility of Professional Photojournalism Compared to User-Generated Content among American News Media Audiences*. Ph. D. Dissertation. Syracuse University.

[30] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems* 27 (2014).

[31] David Gottlieb and Chi-Wang Shu. 1997. On the Gibbs phenomenon and its resolution. *Society for Industrial and Applied Mathematics Review* 39, 4 (1997), 644–668.

[32] Jennifer D Greer and Joseph D Gosen. 2002. How much is too much? Assessing levels of digital alteration of factors in public perception of news media credibility. *Visual Communication Quarterly* 9, 3 (2002), 4–13.

[33] Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. 2021. *Detection Theory: A User's Guide*. Routledge.

[34] Dailan He, Ziming Yang, Hongjiu Yu, Tongda Xu, Jixiang Luo, Yuan Chen, Chenjian Gao, Xinjie Shi, Hongwei Qin, and Yan Wang. 2022. PO-ELIC: Perception-oriented efficient learned image coding. In *Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 1764–1769.

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time–scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).

[36] Nora Hofer. 2024. Increasing trust in image analysis by detecting trellis quantization in JPEG images. In *International Conference on Image Processing*. IEEE, 3834–3840.

[37] Nora Hofer and Rainer Böhme. 2023. Progressive JPEGs in the wild: Implications for information hiding and forensics. In *Workshop on Information Hiding and Multimedia Security*. ACM, 47–58.

[38] Nora Hofer and Rainer Böhme. 2024. A taxonomy of miscompressions: Preparing image forensics for neural compression. In *International Workshop on Information Forensics and Security*. IEEE, 1–6.

[39] Nora Hofer and Rainer Böhme. 2025. Challenging cases of neural image compression: A dataset of visually compelling yet semantically incorrect reconstructions. In *International Conference on Multimedia*. ACM, 13318–13324.

[40] Olivia Holmes, Martin S Banks, and Hany Farid. 2016. Assessing and improving the identification of computer-generated portraits. *ACM Transactions on Applied Perception* 13, 2 (2016), 1–12.

[41] International Telecommunication Union. 2012. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT.500-14* (2012). Available at: https://www.itu.int/rec/R-REC-BT.500-14-201910-I/en.

[42] International Telecommunication Union. 2016. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. *Recommendation ITU-T P.913* (2016). Available at: https://www.itu.int/rec/T-REC-P.913/en.

[43] International Telecommunication Union. 2023. Subjective video quality assessment methods for multimedia applications. *Recommendation ITU-T P.910* (2023). Available at: https://www.itu.int/rec/T-REC-P.910/en.

[44] Farnaz Jahanbakhsh and David R Karger. 2024. A browser extension for in-place signaling and assessment of misinformation. In *CHI Conference on Human Factors in Computing Systems*. 1–21.

[45] Ehsaneddin Jalilian, Heinz Hofbauer, and Andreas Uhl. 2022. Iris image compression using deep convolutional neural networks. *Sensors* 22, 7 (2022), 2698.

[46] Eunchae Jang, Hui Min Lee, Sangwook Lee, Yongnam Jung, and S Shyam Sundar. 2025. Too Good to be false: How photorealism promotes susceptibility to misinformation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[47] Se-Hoon Jeong, Hyunyi Cho, and Yoori Hwang. 2012. Media literacy interventions: A meta-analytic review. *Journal of Communication* 62, 3 (2012), 454–472.

[48] Panqi Jia, A Burakhan Koyuncu, Jue Mao, Ze Cui, Yi Ma, Tiansheng Guo, Timofey Solovyev, Alexander Karabutov, Yin Zhao, Jing Wang, Elena Alshina, and Andre Kaup. 2024. Bit rate matching algorithm optimization in JPEG-AI verification model. In *Picture Coding Symposium*. IEEE, 1–5.

[49] S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. 2021. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist* 65, 2 (2021), 371–388.

[50] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J Nathan Matias, and Jonathan Mayer. 2021. Adapting security warnings to counter online disinformation. In *USENIX Security Symposium*. 1163–1180.

[51] Negar Kamali, Karyn Nakamura, Aakriti Kumar, Angelos Chatzimparmpas, Jessica Hullman, and Matthew Groh. 2025. Characterizing photorealism and artifacts in diffusion model-generated images. In *CHI Conference on Human Factors in Computing Systems*. 1–26.

[52] Mona Kasra, Cuihua Shen, and James F O'Brien. 2018. Seeing is believing: How people fail to identify fake images on the web. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[53] Eric Kee and Hany Farid. 2011. A perceptual metric for photo retouching. *National Academy of Sciences* 108, 50 (2011), 19907–19912.

[54] James E Kelly and Diona Nace. 1994. Digital imaging & believing photos. *Visual Communication Quarterly* 1, 1 (1994), 4–18.

[55] Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. 2021. More real than real: A study on human visual perception of synthetic faces. *Signal Processing Magazine* 39, 1 (2021), 109–116.

[56] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. 2023. Seeing is not always believing: Benchmarking human and model perception of AI-generated images. *Advances in Neural Information Processing Systems* 36 (2023).

[57] Jordan Madden, Lhamo Dorje, and Xiaohua Li. 2025. Bitstream collisions in neural image compression via adversarial perturbations. *arXiv preprint arXiv:2503.19817* (2025).

[58] Brandon Mader, Martin S Banks, and Hany Farid. 2017. Identifying computer-generated portraits: The importance of training and incentives. *Perception* 46, 9 (2017), 1062–1076.

[59] Daniele Mari, Saverio Cavasin, Simone Milani, and Mauro Conti. 2024. Effectiveness of learning-based image codecs on fingerprint storage. In *International Workshop on Information Forensics and Security*. IEEE, 1–6.

[60] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. 2002. A no-reference perceptual blur metric. In *International Conference on Image Processing*, Vol. 3. IEEE, III–III.

[61] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological Methods* 17, 3 (2012), 437.

[62] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. 2020. High-fidelity generative image compression. *Advances in Neural Information Processing Systems* (2020), 11913–11924.

[63] Jaron Mink, Miranda Wei, Collins W Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M Redmiles, and Gang Wang. 2024. It's trying too hard to look real: Deepfake moderation mistakes and identity-based bias. In *CHI Conference on Human Factors in Computing Systems*. 1–20.

[64] Tara Marie Mortensen, Brian P McDermott, and Khadija Ejaz. 2023. Measuring photo credibility in journalistic contexts: Scale development and application to staff and stock photography. *Journalism Practice* 17, 6 (2023), 1158–1177.

[65] Sophie J Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *National Academy of Sciences* 119, 8 (2022), e2120481119.

[66] Yuzhen Niu, Feng Liu, Xueqing Li, and Michael Gleicher. 2010. The complexity of perception of image distortion: an initial study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 3235–3240.

[67] Yuri Ostrovsky, Patrick Cavanagh, and Pawan Sinha. 2005. Perceiving illumination inconsistencies in scenes. *Perception* 34, 11 (2005), 1301–1314.

[68] Basak Oztan, Amal Malik, Zhigang Fan, and Reiner Eschbach. 2007. Removal of artifacts from JPEG compressed document images. In *Color Imaging XII: Processing, Hardcopy, and Applications*, Vol. 6493. SPIE, 60–68.

[69] European Parliament and the Council. 2024. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

[70] Pat Pataranutaporn, Chayapatr Archiwaranguprok, Samantha WT Chan, Elizabeth Loftus, and Pattie Maes. 2025. Synthetic human memories: AI-edited images and videos can implant false memories and distort recollection. In *CHI Conference on Human Factors in Computing Systems*. 1–20.

[71] Yash Patel, Srikar Appalaraju, and R Manmatha. 2019. Human perceptual evaluations for image compression. *arXiv preprint arXiv:1908.04187* (2019).

[72] Tian Qiu, Arjun Nichani, Rasta Tadayontahmasebi, and Haewon Jeong. 2025. Gone with the bits: Revealing racial bias in low-rate neural compression for facial images. In *Conference on Fairness, Accountability, and Transparency*. ACM, 1862–1889.

[73] Lucas Relic, Roberto Azevedo, Markus Gross, and Christopher Schroers. 2024. Lossy image compression with foundation diffusion models. In *European Conference on Computer Vision*. Springer, 303–319.

[74] JPEG (ISO/IEC SC29/WG1). 2024. JPEG AI reference software. https://gitlab.com/wg1/jpeg-ai/jpeg-ai-reference-software, version 7.0.

[75] Victor Schetinger, Manuel M Oliveira, Roberto da Silva, and Tiago J Carvalho. 2017. Humans are easily fooled by digital images. *Computers & Graphics* 68 (2017), 142–151.

[76] Michael Warren Skirpan, Tom Yeh, and Casey Fiesler. 2018. What's at stake: Characterizing risk perceptions of emerging technologies. In *CHI Conference on Human Factors in Computing Systems*. 1–12.

[77] Zhihao Sun, Haipeng Fang, Juan Cao, Xinying Zhao, and Danding Wang. 2024. Rethinking image editing detection in the era of generative AI revolution. In *International Conference on Multimedia*. ACM, 3538–3547.

[78] Frederik Temmermans, Sabrina Caldwell, Deepayan Bhowmik, and Touradj Ebrahimi. 2024. JPEG Trust: an international standard facilitating the assessment of trustworthiness of digital media assets. In *Applications of Digital Image Processing XLVII*, Vol. 13137. SPIE, 99–104.

[79] George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Jona Balle, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. 2020. Workshop and challenge on learned image compression (CLIC2020). In *Conference on Computer Vision and Pattern Recognition Workshops*. IEEE.

[80] Christoph Trattner, Svenja Lys Forstner, Alain D Starke, and Erik Knudsen. 2024. C2PA Provenance labels increase trust in news platforms across western countries. (2024).

[81] Sophie Triantaphillidou, Elizabeth Allen, and R Jacobson. 2007. Image quality comparison between JPEG and JPEG2000. II. Scene dependency, scene analysis, and classification. *Journal of Imaging Science and Technology* 51, 3 (2007), 259–270.

[82] Daria Tsereh, Mark Mirgaleev, Ivan Molodetskikh, Roman Kazantsev, and Dmitriy Vatolin. 2024. JPEG AI image compression visual artifacts: detection methods and dataset. *arXiv preprint arXiv:2411.06810* (2024).

[83] Dhanraj Vishwanath, Ahna R Girshick, and Martin S Banks. 2005. Why pictures look right when viewed from the wrong place. *Nature Neuroscience* 8, 10 (2005), 1401–1410.

[84] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[85] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *Conference on Signals, Systems & Computers*. IEEE, 1398–1402.

[86] Chloe Wittenberg, Ziv Epstein, Adam J Berinsky, and David G Rand. 2024. Labeling AI-generated content: promises, perils, and future directions. *An MIT Exploration of Generative AI* (2024).

[87] Leslie Wöhler, Martin Zembaty, Susana Castillo, and Marcus Magnor. 2021. Towards understanding perceptual differences between genuine and face-swapped videos. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

[88] Ruihan Yang and Stephan Mandt. 2024. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems* (2024), 64971–64995.

[89] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing* 20, 8 (2011), 2378–2386.

[90] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 586–595.

[91] Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang. 2022. The devil is in the details: Window-based attention for image compression. In *Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 17492–17501.

Appendix

A Background

A.1 Principles of Neural Image Compression

The first step of the compression pipeline is the analysis transform, a neural network that encodes the input image into a latent representation. This representation is quantized by rounding and then coded into a bit stream using arithmetic encoding. The most popular construction is called "hyperprior" [7] because it parametrizes the arithmetic encoder by another autoencoder, which predicts the distribution of the latent representation produced by the transform network. This autoencoder's latent space is transmitted for arithmetic decoding. Finally, a synthesis transform network reconstructs the image from the decoded representation. All steps are trained end to end by minimizing a rate–distortion function. To generate the stimulus images in our user study, we used the variational autoencoder used in the original **Hyperprior** proposal by Ballé et al. [7] and three follow-up works. The **HiFiC** model by Mentzer et al. [62] differs in the last step. It uses the decoded latent representation to condition a Generative Adversarial Network (GAN) which reconstructs the image. By contrast, Yang et al.'s **CDC** codec [88] conditions a diffusion model on the decoded latent representation for the reconstruction. Also the **JPEG AI** [5] codec uses the hyperprior construction and feeds the decoded latent representation into a synthesis transform network to reconstruct the image. Unlike other codecs, JPEG AI converts the image into the *YUV* color space before transformation.

A.2 Acknowledgement of Semantic Changes

In their limitation section, Relic et al. [73, p. 316] mention the challenge of "*misgeneration of content*". They remark: "*Specifically at very low bitrates, identities, text, or lower-level content can vary from the original image, and thus may raise ethical concerns in specific scenarios.*" Also, Mentzer et al. [62, p. 10] mention failure cases, such as small text or faces, and emphasize that "in theory" their generator can "*[…] produce images that are very different from the input*" and that their method is "*not suitable for sensitive image contents, such as, e.g., storing medical images, or important documents.*" Most notably, Agustsson et al. note that "*Since the realism constraint might produce reconstructions that are far away from the input, these systems might be looked at with suspicion because it is not clear which details are in the original and which were added by the architecture*" [2, p. 22325]. None of the above studies measure the perception of the semantic differences.

B Method

B.1 Participant Briefing

[Translated from German and *[anonymized]*]

Welcome to the study **Perception of disturbances in digital images** conducted by *[institute]*!

During the transmission of images on the internet, visible and invisible "distortions" can occur. Researchers worldwide are working on new standards to improve the quality of transmitted images. In this study, you will compare original images with transmitted ones and evaluate any differences that may appear.

The participation is voluntary, and you have the right to withdraw from the study at any time. Your responses and the answering time per question will be recorded. All data will be fully anonymized. The study will take approximately 15 minutes to complete. It is part of the *[research project and funding organization]*. The participation in the study involves no risks.

If you have questions during the study, please contact the pro seminar teacher. For questions after completing the study, you can reach out to *[name and contact information of the project leader]*. At the end of the study, you will have the opportunity to provide your email address to receive updates on the study's findings.

Your participation helps us improve future image formats. Thank you for your contribution!

[Declaration of Consent]

B.2 Questionnaire

[Translated from German. *Answer options are listed in italics.*]

Part 1: Introduction

I1: [Same as printed Briefing]

Part 2: Demographics

D1: What year were you born? *Answer must be between 1925 and 2006.*

D2: Which gender do you identify with? *female, male, non-binary, would not like to specify*

D3: Eyesight *yes, no*

(1) Is your vision impaired, *e.g.*, due to long/short-sightedness or color blindness?
(2) Do you use optical aids to compensate for this while completing this study, *e.g.*, glasses or contact lenses?

Part 3: Image Comparison

Suppose you took a picture some time ago and uploaded it to a social media platform. In the meantime, the image has spread across the internet and a second person discovers it in their feed on another platform.

[Stimulus 1] image taken by you
[Stimulus 2] image discovered by the other person

S1: The two images are not identical. Can you see at least one difference? *yes (condition for S2 and S3), no*

S2: What are the effects of the differences? *certainly, very likely, likely, unlikely, very unlikely, certainly not*

(1) Could the differences lead to **misunderstandings** between you and the other person?

S3: What are the causes of the differences? *certainly, very likely, likely, unlikely, very unlikely, certainly not*

(1) Would you attribute the differences to **intentional editing**, *e.g.*, retouching, filters, or manipulation?
(2) Would you attribute the differences to **uncontrolled distortions**, *e.g.*, transmission errors or image compression?

A1: Action page.
[Appeared once after the first image comparison block.]
Thank you – We will now repeat this for different images.
Please don't get distracted and evaluate each picture carefully.

Part 4: Control Variables

C1: Please indicate your **theoretical knowledge** of the following technologies *never heard of it, have heard of it, can explain how it works, profound knowledge*

(1) Image retouching, *e.g.*, color corrections, removal of "blemishes", smoothing, or sharpening of textures
(2) Image montage, *e.g.*, adding, modifying, or removing objects with the cloning tool
(3) AI-supported image generation, *e.g.*, DALL·E, MidJourney, or Stable Diffusion
(4) Generative inpainting, *e.g.*, with DeepFill, Adobe Firefly Generative Fill, or DALL·E
(5) Digital photography, *e.g.*, with a smartphone or digital camera
(6) Conventional lossy image compression, *e.g.*, JPEG or WEBP
(7) Neural image compression, *e.g.*, with algorithms like JPEG AI or HiFiC
(8) Virtual image compression, *e.g.*, with codecs like VBC Optimizer or SpectraZip

C2: Please indicate your **practical experience** with the following technologies. *no practical experience, have tried it, use occasionally, use regularly*
[*cf.* enumeration from C1]

C3: Please indicate how often you verify the authenticity of images from various sources, *e.g.*, by zooming in on the image. I verify in image … … *almost always., … often., … occasionally., … almost never., not applicable*

(1) Social network profile of a public person or organization I know
(2) Social network profile of a public person or organization I **don't** know
(3) Social network profile of a private person I know
(4) Social network profile of a private person I **don't** know
(5) Private direct message (e.g., SnapChat, Instagram, WhatsApp)
(6) Reputable online news site

C4: What is your experience with the distribution of images on the internet? *happens to me regularly, has happened to me, could happen to me, could rather not happen to me, don't know*

(1) How realistic is the scenario with the photo that gets widely distributed online?
(2) And how realistic is it that the photo gets modified in this process?

F1: Please take a moment to share your feedback on the study, *e.g.*, the clarity of the questions, the selection and presentation of the images, any difficulties in answering, *etc.*

[Open question field]

End page

Thank you for your participation!
If you are interested in the results of the study, please provide us with your email address. The address will be stored separately from your responses. [Link to survey]

B.3 Stimuli Compression

The source images were taken from two widely-used image compression benchmark datasets [3, 79]. For the compression of our test images we selected four state-of-the-art neural image compression codecs that cover the range of technologies applied in the neural compression literature. For the compression with the Hyperprior [7] we used the hierarchical mode optimized for MSE at compression intensity 3 out of 8. For the compression with the GAN based HiFiC model [62] we used the intensities Hi or Lo. We used TensorFlow Compression library [4] (TFC) for the compression with these two models and pretrained weights. For the compression with JPEG AI [5], we used the verification model of its Reference Software [74] at version 7.0, commit 50EC1478. During encoding we use high operation point (HOP)[48] and disable all tools. We include images for the target bits per pixel (BPP) values 0.25 and 0.75. Lastly, we use the code repository of the authors and their pretrained weights to compress images with the diffusion based CDC model [88] using x-parameterization for LPIPS weight 0.9 and Lagrangian multiplier to control the compression quality of 2048 and 0512. All codecs were executed on a shared GPU cluster using a 64-core Nvidia A100 GPU.

To compress the JPEG control images we used the Python Imaging Library (PIL) version 11.1.0 with `libjpeg-turbo`, version 3.0 [8] at the default quality factor 75.

[4] https://github.com/tensorflow/compression

B.4 Stimuli Overview

Table 2: Compression specification and verbal description of semantic changes of stimulus images

Test image	Crop Size	Compression codec	BPP[1)]	Position in instrument	Source dataset	Semantic change
Bag–M	256^2	CDC-2048x09	0.30	11	CLIC	The color of the bag changes from purple to blue.
Bracelet–M	256^2	CDC-2048x09	0.29	3	CLIC	The texture changes from a beaded to a knotted look.
Bracelet–J	256^2	JPEG-75	1.03	5	CLIC	-
Brake lights–M	128^2	HiFiC-Hi	0.47	9	DIV2K	The van's brake lights are turned off.
Brake lights–J	128^2	JPEG-75	1.27	9	DIV2K	-
Burj Khalifa–M	256^2	Hyper-MSE 3	0.19	3	DIV2K	Color appears on the façade that could resemble paint.
Burj Khalifa–C	256^2	JPEG AI-0.25	0.22	10	DIV2K	-
Camera–M	256^2	HiFiC-Lo	0.09	1	DIV2K	The engraved number 8 changes into the number 6.
Cardigan–C	512^2	CDC-2048x09	0.28	4	CLIC	-
Cardigan–J	512^2	JPEG-75	1.35	9	CLIC	-
Church–M	256^2	HiFiC-Lo	0.08	6	CLIC	The steeple's cross changes into a star.
Church–C	256^2	CDC-0512x09	0.47	10	CLIC	-
Paris–M	512^2	HiFiC-Lo	0.17	8	CLIC	The people in the grass disappear, and the crowd on the stairs turns into a black blur.
Paris–J	512^2	JPEG-75	1.25	4	CLIC	-
Rings–M	256^2	HiFiC-Lo	0.14	10	CLIC	The thumb ring and the necklace pendant disappear.
Rings–C	256^2	Hyper-MSE 3	0.20	4	CLIC	-
Road–U	512^2	-	-	12	CLIC	-
Rock house–M	128^2	HiFiC-Lo	0.14	10	DIV2K	The house built into the rock is unrecognizable because its door, windows, and roof disappear.
Rock house–J	128^2	JPEG-75	1.06	5	DIV2K	-
Santorini–J	256^2	JPEG-75	1.24	2	DIV2K	-
Suit–M	256^2	HiFiC-Lo	0.17	7	DIV2K	The man's hand and facial features disappear.
Suit–C	256^2	JPEG AI-0.25	0.26	3	DIV2K	-
Tattoo–M	256^2	CDC-2048x09	0.29	9	CLIC	The tattoo turns from a crescent moon into a full moon.
Tattoo–C	256^2	CDC-0512x09	0.69	7	CLIC	
Traffic l.–M (arrow)	128^2	CDC-2048x09	0.22	3	CLIC	The arrow-shaped traffic light has turned into a regular, round traffic light.
Traffic l.–M (sign)	128^2	HiFiC-Hi	0.36	5	CLIC	The "No Cars" sign has turned into a "No Cameras" sign.
Traffic l.–C	128^2	JPEG AI-0.25	0.26	4	CLIC	
Train–M	256^2	HiFiC-Lo	0.17	6	DIV2K	The head of a person leaning out the window disappears.
Train–J	256^2	JPEG-75	1.18	6	DIV2K	
Watch (broken)–M	256^2	CDC-2048x09	0.19	8	CLIC	The smartwatch looks broken.
Watch (off)–M	256^2	HiFiC-Lo	0.09	7	CLIC	The display of the smartwatch is turned off.
Watch–J	256^2	JPEG-75	0.72	8	CLIC	-
Wind sock–M	256^2	JPEG AI-0.25	0.22	6	CLIC	The red wind sock disappears.
Wind sock–C	256^2	HiFiC-Lo	0.08	8	CLIC	-
Window shades–M	128^2	HiFiC-Hi	0.36	7	CLIC	The open window shades look closed.
Window shades–J	128^2	JPEG-75	0.61	5	CLIC	-

[1)] measured for the full image.

B.5 Images

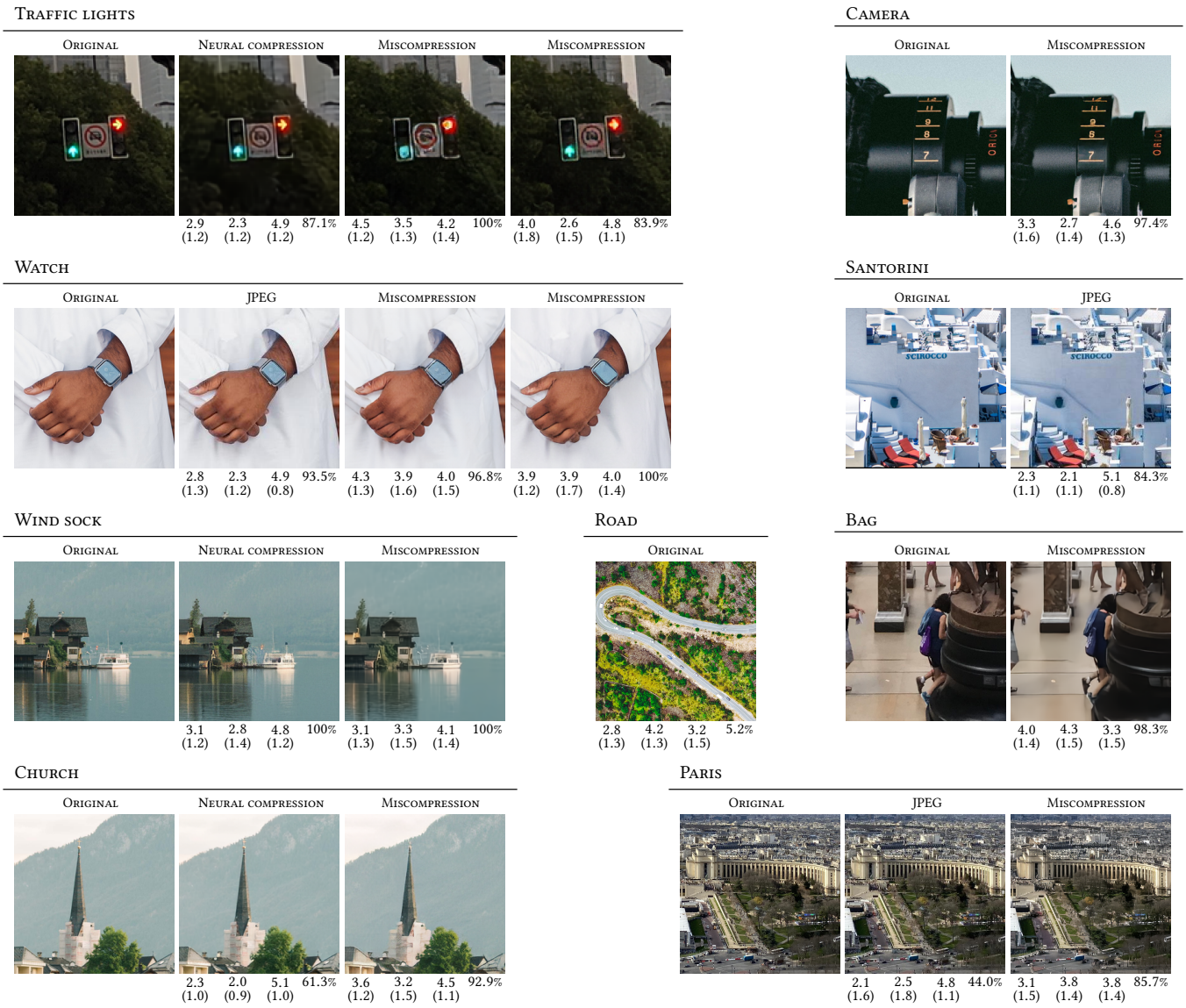
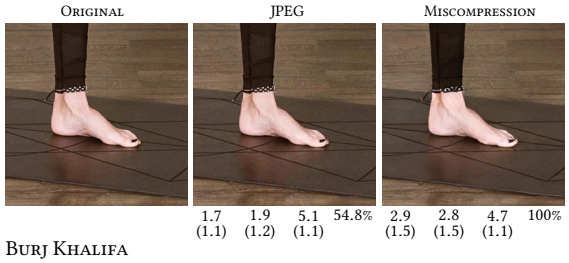
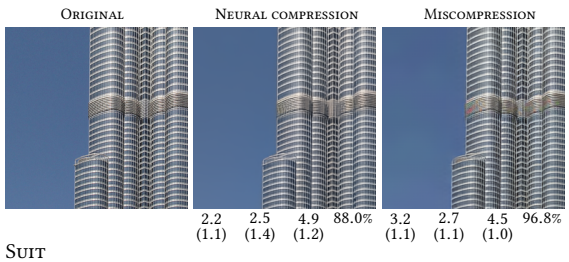


Figure 12: Stimulus material with per-image descriptive statistics. Camera, Bag, Santorini and Road were shown to all groups. Images are best viewed on screen and magnified. See Figure 13 on the next page for the legend.

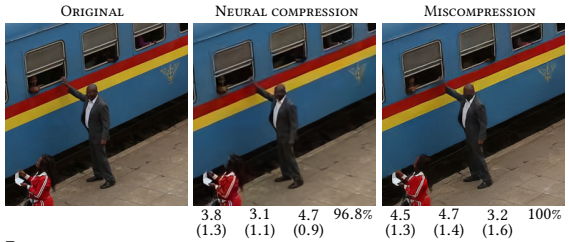
Bracelet



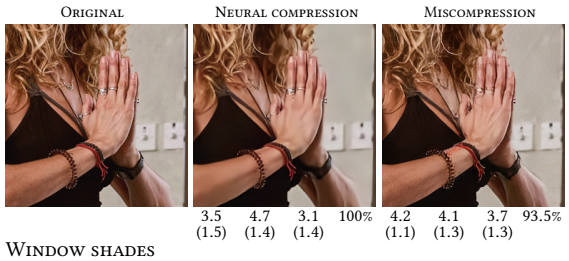
Burj Khalifa



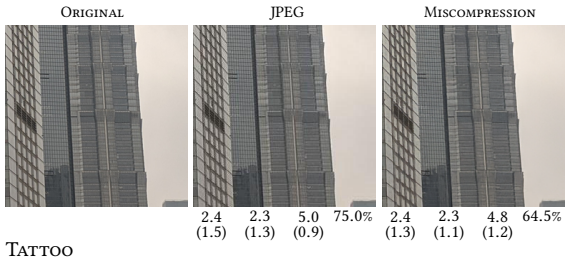
Suit



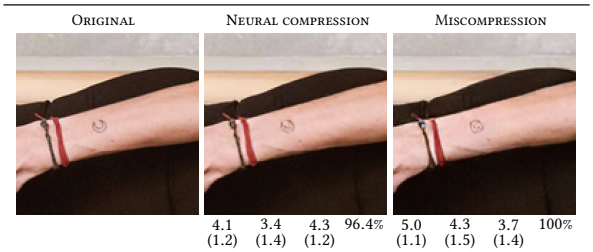
Rings



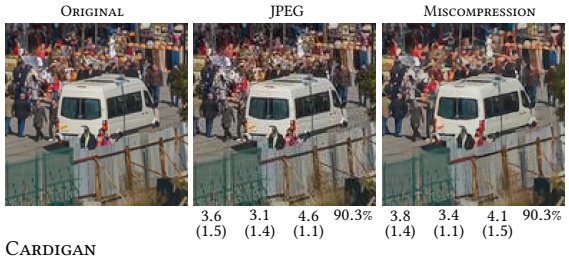
Window shades



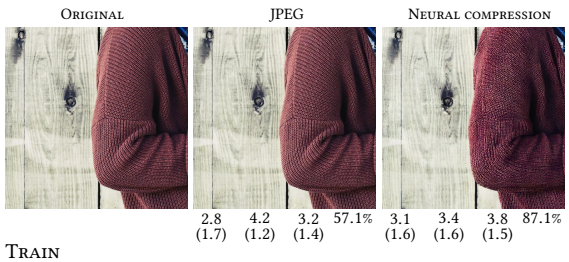
Tattoo



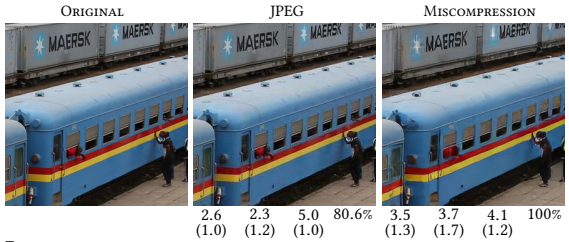
Brake lights



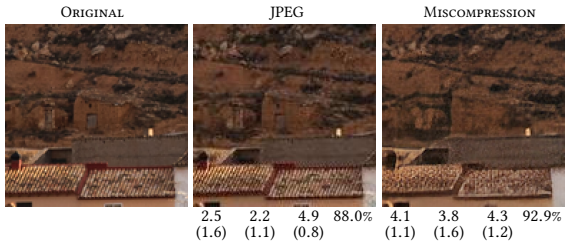
Cardigan



Train



Rock house



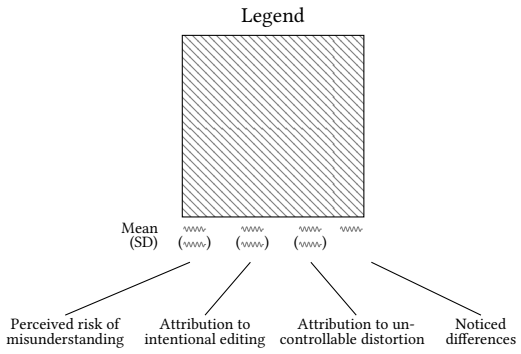


Figure 13: Stimulus material with per-image descriptive statistics (continued).

C Results

C.1 Supplemental Figure

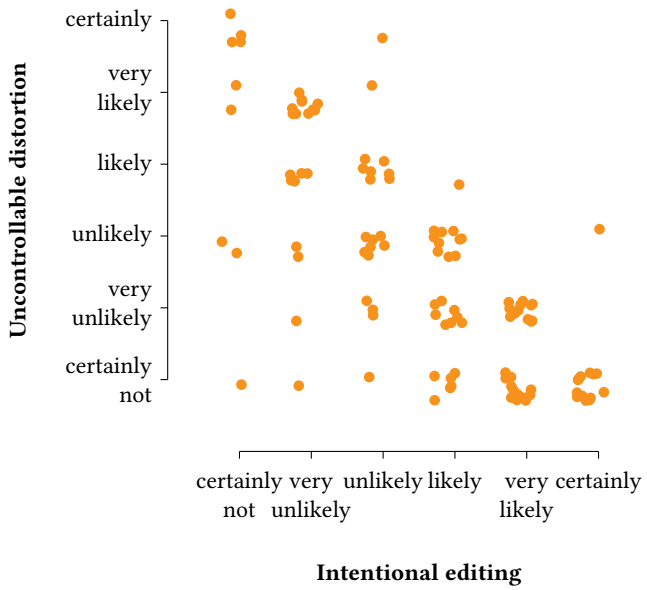


Figure 14: Suspected causes of the differences in the M-Bag. Unlike Figure 11, this scatter plot shows responses of individual participants (with jitter applied for visibility). The negative correlation is not perfect as rejecting one cause does not imply the support of the other.

C.2 Supplemental Tables of Control Questions

Table 3: Descriptive statistics of the control questions

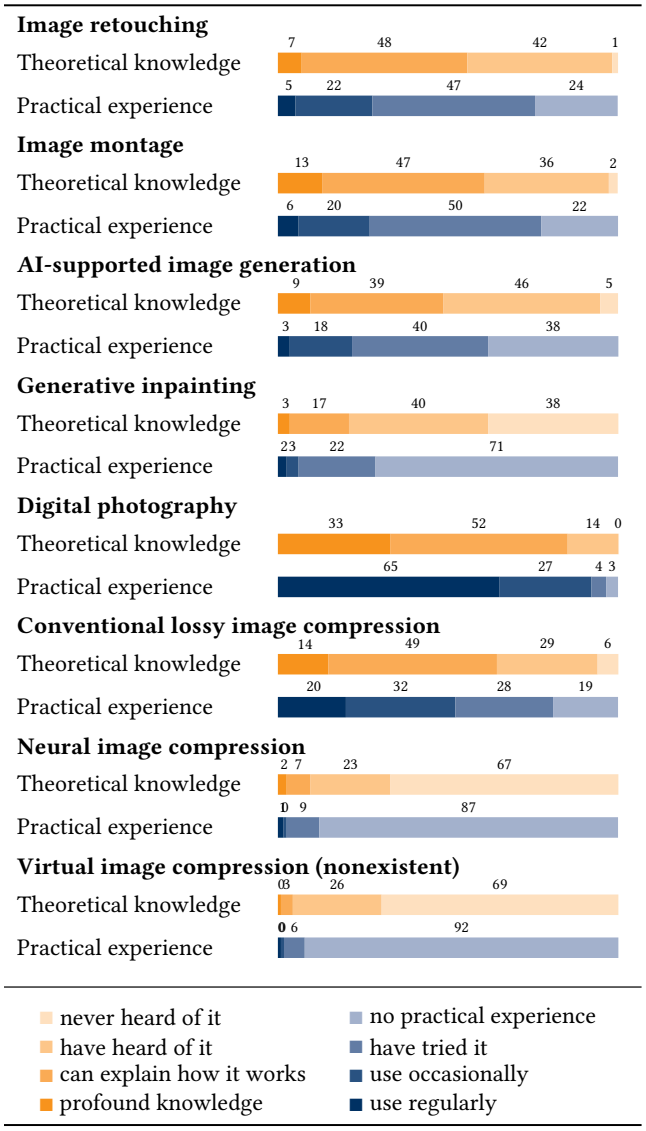


Table 4: Descriptive statistics of the control questions (cont'd)

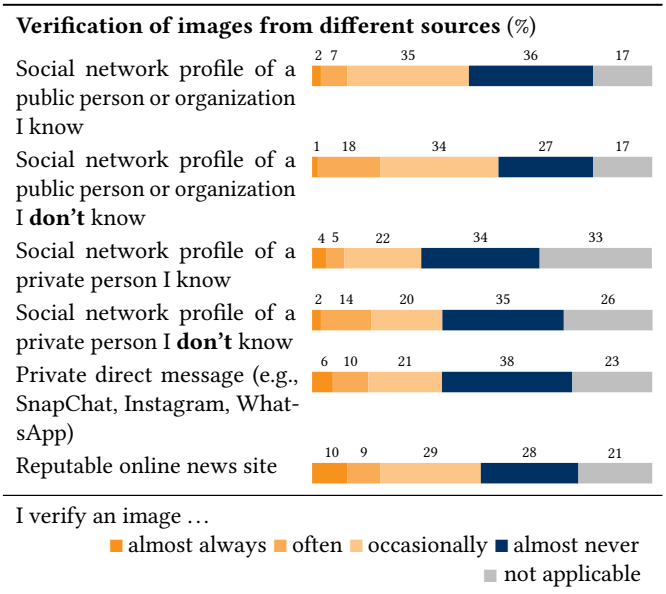


Table 5: Descriptive statistics of the control questions (cont'd)