

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Exploring the Influence of Verbal and Nonverbal Similarities on the Verbal Overshadowing Effect in Facial Recognition

Permalink

<https://escholarship.org/uc/item/5tb445m7>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Tsukamura, Yuki
Okada, Kensuke

Publication Date

2024

Peer reviewed

Exploring the Influence of Verbal and Nonverbal Similarities on the Verbal Overshadowing Effect in Facial Recognition

Yuki Tsukamura (tsukkacogsci@gmail.com)

Graduate School of Education, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.
Japan Society for the Promotion of Science
5-3-1 Kojimachi, Chiyoda-ku, Tokyo 102-0083, Japan.

Kensuke Okada (ken@p.u-tokyo.ac.jp)

Graduate School of Education, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

Abstract

The verbal overshadowing effect, a phenomenon where verbal descriptions of an encoded face hinder subsequent recognition, has been linked to the similarity in facial image sets used in recognition tasks. However, the specific aspects of similarity that influence this effect remained underexplored. This study, therefore, employed the Stable Diffusion image-generation model to create image sets that are similar either verbally or nonverbally. Experimental results using these sets revealed the presence of the verbal overshadowing effect in the verbally-similar set, but it was not evident in the nonverbally-similar set. These findings align with existing explanations of the verbal overshadowing effect and contribute to enhancing its predictability.

Keywords: facial recognition, verbal overshadowing, verbal similarity, image-generation models

Introduction

The verbal overshadowing effect denotes the disruption in facial memory due to the verbalization of facial features before recognition (Schooler & Engstler-Schooler, 1990). This effect is crucial for understanding the structure of human memory representation, but the cognitive processes leading to it are still under debate (Baker & Reysen, 2021).

Two main theoretical accounts explain the verbal overshadowing effect: the retrieval-based interference account and the transfer-inappropriate processing shift account. The retrieval-based interference account suggests that verbalization generates incorrect information about the encoded facial features, thereby negatively impacting recognition accuracy. Evidence for this theory includes the negative correlation between erroneous verbalization and successful recognition, and the positive correlation between accurate verbalization and recognition performance (Schooler & Engstler-Schooler, 1990; Meissner, Brigham, & Kelley, 2001). However, some studies report the absence of this correlation (Baker & Reysen, 2021; Kitagami, Sato, & Yoshikawa, 2002), raising doubts about the validity of the retrieval-based interference account.

The transfer-inappropriate processing shift account emphasizes the difference between configural processing, used in facial encoding and recognition, and featural processing, used in the verbal description of faces (Schooler, Fiore, & Brandimonte, 1997). This account proposes that engaging in featural processing, such as during verbalization, temporarily impedes the availability of configural processing, leading to the

























verbal overshadowing effect. Numerous findings can be explained by the transfer-inappropriate processing shift account but not by the retrieval-based interference account (for a detailed explanation, see Baker & Reysen, 2021). These include the interference effects to non-verbalized targets (Dodson, Johnson, & Schooler, 1997; Brown & Lloyd-Jones, 2002; Lloyd-Jones & Brown, 2008), the enhancement effects resulting from engagement in facial composite recall tasks that include visual processing (Meissner & Brigham, 2001), the gradual decrease of the verbal overshadowing effect over time after verbalization (Meissner et al., 2001; Finger & Pezdek, 1999), and the weakening of the verbal overshadowing effect after engaging in non-verbal tasks (Finger, 2002).

However, it is crucial to note that while there is substantial indirect evidence supporting the transfer-inappropriate processing shift account, direct evidence is relatively limited (Baker & Reysen, 2021). Furthermore, this account is abstract and lacks predictive power in specific scenarios. For instance, the distinction between configural and featural processing is not clearly defined (e.g., Piepers & Robbins, 2012). As a result, it is hard to predict the size of the verbal overshadowing effect produced by a given set of images, at least directly, by the transfer-inappropriate processing shift account.

In this study, we focus on the influence of stimulus characteristics on the verbal overshadowing effect. Large-scale replication experiments have confirmed the verbal overshadowing effect as a consistent phenomenon, particularly with certain stimuli (Alogna et al., 2014). However, some studies have shown that verbalization of facial features facilitates facial recognition (Brown & Lloyd-Jones, 2005; Jones, Armstrong, Casey, Burson, & Memon, 2013). Several studies also reported that the verbal overshadowing effect was not replicated (for some unpublished cases, see Kitagami et al., 2002), suggesting that the effect may vary depending on the test set used.

Kitagami et al. (2002) highlighted how the verbal overshadowing effect is influenced by the similarity between the target stimulus and the distractor stimuli on the verbal overshadowing effect. They used morphing techniques to create composite facial images and reported that images closely resembling the average face (i.e., strongly morphed) induced the verbal overshadowing effect, whereas this effect was not

Table 1: The sets of generated images used in the experiment.

condition	set	target	distractors				
verbally-similar	1						
	2						
nonverbally-similar	3						
	4						

evident in images further from the average. This suggests that the verbal overshadowing effect depends on the similarity of the stimuli. However, the exact meaning of “similarity” in this context remains ambiguous. It might imply a holistic visual resemblance in facial characteristics, often difficult to verbalize, or it may relate to the similarity in verbal descriptions of the faces. Previous studies have not clearly distinguished these different meanings of “similarity.”

In this study, the effects of verbal overshadowing are examined under two conditions: a verbally-similar condition, where the easily verbalized features of the target stimulus are shared with the distractors, but the difficult-to-verbalize features are not; and a nonverbally-similar condition, where the easily verbalized features are not shared, but the difficult-to-verbalize features are. Given the limited availability of facial images, controlling these similarities poses a challenge. This study addresses this issue by employing face images generated from an image-generation model as stimuli. We hypothesize that the verbal overshadowing effect is observable in the verbally-similar condition. Conversely, based on the previous studies suggesting that the verbal overshadowing effect does not arise in verbalization-dependent tasks (Schooler & Engstler-Schooler, 1990; Schooler et al., 1997), we hypothesize that this effect is not observed in nonverbally similar conditions.

Method

The experimental protocols in this study were approved by the ethics committee of the first author’s institution. The

ethics review was performed in accordance with the principles of the Declaration of Helsinki.

Participants

The participants ($N = 76$) were undergraduate students at a Japanese university, aged 19-22 years ($M = 19.38, SD = 0.66$). They participated in the experiment as part of a psychological experiment class. Of the participants, 29 were male, 46 were female, and one chose not to disclose their gender.

Participants were randomly assigned to either the verbalization ($N = 37$) or control ($N = 39$) groups. Data from two participants in the verbalization group were excluded from the analysis because corruption occurred in the data due to encoding issues. Additionally, based on previous results indicating that the size of the verbal overshadowing effect varies with perceptual expertise (Fallshore & Schooler, 1995), data from two individuals (one each from the verbalization and control groups) who had resided in Asia for less than 90% of their lives were also excluded.

Design

The experiment employed a two-factor mixed design. The within-participants factor was the similarity between the target and distractors (2 levels: verbally-similar and nonverbally-similar), and the between-participants factor was whether the target images were verbalized (2 levels: verbalization and control). Note that the 1st factor (similarity) was repeated.

Materials and Image Generation

Each trial in this experiment used six images, referred to as a “set.” Both groups underwent two trials in the verbally-similar condition and two trials in the nonverbally-similar condition. Thus, four sets, totaling 24 facial images, were created.

The image generation model, Stable Diffusion version 2.1 (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022), was used to create the images. All facial images were generated using “Japanese” as prompts. Additionally, two sets under each condition were generated using “male” and “female” as prompts, respectively. Furthermore, within each set, the clothing, background, and poses were made as uniform as possible. ControlNet OpenPose was used for controlling pose (Zhang, Rao, & Agrawala, 2023)¹. The image used for controlling pose was created with reference to the test stimuli used in the large-scale replication experiment (Alogna et al., 2014).

Images used in the verbally-similar condition were generated by maintaining a consistent prompt within the set while varying the initial seed. For instance, all images in set 1 (Table 1) were generated with a positive prompt of “a portrait of a short-haired Japanese male college student in black suit with white background, 20 years old, looking at camera.” Given the frequent production of images featuring smiles, a negative prompt “smile” was also utilized concurrently to control the facial expressions. Similarly, images in set 2 were generated using “female, beautiful, a portrait of a long-haired Japanese woman in gray suit with white background, 20 years old, looking at camera” as a positive prompt. The terms “female, beautiful” were added because generating natural-looking female faces was challenging in the earlier setting.

Conversely, the images used in the nonverbally-similar condition were generated by maintaining a constant initial seed within the set while modifying the prompt. This approach was adopted because Stable Diffusion tends to produce similar images with the same initial seed (e.g., Dehouche & Dehouche, 2023). Elements such as hair and beards were altered among the images in the prompt, as these features can be consistently described and easily manipulated. For instance, to generate the target image of set 3 (Table 1), the positive prompt “a portrait of a short-haired Japanese male college student in white collared shirt with white background, black eyes, 20 years old, looking at camera” was used, along with the negative prompt “smile,” for reasons identical to those for set 1. The distractor images in set 3 were created by combining the following three alterations: changing “short-haired” to “long-haired,” adding “bearded” to the positive prompt, and incorporating “bangs” into the negative prompt. Similarly, in set 4, the target image was generated using the prompt: “a portrait of a brown wavy-haired long-

haired Japanese woman in gray suit with white background, 20 years old, looking at camera.” The distractors were created by varying this prompt in three ways: changing “short-haired” to “long-haired,” “brown” to “black,” and removing “wavy-haired.”

The image generation process involved iterative modification of the seed until images that satisfied the criteria and appeared coherent were produced. Notably, for images under the verbally-similar condition, techniques such as img2img and inpainting were utilized to refine aspects like facial features and attire.

Procedures

The experiment was conducted over six class sessions between May and September of 2023. Participants were organized into six groups (one group for each class session), each consisting of 10 to 14 individuals. Participants used their own computers for the experiment. Initially, they provided information about their age, duration of residence in Asia, and gender.

The main trials commenced thereafter. At the start of each trial, participants were shown a target image for five seconds. Subsequently, participants in the verbalization group were instructed to write down as many details as possible about the facial features of the image over five minutes. Conversely, participants in the control group were asked to engage in an unrelated task for the same duration, which involved listing names of prefectures in Japan, Japanese towns and cities, countries around the world, and city names around the world. After the verbalization or unrelated task, both groups participated in a recognition task. In this task, they were presented with a set of six facial images, which consists of the initially shown image and five distractors, arranged in a 2×3 grid in a random order. They were asked to either identify the original image or indicate that it was not present. This procedure was repeated four times, once for each set, with the order of the sets randomized for each participant. The procedures were adapted from the methodology used in Experiment 6 of Schooler and Engstler-Schooler (1990). However, unlike the original experiment, which included a five-minute filler task prior to verbalization, our experiment omitted this due to time constraints.

Finally, the participants received feedback that the images employed in the study were created by an image-generation model. Subsequently, they were asked whether they had realized that the images were generated by such a model and were not photographs of real people. Participants could respond with “Yes,” “No,” or “Uncertain.” Those who answered “Yes” were also asked what had led them to the realization.

Results

FaceNet-based similarity of facial images

Prior to analyzing the experimental results, the similarity between facial images was assessed using a different metric by employing FaceNet (Schroff, Kalenichenko, & Philbin,

¹We used a trained ControlNet OpenPose model for Stable Diffusion 2.1: https://huggingface.co/thibaud/controlnet-sd21/blob/main/control_v1lp_sd21_openposev2.safetensors



Figure 1: The cosine similarity matrices for FaceNet-computed embeddings of facial images in each stimulus set.

2015) to evaluate facial resemblance. Each facial image was cropped using the Multi-Task Cascaded Convolutional Neural Networks (MTCNN) face detection model and then transformed into 512-dimensional embeddings by a pre-trained FaceNet model². The cosine similarity of these vectors was used to define the FaceNet-based similarity between facial images. The FaceNet-based similarity for all possible pairings in each stimulus set is presented in Figure 1. Specifically, the FaceNet-based similarity between the target image and each of the five distractors was focused. The average of these five computed values was 0.614 and 0.604 for verbally-similar sets 1 and 2, respectively. For the nonverbally-similar sets 3 and 4, the values were 0.614 and 0.857, respectively. This suggests that the FaceNet-based similarity between the target and distractors is on par or higher for nonverbally-

similar sets.

Recognition Performance

The analysis was conducted using R (version 4.3.2). Figure 2 depicts the recognition accuracy for each group under every condition. We used a generalized linear mixed model (GLMM) to examine the influence of group and condition on the likelihood of correct responses. The model incorporated group, condition, and their interaction as fixed effects, with the effect of participants as a random effect. Given the binary nature of the dependent variable (correct/incorrect), a Bernoulli distribution was specified. The equation for the model is as follows:

$$\text{logit}(P(\text{correct})) = \beta_0 + \beta_1 \times \text{group} + \beta_2 \times \text{condition} + \beta_3 \times \text{group} \times \text{condition} + u_i \quad (1)$$

²We used a FaceNet model pretrained on VGGFace2: <https://github.com/timesler/facenet-pytorch>

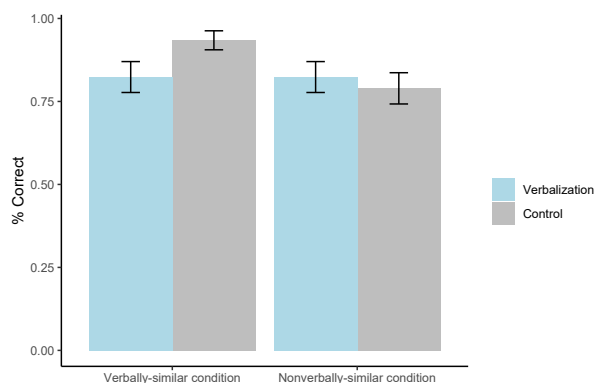


Figure 2: Percentages of correct face recognitions of each group \times condition. Error bars indicate the standard errors.

Note that u_i in this equation indicates the random effect for participant i .

The GLMM analysis was performed using the lme4 package (version 1.1.35.1). The model was fitted using maximum likelihood estimation with Laplace Approximation. The results are presented in Table 2. Notably, the fixed effect of the interaction between group and condition was not statistically significant ($\beta_3 = 1.39, z = 1.92, p = .054$), indicating that the differences between groups (i.e., the size of the verbal overshadowing effect) did not vary significantly across conditions.

A primary objective of this research was to investigate the magnitude of the verbal overshadowing effect in each similarity condition. Consequently, an additional GLMM analysis using a Bernoulli distribution was conducted for each condition. The equation for the model is as follows:

$$\text{logit}(P(\text{correct})) = \beta_0 + \beta_1 \times \text{group} + u_i. \quad (2)$$

Table 3 presents the results. Notably, the effect of the group was statistically significant in the verbally-similar condition ($\beta_1 = 1.11, z = 1.98, p = .047$) but not in the nonverbally-similar condition ($\beta_1 = 0.27, z = 0.47, p = .635$). These results suggest that the verbal overshadowing effect was observed for verbally-similar images, while its presence for nonverbally-similar images was not as evident.

Effect of Recognizing Machine-Generation

Finally, a brief analysis was conducted to determine whether the participants' realization that the images were machine-generated was related to their recognition performance.

To the question whether they noticed that the images used were machine-generated, 30 (41.7%) of 72 participants answered "Yes," and 32 (44.4%) answered "No." Incorporating these responses as an independent variable in the GLMM model (1) did not result in an improvement in the model's Akaike Information Criterion ($\Delta AIC = 3.68$).

Discussion

The study aimed to elucidate the types of "similarity" in stimuli that are associated with the verbal overshadowing effect. Utilizing the Stable Diffusion image-generation model, we created sets of images that were either verbally-similar or nonverbally-similar. Experimental investigations using these stimuli revealed the occurrence of the verbal overshadowing effect in the verbally-similar set, while no evidence of the effect was found in the nonverbally-similar set.

Interestingly, in terms of FaceNet-based similarity, the similarity was either equivalent or higher in the nonverbally-similar set. On the contrary, the verbal overshadowing effect was actually observed in the verbally-similar condition. Previous studies have posited that a high degree of "similarity" leads to the verbal overshadowing effect, but our findings indicate a more complex nature of "similarity" and suggest that "verbal similarity" within the recognition task set is a key factor that contributes to the verbal overshadowing effect. However, the use of FaceNet involves cropping the facial region, which implies that the modified parts within verbally-similar sets such as the hair or beard may not be entirely processed by FaceNet. This could potentially affect the validity of FaceNet-similarity. To improve this, altering the image generation process to ensure valid similarity calculations by FaceNet, or the use of more appropriate similarity metrics, may be required.

These findings align with the transfer-inappropriate processing shift account. This account posits that during a verbalization task, configural processing becomes temporarily unavailable, leading to the activation of featural processing during recognition. According to this account, the verbal overshadowing effect in the verbally-similar sets can be interpreted as occurring due to the difficulty of distinguishing based solely on featural characteristics. In contrast, in the nonverbally-similar sets, the featural attributes alone were sufficient for distinction, hence no significant effect was observed. This interpretation is also consistent with the results of previous studies that the verbal overshadowing effect is manipulated by whether the verbalized information is important for face identification (Hatano, Kitagami, & Kawaguchi, 2014).

This research provides key insights for understanding when interference effects occur, thereby enhancing the predictability of verbal overshadowing. While the results do not allow for immediate prediction of the effects of verbal overshadowing on given targets and distractors, the clarification of the meaning of "similarity" moved us towards a more predictable framework. Future research is expected to build on these findings to create models that can predict verbal overshadowing from given stimuli.

Use of Image Generation Models

A distinguishing feature of this study is its use of images generated by image-generation models in its experiments. Despite these technologies still being in development, as demon-

Table 2: Summary of fixed effects estimates in the GLMM analysis for the entire data.

Coefficient	Estimate (SE)	<i>z</i> value	<i>p</i> value
β_0 (Intercept)	1.45 (0.34)	4.23	< .001
β_1 (Group)	0.23 (0.47)	0.49	.625
β_2 (Condition)	1.39 (0.56)	2.50	.012
β_3 (Group \times Condition)	1.39 (0.72)	1.92	.054

Table 3: Summary of fixed effects estimates in the GLMM analysis for each similarity condition.

Condition	Coefficient	Estimate (SE)	<i>z</i> value	<i>p</i> value
Verbally-similar	β_0 (Intercept)	2.65 (0.46)	5.73	< .001
	β_1 (Group)	1.11 (0.56)	1.98	.047
Nonverbally-similar	β_0 (Intercept)	1.73 (0.51)	3.41	< .001
	β_1 (Group)	0.27 (0.56)	0.47	.635

strated, they hold potential benefits for cognitive science studies.

Fundamentally, for empirical experiments of the verbal overshadowing effect, the use of facial images as stimuli is essential. However, the use of facial images is subject to various constraints. For instance, (1) there is a finite number of faces in the world, (2) rights-related restrictions may apply even when a desired face is available, and (3) even if legal issues are resolved, considerable effort may be required to modify images appropriately for research.

The application of image-generation models theoretically has the potential to overcome these limitations. This includes: (1) approximating the distribution of facial images through deep learning, thereby enabling the generation of plausible yet non-existent facial images; (2) typically encountering fewer legal restrictions, allowing for more unrestricted application; (3) the ability to easily control the characteristics of the generated images through conditioning on variables such as text or poses.

Furthermore, the use of image generation models in research is expected to enhance the open-access availability of research materials. In fact, in studies on the verbal overshadowing effect, experimental stimuli are often not disclosed due to various constraints. On the other hand, images generated by Stable Diffusion 2.1 are available for use under the CreativeML Open RAIL++-M license.

However, it is crucial to acknowledge that images resembling realistic facial photos are not always processed in the same way as actual faces, which presents a complex problem. We can, at least, partially confirm the impact of realizing that the images were machine-generated can be partially confirmed from the data. Inputting the self-reports on whether or not they noticed this into the model did not improve the AIC. This suggests that recognizing the machine-generation does not significantly influence recognition performance in this study.

Limitations

The results of the primary statistical analyses in this study approach significance, with *p* values nearing .05. This necessitates additional replication studies to confirm the robustness of the effect. The borderline nature of these results could be attributed to various factors.

First, the timing of verbalization might have influenced the results. Unlike Experiment 6 of Schooler and Engstler-Schooler (1990), which incorporated a 5-minute filler task between the encoding of images and verbalization, our study omitted such a task, leading to immediate verbalization after image viewing. Alogna et al. (2014) reported that the verbal overshadowing effect is more pronounced when a filler task is used between encoding and verbalization than when it is used between verbalization and recognition, indicating a potential limitation in our study design.

Additionally, a ceiling effect was observed, as the average recognition rate across all groups and conditions exceeded 0.75 (as shown in Figure 2). Notably, the control group in the verbally-similar condition exhibited an exceptionally high average recognition rate of 0.934. Factors contributing to this ceiling effect include a shorter interval between stimulus encoding and recognition compared to previous studies, and the use of images with greater clarity than those in earlier research such as Kitagami et al. (2002).

Acknowledgments

This work was supported by JSPS KAKENHI Grant number JP23KJ0793.

References

- Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., . . . Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9(5), 556–578.
- Baker, M. A., & Reysen, M. B. (2021). Using intentional and incidental encoding instructions to test the transfer inappropriate processing shift account of verbal overshadowing. *Journal of Cognitive Psychology*, 1–16.

- Brown, C., & Lloyd-Jones, T. J. (2002). Verbal overshadowing in a multiple face presentation paradigm: effects of description instruction. *Applied Cognitive Psychology, 16*(8), 873–885.
- Brown, C., & Lloyd-Jones, T. J. (2005). Verbal facilitation of face recognition. *Memory & Cognition, 33*(8), 1442–1456.
- Dehouche, N., & Dehouche, K. (2023). What’s in a text-to-image prompt? the potential of stable diffusion in visual arts education. *Heliyon, 9*(6), e16757.
- Dodson, C. S., Johnson, M. K., & Schooler, J. W. (1997). The verbal overshadowing effect: why descriptions impair face recognition. *Memory & Cognition, 25*(2), 129–139.
- Fallshore, M., & Schooler, J. W. (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(6), 1608–1623.
- Finger, K. (2002). Mazes and music: using perceptual processing to release verbal overshadowing. *Applied Cognitive Psychology, 16*(8), 887–896.
- Finger, K., & Pezdek, K. (1999). The effect of the cognitive interview on face identification accuracy: release from verbal overshadowing. *The Journal of Applied Psychology, 84*(3), 340–348.
- Hatano, A., Kitagami, S., & Kawaguchi, J. (2014). Verbalizing information salient to face identification does not cause verbal overshadowing. *Comprehensive Psychology, 3*.
- Jones, T. C., Armstrong, R., Casey, A., Burson, R. A., & Memon, A. (2013). Verbal description benefits for faces when description conditions are unknown a priori. *Quarterly Journal of Experimental Psychology, 66*(9), 1818–1839.
- Kitagami, S., Sato, W., & Yoshikawa, S. (2002). The influence of test-set similarity in verbal overshadowing. *Applied Cognitive Psychology, 16*(8), 963–972.
- Lloyd-Jones, T. J., & Brown, C. (2008). Verbal overshadowing of multiple face recognition: Effects on remembering and knowing over time. *The European Journal of Cognitive Psychology, 20*(3), 456–477.
- Meissner, C. A., & Brigham, J. C. (2001). A meta - analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15*(6), 603–616.
- Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition, 29*(1), 176–186.
- Piepers, D. W., & Robbins, R. A. (2012). A review and clarification of the terms “holistic,” “configural,” and “relational” in the face perception literature. *Frontiers in Psychology, 3*, 559.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10684–10695).
- Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology, 22*(1), 36–71.
- Schooler, J. W., Fiore, S. M., & Brandimonte, M. A. (1997). At a loss from words: Verbal overshadowing of perceptual memories. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 37, pp. 291–340). San Diego, CA, US: Academic Press.
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 815–823*.
- Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3836–3847).