Received: 18 April 2023 | Accepted: 29 November 2023

DOI: 10.1111/bjet.13419

British Journal of Educational Technology BERA

ORIGINAL ARTICLE

# How to train students to engage in text-picture integration for multimedia lessons

Xiaoxue Leng[1] | Fuxing Wang[1] | Richard E. Mayer[2] | Tingting Zhao[1]

[1]School of Psychology, Central China Normal University, Wuhan, China

[2]Department of Psychological and Brain Sciences, University of California, Santa Barbara, Santa Barbara, California, USA

**Correspondence**
Fuxing Wang, School of Psychology, Central China Normal University, 152 Luoyu Street, Wuhan, Hubei, 430079, P.R. China.
Email: fxwang@ccnu.edu.cn

**Funding information**
National Natural Science Foundation of China, Grant/Award Number: 62277025

This study investigated the effectiveness of visual training or verbal training on how to use a text-picture processing strategy for learning from computer-based multimedia instructional material. Sixty-nine university students were randomly assigned to the verbal training group (students received text-based instruction for a text-picture processing strategy), the visual training group (students observed a video depicting an expert's eye fixations while using a text-picture processing strategy for an initial portion of a multimedia lesson) or the control group (students did not receive any instruction). During reading a multimedia lesson on biology, students' eye movements were tracked; and after the lesson, students took a posttest. Concerning learning outcomes, both visual and verbal training helped students perform better than the control group on a recall test and the verbal training group perform better on a transfer test. Concerning learning processes, both visual and verbal training caused students to attend less to on-screen text and more to on-screen pictures as compared to the control group. Mediation analysis showed that increased attention to pictures was a mediator for better learning outcomes. Practical and theoretical implications are discussed.

**KEYWORDS**
eye movement modelling examples, multimedia learning, pre-training principle, strategy acquisition, text-picture integration

**Practitioner notes**

What is already known about this topic

- Pre-training on key concepts or terms improves learning, but little is known whether and how pre-training on strategy acquisition supports learning.
- Mayer's multimedia principle suggests people learn better from illustrated text than from text alone; however, learners sometimes fail to integrate text and picture.

What this paper adds

- Pre-training on text-picture processing strategy is effective.
- Verbal and visual training foster text-picture processing strategy acquisition.
- Verbal training improves both recall and transfer test performance, and visual training improves only recall test performance.
- Verbal training is better in improving outcomes.
- Fixation time on pictures mediates the effects of training on learning outcomes.

Implications for practice and/or policy

- Pre-training should be used to support learners' strategy acquisition.
- This study also provides suggestions on how to design pre-training on strategy acquisition.

# INTRODUCTION

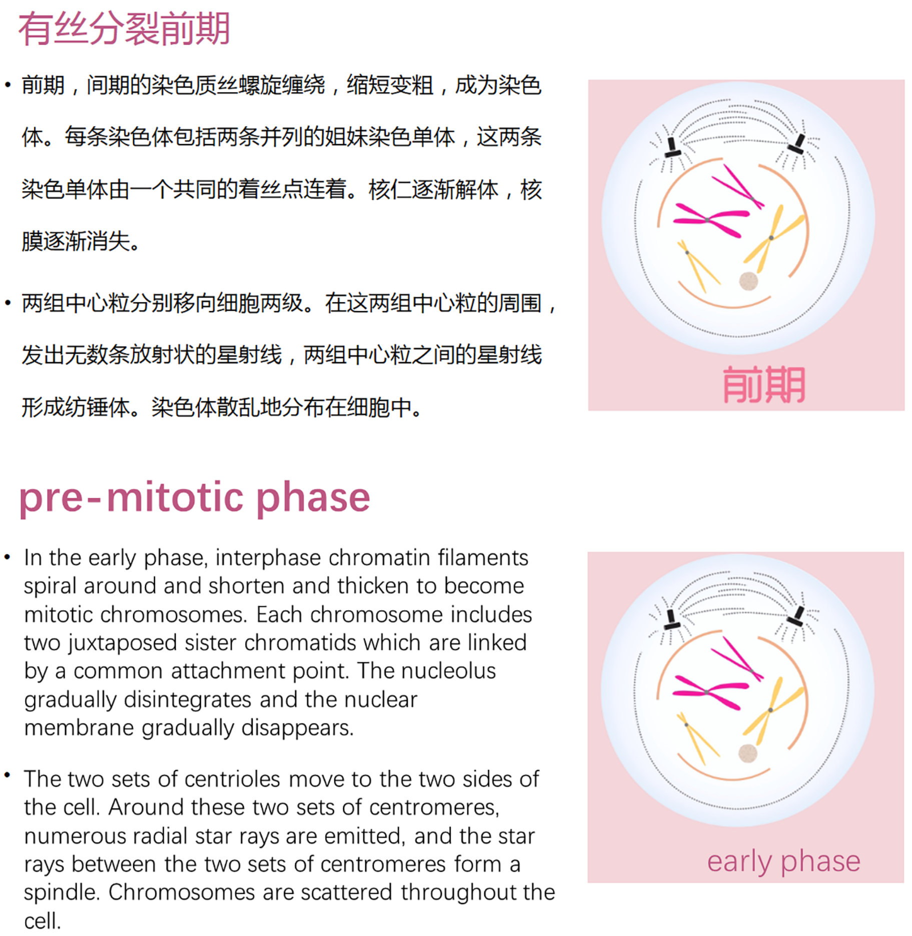
## Objective and rationale

Suppose a student receives a computer-based multimedia lesson on a scientific topic that consists of a series of slides containing printed words and corresponding illustrations, such as exemplified in Figure 1. According to current theories of multimedia learning, a major cognitive challenge for the student is to engage in the process of text-picture processing [Ainsworth's (2022) DeFT framework; Mayer's (2021, 2022a) Cognitive Theory of Multimedia Learning; Schnotz's (2022, 2023) Integrated Model of Text and Picture Comprehension]. Text-picture processing includes selection of relevant text and pictorial elements, organization of text and pictures and integration of text and pictorial information. In text-picture processing, the learner attends to corresponding aspects of the text and picture, and mentally connects them in working memory. For example, in examining the slide in Figure 1, a step in text-picture integration occurs when the learner attends to and mentally connects the printed words ‘mitotic chromosomes’ and the part of the illustration showing mitotic chromosomes. Engaging effectively in the process of text-picture processing is the hallmark of meaningful learning (Mayer, 2021, 2022a; Schnotz, 2022, 2023). Therefore, it is worthwhile to train students in effective strategies for text-picture processing.

The present study is intended to contribute to this search for ways to train students to engage in effective text-picture processing. The current study attempts to investigate how visual or verbal training methods prompt strategy use in computer-based multimedia learning, and therefore improve learning with educational technology.

## The case for teaching of text-picture processing strategies

Research shows that people can learn better from printed text and illustrations than from printed text alone, which is an example of the *multimedia principle* (Mayer, 2021, 2022b).



有丝分裂前期

- 前期，间期的染色质丝螺旋缠绕，缩短变粗，成为染色体。每条染色体包括两条并列的姐妹染色单体，这两条染色单体由一个共同的着丝点连着。核仁逐渐解体，核膜逐渐消失。
- 两组中心粒分别移向细胞两级。在这两组中心粒的周围，发出无数条放射状的星射线，两组中心粒之间的星射线形成纺锤体。染色体散乱地分布在细胞中。

前期

pre-mitotic phase

- In the early phase, interphase chromatin filaments spiral around and shorten and thicken to become mitotic chromosomes. Each chromosome includes two juxtaposed sister chromatids which are linked by a common attachment point. The nucleolus gradually disintegrates and the nuclear membrane gradually disappears.
- The two sets of centrioles move to the two sides of the cell. Around these two sets of centromeres, numerous radial star rays are emitted, and the star rays between the two sets of centromeres form a spindle. Chromosomes are scattered throughout the cell.

early phase

FIGURE 1 One example from the lessons on the Mitosis and Meiosis (top: Chinese version; bottom: English translation).

However, illustrated text is not always advantageous, because the advantage of illustrated text depends on how well learners can successfully integrate corresponding text and pictures. Previous research has shown that students were often unable to effectively integrate text and pictures (Bodemer et al., 2005; Butcher, 2006; Hegarty & Just, 1993; Mayer, 2022b; Paas et al., 2010). For example, Hannus and Hÿonä (1999) found that learners usually ignored the illustrations. Mason et al. (2013) found students who had relatively lower prior knowledge were less likely to integrate text and pictures successfully.

To foster the integration of text and pictures, previous research has mainly focused on the optimization of instructional design. For instance, there is substantial evidence for the *spatial contiguity principle*, which states that the adjacent presentation of corresponding printed text and pictures on the screen is more conducive to learning than distant presentation of text and corresponding graphics on the screen (Ayres & Sweller, 2022; Mayer, 2021). Nevertheless, under real-world circumstances, learners cannot always encounter well-designed and principle-based materials. For example, in Figure 1, the text is presented on one side of the screen and the illustration is presented on the other side of screen. When illustrated text materials are not optimally designed, learners need to rely on their own skill in active text-picture processing. Thus, it was essential for learners to learn how to use text-picture processing strategies for multimedia instructional materials.

One approach is to provide students with pre-training in how to engage in text-picture processing. According to the *pre-training principle*, people learn more deeply when they know key concepts before being exposed to the lesson (Mayer, 2021; Mayer & Fiorella, 2022). In addition to the key concepts and terms, we can also apply the pre-training principle to train learners to master some learning strategies, such as how to integrate the text and pictures in an online lesson.

Verbal and visual training have been used to foster the acquisition of text-picture processing strategies (eg, Kombartzky et al., 2010; Krebs et al., 2019; Mason et al., 2015, 2016; Schlag & Ploetzner, 2011). In verbal training, the learner receives prompts or instruction in the text format that explains how to complete a task and directly tells what to do. In visual training, or eye movement modelling examples (EMMEs; Jarodzka et al., 2013), the learner views a video in which recordings of an expert's eye fixations when learning from an on-screen lesson are superimposed on the original learning material.

Although some studies have investigated the effects of visual or verbal training on strategy acquisition (eg, Kombartzky et al., 2010; Krebs et al., 2019; Mason et al., 2015, 2016, 2017; Schlag & Ploetzner, 2011), less is known about whether these two forms on text-picture processing strategy training differ in their effectiveness. Therefore, in the present study, we examined whether visual and verbal training effectively fostered the acquisition of text-picture processing strategies and enhanced learning outcomes, as compared to a control group. We also compared the effects of these two training methods on strategy use during learning and on learning outcomes.

## Theoretical framework for integration of text and pictures

Both the cognitive theory of multimedia learning (CTML; Mayer, 2021, 2022a) and cognitive load theory (CLT; Sweller et al., 1998, 2019) provide a theoretical framework for effective cognitive processes in multimedia learning involving the integration of text and pictures.

The CTML states three assumptions: dual channels, limited capacity, and active processing (Mayer, 2021, 2022a). The dual channels assumption is that humans process visual information (eg, on-screen text, pictures, and video) and verbal/auditory information (eg, narration) in separate channels. The limited capacity assumption is that working memory capacity is limited so that humans can process only limited amounts of information in a channel at one time. The active processing assumption states that to construct a meaningful mental model, learners need to actively get involved in appropriate cognitive processing, that is, they need to select relevant text and pictorial information in the presented material, organize the text and pictorial information in the working memory into coherent mental representations, and finally, integrate text and picture representations into a mental model along with prior knowledge activated from long-term memory.

According to CTML, people may fail to form a coherent mental model if they are unable to select relevant text and pictorial information and integrate text information with corresponding pictorial information. For example, when studying illustrated text, students often ignore the illustrations, making them fail to integrate text and picture information and form a meaningful mental model (Hannus & Hÿonä, 1999; Renkl & Scheiter, 2017). Therefore, it is important to train learners on how to use appropriate strategies to manage their cognitive processing in multimedia learning, particularly how to select and integrate corresponding words and pictures.

The pre-training principle is a solution to guiding appropriate cognitive processing in multimedia learning, which states that when learners know key concepts, they will learn from multimedia materials more deeply (Mayer & Fiorella, 2022; Mayer & Pilegard, 2014; Meyer et al., 2019). Several reviews have found a positive effect of the pre-training principle ($d=0.46$, Mayer, 2017; $d=0.75$, Mayer & Pilegard, 2014). The present study is based on the idea that when learners are pre-trained on effective cognitive processing strategies, they put more attention on relevant elements, appropriately manage working memory to select necessary text and pictorial information and then integrate corresponding text and pictures, therefore supporting deeper cognitive processing.

CLT suggests there are three types of cognitive load that influence the cognitive processing in multimedia learning: intrinsic, extraneous and germane cognitive load (Sweller et al., 1998, 2019). Mayer (2021, 2022a) proposed three kinds of demands on the learner's information processing system: extraneous processing, essential processing and generative processing, which are analogous to extraneous, intrinsic and germane cognitive load, respectively. Extraneous processing, which is caused by poor instructional design, is cognitive processing that is irrelevant to instructional goal. Essential processing, which depends on the complexity of the learning materials, is cognitive processing required by meaningful mental construction of the essential material. Generative processing is exerting effort to make sense of the learning content. People with low ability spend much time on irrelevant information, for example, blank spaces between and around text and illustrations (Hannus & Hÿonä, 1999). Directing people's attention to relevant elements by using cues is a common solution (eg, Arslan-Ari, 2018).

Rather than directly guiding attention, learning a strategy equips people with knowledge about how to select relevant elements, organize and integrate text and pictures. Using an appropriate text-picture processing strategy can help to reduce attention to irrelevant information, thereby allowing for learners to reduce extraneous processing and allocate their cognitive resources to manage essential processing and foster generative processing—the foundations of meaningful learning. In this case, generative processing will increase since people are more involved in meaningful learning. In the present study, we investigated training tools intended to foster efficient text-picture processing, which could reduce extraneous processing, guide essential processing and increase generative processing.

## Learning text-picture processing strategy by visual training

According to the social-cognitive learning theory (Bandura, 1986), to a large extent, cognitive skills are learned from observation of other people who exemplify what is to be learned. Eye movement modelling examples (EMMEs) are a kind of visual strategy training for processing onscreen material (Emhardt et al., 2023; Van Gog & Rummel, 2010). EMMEs take the form of a video showing the eye fixations of an expert learner, for example, rendered as red dots that grow in size as duration increases. For example, in the present study, a sequence of red dots growing in size was superimposed on the instructional materials.

Eye movements reflect one's strategy for cognitive processing (eg, Jarodzka, 2022; Levin et al., 2021; Scholz et al., 2015). Watching others' eye gaze such as in EMMEs helped regulate learners' processing behaviour and thereby helped them to acquire new processing strategies (Brennan et al., 2008; Litchfield et al., 2010). Studies in sports and medical science have demonstrated that experts' gaze was able to train novices to use specific strategies, and thus improve their task performance (eg, Vine et al., 2011; Wilson et al., 2011). Visual training can provide direct access to an expert's strategies for cognitive processing. Learners can be shown the relevant parts the expert looked at, how to shift gaze and the duration of time focused on any part. This can help learners gain insights into the expert's covert cognitive process, and help them acquire corresponding strategies (Frischen et al., 2007; Krebs et al., 2019).

Previous research demonstrated that visual training was an effective training tool to foster the use of text-picture processing strategies (eg, Emhardt et al., 2023; Krebs et al., 2019, 2021; Mason et al., 2015, 2016, 2017; Scheiter et al., 2018). As a training tool, visual training promoted integrative processing behaviour (ie, more fixation time on pictures and more frequent transitions between text and pictures) and improved learning outcomes. In a series of research studies by Mason et al. (2015, 2016, 2017), 7th-grade students benefited from watching visual training. For example, in the research of Mason et al. (2015), an EMME, in which the expert used a text-picture processing reading strategy, was presented to a group of students when learning material about the water cycle; then students learned about new material about the food chain on their own without EMME. Results showed that students who watched the EMME shifted their eye-gaze between text and pictures more often and performed better on both recall and transfer tests than those who did not. Krebs et al. (2019, 2021) also found that university students in the visual training condition gazed at pictures more and had better learning performance than those in the control condition.

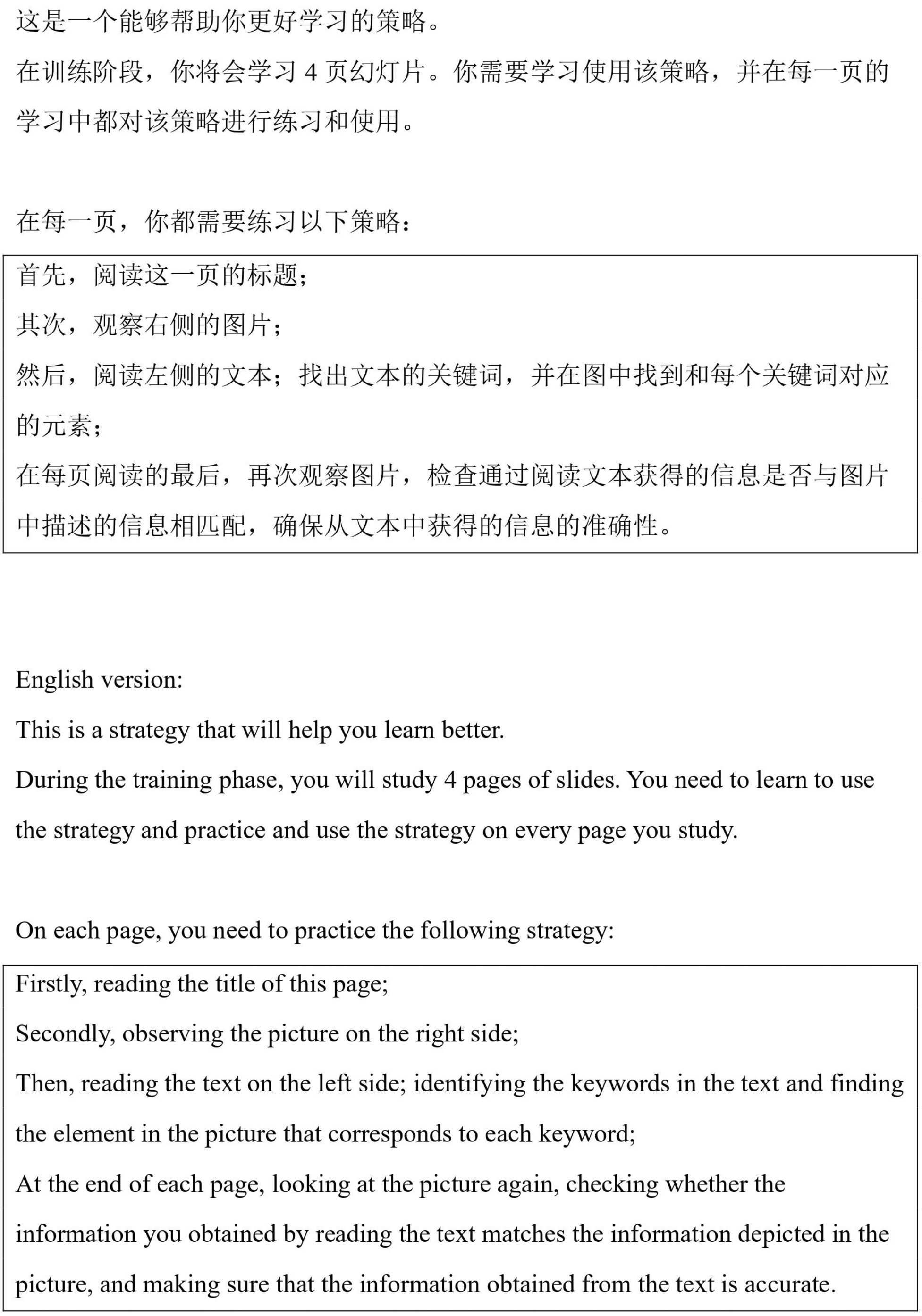
## Learning text-picture processing strategy by verbal training

Verbal training, namely verbal prompts or instructions, provides a kind of explicit description of text-picture processing strategies, including specific procedures that directly tell learners what to do. For instance, verbal training in the text-picture processing strategy describes how learners need to select portions of printed text, find corresponding parts of pictures and connect them together. For example, Figure 2 shows the verbal training strategy we used in this study.

Previous studies have investigated whether verbal training was beneficial in motor learning (eg, Meier et al., 2020; Van Duijn et al., 2019), self-regulated learning (eg, Azevedo & Cromley, 2004) and strategy use (eg, Jian, 2018, 2021; Kombartzky et al., 2010; Larson et al., 1986; Schlag & Ploetzner, 2011; Seufert, 2019). Some studies on the effectiveness of verbal training on text-picture processing showed a positive effect on learning performance (eg, Jian, 2018, 2021; Kombartzky et al., 2010; Larson et al., 1986; Stalbovs et al., 2015). For instance, Kombartzky et al. (2010) found that learners who received verbal strategy instruction learned significantly more than those who did not. Similarly, Schlag and Ploetzner (2011) found that learners who received verbal strategy instruction before learning achieved higher scores on post-learning tests. Jian (2018, 2021) also found verbal training on a three-step text-picture processing strategy improved young children's reading comprehension performance. Stalbovs et al. (2015) used a strategy training method based on specific verbal rules to support text-picture processing for multimedia instruction, and found this kind of strategy training improved learning. In the present study, learners received verbal training and practiced on how to use the text-picture processing strategy during the training phase.

## Visual training versus verbal training

Visual training provides bottom-up control and trains leaners by guiding attention to specific cognitive processing, whereas verbal training provides top-down control and trains learners by teaching them cognitive processing and asking them actively engage in them. Verbal training provides an explicit form of knowledge, which contains more specific, detailed information and is easy to understand and follow. After receiving verbal training, learners can transform provided declarative knowledge about text-picture processing strategy into procedural knowledge through three stages: the cognitive (interpreting information or procedures), associative (converting declarative to procedural knowledge) and autonomous



这是一个能够帮助你更好学习的策略。

在训练阶段，你将会学习 4 页幻灯片。你需要学习使用该策略，并在每一页的学习中都对该策略进行练习和使用。

在每一页，你都需要练习以下策略：

首先，阅读这一页的标题；

其次，观察右侧的图片；

然后，阅读左侧的文本；找出文本的关键词，并在图中找到和每个关键词对应的元素；

在每页阅读的最后，再次观察图片，检查通过阅读文本获得的信息是否与图片中描述的信息相匹配，确保从文本中获得的信息的准确性。

English version:

This is a strategy that will help you learn better.

During the training phase, you will study 4 pages of slides. You need to learn to use the strategy and practice and use the strategy on every page you study.

On each page, you need to practice the following strategy:

Firstly, reading the title of this page;

Secondly, observing the picture on the right side;

Then, reading the text on the left side; identifying the keywords in the text and finding the element in the picture that corresponds to each keyword;

At the end of each page, looking at the picture again, checking whether the information you obtained by reading the text matches the information depicted in the picture, and making sure that the information obtained from the text is accurate.

FIGURE 2 Verbal instruction presented in the verbal training group.

stages (after enough practice, skill using becomes automatic) (Fitts & Posner, 1967). Once becoming automatic, strategies will be easily accessed and activated.

By contrast, visual training involves a direct demonstration of cognitive strategy use. Eye movements contain important information on what to look and when to look at it. However, knowing when to look at the right element does not mean that learners are able to actively implement it. Visual training that only includes eye movements may be difficult for learners to understand (Van Gog et al., 2009); hence, they may not know why they should shift their gaze.

Considering the lack of previous study directly comparing visual and verbal training on text-picture processing strategy acquisition, it is worthwhile to compare the effectiveness of the two methods in promoting the acquisition of new strategies and learning outcomes. The present study addresses this need.

The two training methods use the same text-picture processing strategy based on Krebs et al. (2019). For each page, firstly, students will read the title. This is because previous research has found that this kind of global text processing strategy (ie, reading the title) improved learners' recall of the content and the summarizing skills of the content (Hyönä et al., 2002; Lorch et al., 2013; Sanchez et al., 2001). Secondly, students observe the picture on the right side, so that they can obtain the pictorial representation more coherently and comprehensively (Eitel et al., 2013). Additionally, observing pictures first can foster picture processing and text-picture integration behaviour (Mason et al., 2017). Thirdly, students read the text, identify significant keywords and find corresponding elements in the pictures. This process, which has been proven to be effective, includes selection of relevant text and pictorial information, text organization, picture organization and integration of text and pictures (Krebs et al., 2019; Mason et al., 2015, 2016). At the end of each page, students observe the picture again, check whether the information obtained by reading the text matches the information depicted in the picture and make sure that the information obtained from the text is accurate (Hegarty & Just, 1993; Krebs et al., 2019).

Specifically, in the visual training group, an expert used this strategy to read materials, and the eye movement recording was superimposed on the learning materials, which can provide detailed insight into cognitive processing when using this strategy. People are be able to follow eye movements of an expert, adjust and regulate their own behaviour and finally acquire the new strategy. In the verbal training group, students are provided instruction that describes this strategy. People interpret the instruction first, convert declarative knowledge to procedural knowledge by using this strategy when reading materials, adjust the strategy use to suit their pace and finally acquire this new strategy.

Eye-tracking technology is now widely used in multimedia learning research (Alemdag & Cagiltay, 2018; Cagiltay & Coskun, 2022; Jarodzka, 2022; Mayer, 2010; Van Gog & Scheiter, 2010). It has been used to reveal where people look during the integration of text and pictures and has provided opportunities to investigate internal cognitive processes (Hyönä, 2010; Scheiter & Van Gog, 2009; Tabbers et al., 2008). In studies with EMME, researchers generally used eye-tracking technology to explore learners' underlying cognitive processes during learning (Emhardt et al., 2023). Whereas, in the research of verbal training, to our knowledge, only Jian (2018, 2019, 2021) used an eye tracker to investigate young children's integrative processing behaviour. Therefore, in the present study, using an eye tracker, we investigated whether using visual or verbal training could help college students acquire the text-picture processing reading strategy by focusing more on pictures and producing more integrative processing behaviour, thereby leading to improved learning outcomes.

## Hypotheses for the current study

The current study examined whether verbal and visual training methods improved performance compared to the control group on learning outcomes (as measured by recall and transfer tests) and visual attention (as measured by text fixations, picture fixations and transitions). Based on theories of multimedia learning, we expected that both visual and verbal training would be effective in improving learning outcomes and directing visual attention. Thus, we made the following hypotheses.

> **Hypothesis 1.** The first hypothesis is that verbal training in text-picture processing improves learning outcomes. Thus, the verbal training group will score higher than the control group on recall score (hypothesis 1a) and transfer score (hypothesis 1b).

> **Hypothesis 2.** The second hypothesis is that visual training in text-picture processing improves learning outcomes. Thus, the visual training group will score higher than the control group on recall score (hypothesis 2a) and transfer score (hypothesis 2b).

> **Hypothesis 3.** The third hypothesis is that verbal training directs visual attention. Specifically, the verbal training group will display less fixation time on text (hypothesis 3a), more fixation time on pictures (hypothesis 3b) and more transitions between text and pictures (hypothesis 3c) than the control group.

> **Hypothesis 4.** The fourth hypothesis is that visual training directs visual attention. Specifically, the visual training group will display less fixation time on text (hypothesis 4a), more fixation time and pictures (hypothesis 4b) and more transitions between text and pictures (hypothesis 4c) than the control group.

We also explored two research questions:

> Research question 1: Is verbal training of text-picture processing more effective than visual training, in improving learning outcomes and guiding visual attention? To answer this question, we compared the verbal training group and the visual training group on recall test score, transfer test score, fixation time on text, fixation time on pictures and number of transitions between text and pictures.

> Research question 2: What mediates the effects of verbal and visual training on learning outcomes? To answer this question, we conducted mediational analysis.

# METHOD

## Participants and design

In the current study, we recruited 69 participants from a university in China. A prior analysis based on G*Power 3.1 (Faul et al., 2007) showed that, to conduct ANCOVAs with 3 covariates, 64 participants were required when using a large effect size $f = 0.40$ (Jian, 2018; Wang et al., 2020), $\alpha$ was set to 0.05 and statistical power 1-$\beta$ was set to 0.80. The experiment used a one-factor three-level between-subjects design. Participants were randomly assigned to three groups: 24 students served in the verbal training group, 23 students in the visual training group and 22 students in the control group. In order to exclude participants with high prior knowledge, we only included participants whose pretest score was 60% or less (ie, 0–6 score). All participants had normal or corrected-to-normal vision. The experiment was approved by the university's ethics committee, and signed informed consent was obtained for all participants.

A total of 66 participants were included in the analysis of learning outcomes (3 participants were excluded because their posttests were not finished), leaving 22 students in each group. There was no significant difference among the three groups in the proportion of men and women, $\chi^2(2)=4.27$, $p>0.05$. There was no significant difference among the groups in pretest score, $F(2, 63)=0.68$, $p>0.05$. However, there was a significant difference in age, $F(2, 63)=3.35$, $p=0.041$, so age was used as a covariate in subsequent data analyses on learning outcome.

A total of 67 participants were included in the analysis of eye-tracking data (2 participants were excluded because of poor eye-tracking data quality, ie, sampling rate less than 70%), leaving 24 students in the verbal training group, 23 students in the visual training group and 20 students in the control group. There was no significant difference among the three groups on the proportion of men and women, $\chi^2(2)=4.43$, $p>0.05$. There was no significant difference among the groups in pretest score, $F(2, 64)=0.97$, $p>0.05$. However, there was a significant difference in age, $F(2, 64)=3.61$, $p=0.033$. Thus, age was also used as a covariate in the subsequent eye movement data analyses.

A total of 64 participants were included in the mediation analyses. There were 22 participants in the verbal training group, 22 participants in the visual training group and 20 participants in the control group.

## Instructional materials

There were three versions of instructional materials: control, verbal training and visual training. For the control group, the learning materials were slides explaining the steps of mitosis and meiosis. The content of instructional materials was compiled from several books related to cell biology, and experts in related fields were invited to review the accuracy of the content. The experimental materials were presented in the form of printed text and pictures (as shown in Figure 1), including a total of 16 slides. The content can be divided into 2 parts: basic concepts (4 slides) and specific steps of mitosis and meiosis (12 slides).

For the verbal training group, the learning materials were the same as in the control group. Before the lesson, students were given verbal instructions for text-picture processing as shown in Figure 2. In the first part of verbal training, the first four slides were used to practice strategy use. Students would read the verbal instructions for how to integrate printed words and pictures, practice using the strategy to read multimedia materials and become familiar with the strategy; then in the second part, no verbal instructions were provided, and the learner was asked to use the strategy they learned to read the learning material (ie, the last 12 slides).

For the visual training group, students received a video showing EMME for the first 4 slides and then read the following 12 slides without EMME. Specifically, to demonstrate the text-picture processing strategy in line with a previous study (Krebs et al., 2019), the first part of the learning materials (ie, 4 slides on basic concepts) was a video showing the eye-tracking recordings of an expert reading the first four slides within the text-picture processing strategy. The steps of text-picture processing reading strategies used by the expert were based on the strategy instructions shown in Figure 2 (Krebs et al., 2019). Notably, the strategy in the visual training group was exactly the same as that in the verbal training group. The visual training video adopted the eye movements of an expert (a female research assistant) who was very familiar with the material. Before recording, the expert learned the specific steps of text-picture processing strategy as shown in Figure 2. She was asked to use this strategy when watching the lesson and the eye-tracking recording would be presented to other learners. The visual training video showed fixations with a duration of more than 100 ms, which were visualized as translucent red dots. The total duration of the EMME recoding involving the first 4 slides was 206 seconds. The remaining 12 slides were presented the same as for the control group. The assessment covered the material presented in these 12 slides, which was the same for all groups.

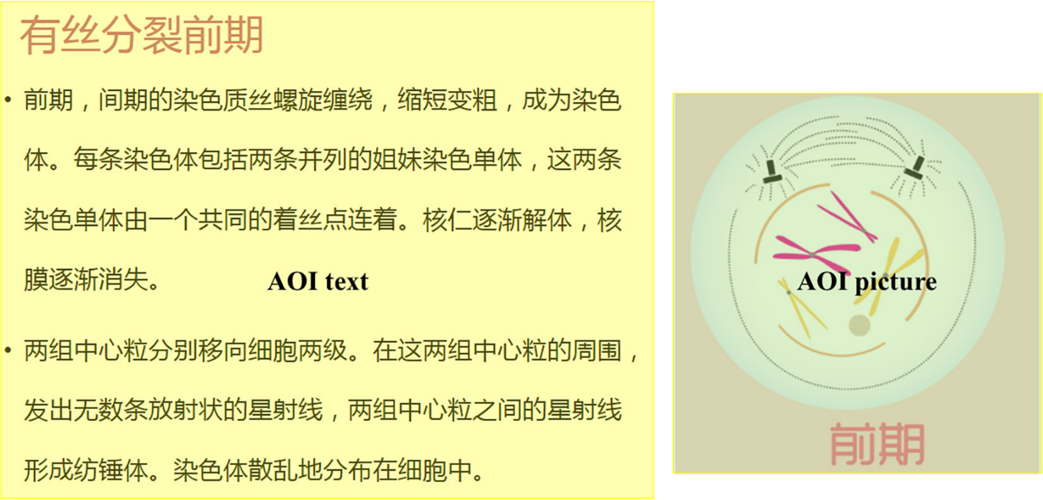
## Assessment instruments

The assessment instruments consisted of the pretest and posttest. The pretest consisted of two parts. The first part collected students' demographic information such as gender and age. The second part was used to measure students' prior knowledge about mitosis and meiosis, which included 10 multiple-choice questions with 4 choices and one correct answer per question ($\omega = 0.50$). Each correct answer was awarded one point. The internal consistency was not high. The reason was that the prior knowledge test included several subtopics and details that were not connected to each other, rather than similar questions.

The posttest included recall and transfer tests, which only related to the last 12 slides learned in the learning phase. The recall test consisted of two questions that require students to write as much as possible about the specific steps of mitosis and meiosis. Participants received one point for expressing each of 30 information units (19 points for mitosis and 11 points for meiosis), regardless of specific wording. The transfer test consisted of 7 multiple-choice questions (eg, 'One cell of corn has 10 pairs of chromosomes. After two successive mitotic divisions, how many chromosomes are in the daughter cells?') and 2 open-ended questions (eg, 'Some people may be born with a disease when something wrong happened in the process of sperm formation. The patient who has this disease has 3 chromosomes, rather than one normal pair. Please try to use the knowledge about meiosis to analyse what happened in the process of sperm formation'), with a maximum of 19 points. Each multiple-choice question had 4 choices and one correct choice, and one point was awarded for each correct answer. The two open-ended questions required students to answer novel questions about mitosis and meiosis based on what they had learned. Students received one point for each acceptable statement, with a maximum of 12 points for two open-ended questions. Two well-trained raters scored the posttest independently. The average score of the two scorers was taken as the final score. Inter-rater reliabilities between two raters for the recall and transfer test scores were 0.99 and 0.98 ($ps < 0.001$), respectively. For the reliability of posttests, we calculated McDonald's omega for the recall test ($\omega = 0.67$) and transfer test ($\omega = 0.79$).

## Gaze measures

In order to examine whether verbal training and visual training would guide students' attention to text or pictures and foster text-picture integration processing behaviour, this study analysed students' gaze on each slide of the second part of instructional material (ie, the last 12 slides). We defined the Area of Interests (AOIs) for each slide as exemplified in Figure 3. Specifically, in each slide, we defined a picture AOI and a text AOI. Previous studies have shown that total fixation time on text reflected processing of text, total fixation time on pictures reflected processing of pictures and the number of text-picture transitions (ie, the number of times the learners gaze when from text to picture or from picture to text) reflected students' attempts to integrate text-picture information (Johnson & Mayer, 2012; Krebs et al., 2019). Thus, three eye movement indices were chosen: total fixation time on the text, total fixation time on the picture and the number of transitions between text and picture, all of which are based on the overall values across all the 12 slides.



有丝分裂前期

- 前期，间期的染色质丝螺旋缠绕，缩短变粗，成为染色体。每条染色体包括两条并列的姐妹染色单体，这两条染色单体由一个共同的着丝点连着。核仁逐渐解体，核膜逐渐消失。
- 两组中心粒分别移向细胞两级。在这两组中心粒的周围，发出无数条放射状的星射线，两组中心粒之间的星射线形成纺锤体。染色体散乱地分布在细胞中。

FIGURE 3 One example of Areas of Interests (AOIs).

## Apparatus

The eye movements of students were collected by the SMI RED 250 desktop eye tracker (SensoMotoric Instruments, Germany) with a sampling rate of 250 Hz and a spatial resolution of fewer than 0.1 degrees. The resolution of the monitor presenting learning material was 1680 × 1050 pixels, and participants were seated approximately 65 cm away from the monitor.

## Procedure

Participants were tested individually and randomly assigned to one of the three groups. First, participants completed the pretest. Second, participants sat in front of the eye tracker with their chins on the headrest, and they were required to keep their heads still in the experiment as much as possible. Participants' eye fixations were calibrated using a 9-point calibration procedure.

Next, students received pre-training based on their assigned group. Participants in each of the two training groups were told that, in the training phase, they should learn a strategy and they were required to use this strategy to read each page. Students in the verbal training group were provided with paper-based instructions for the text-picture processing strategy (as shown in Figure 2). Then they watched the learning content based on instruction and could read the instruction anytime during the training phase. Students in the visual training group were told that the red dots in the learning materials represent the eye movements of an expert, and the dot increased in size to reflect increases in the expert's fixation duration, and then, they watched a video showing the learning content superimposed by expert's eye movements for the first 4 slides. Students in the control group watched the learning content without any instruction. After the training phase, the experimenter asked each participant whether he/she knew how to use this strategy. Every participant was confirmed to know the text-picture processing strategy. Then, to keep eye tracking quality, participants' eye fixations were recalibrated using a 9-point calibration procedure. In the learning phase, participants were told to use strategies they had trained to learn. There were no time limits in the learning phase, and participants could press to turn to the next slide but not to turn back to the last slide. In the end, all students completed the posttest at their own pace. The whole experiment lasted about 40 minutes and participants were paid for their participation.

# RESULTS

To compare differences among the three groups, one-way ANCOVAs were performed for each variable with prior knowledge, learning time and age as covariates. In all ANCOVAs, all assumptions were met (skewness and kurtosis in each group; see Table S1 in supplementary material). To adjust post-hoc tests, Benjamini and Hochberg's (1995) method was used to correct the *p* values for multiple comparisons by using the *stats* package in R software (R Core Team, 2021) since this method was more powerful than the Bonferroni procedure (Glickman et al., 2014). In addition, we also provided results from linear mixed models (see Supplementary material A). Also, mediation models were conducted to test whether the impact of conditions on performance was mediated by participants' visual processing by PROCESS (Model 4 with a 95% confidence interval and 5000 bootstrap samples; Hayes, 2013).

## Preliminary analyses

One-way ANOVAs were conducted to detect any differences in participants' prior knowledge, learning time and age. Concerning 66 participants in the comparison of learning outcomes, there was no significant difference among three groups on prior knowledge, $F(2, 63)=0.68$, $p>0.05$, and learning time, $F(2, 63)=1.40$, $p>0.05$, and there was a significant difference in age, $F(2, 63)=3.35$, $p=0.041$. Concerning 67 participants in the comparison of eye tracking data, there was no significant difference among three groups on prior knowledge, $F(2, 64)=0.97$, $p>0.05$, and learning time, $F(2, 64)=1.58$, $p>0.05$, and there was a significant difference in age, $F(2, 64)=3.61$, $p=0.033$.

Learning time was significantly correlated with text fixation time ($r=0.75$, $p<0.001$), picture fixation time ($r=0.43$, $p<0.001$), and transitions between text and pictures ($r=0.55$, $p<0.001$). Therefore, we included learning time as a covariate to control its potential influence. Although there were no differences among groups, we still involved prior knowledge as a covariate. Specifically, because prior knowledge can moderate the effects of EMME (Krebs et al., 2019, 2021; Scheiter et al., 2018), we attempted to control its potential influence. Since participants' age were significantly different among three groups, we also included age as a covariate.

## Learning outcomes

The top row of Table 1 shows the mean recall scores and standard deviations for the three groups. An ANCOVA indicated significant differences among the three groups on the recall test, $F(2, 60)=4.63$, $p=0.013$, $\eta^2_p=0.13$. Consistent with hypothesis 1a, a post-hoc test showed that the recall score of the verbal training group ($p=0.021$) was better than the control group. Consistent with hypothesis 2a, a post-hoc test showed that the visual training group scored significantly higher than the control group on the recall test ($p=0.032$).

The second row of Table 1 shows the mean transfer scores and standard deviations for the three groups. An ANCOVA indicated significant differences among the groups on the transfer test, $F(2, 60)=3.17$, $p=0.049$, $\eta^2_p=0.10$. Consistent with hypothesis 1b, post-hoc

TABLE 1 Means and standard deviations of variables in each group.

| | Visual training | | Verbal training | | Control | |
|---|---|---|---|---|---|---|
| Variables | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Learning outcomes | | | | | | |
| Recall | 14.84 | 5.32 | 15.34 | 4.90 | 11.11 | 5.09 |
| Transfer | 9.00 | 2.51 | 10.77 | 3.37 | 8.20 | 3.27 |
| Eye movements | | | | | | |
| Fixation time on text (s) | 285.44 | 78.79 | 331.19 | 92.54 | 390.38 | 118.79 |
| Fixation time on pictures (s) | 159.41 | 58.88 | 182.92 | 84.34 | 122.88 | 28.69 |
| Transitions between text and pictures | 165.13 | 54.28 | 197.04 | 51.20 | 190.05 | 57.20 |

results showed the verbal training group outperformed the control group ($p = 0.048$). There was no significant difference between the visual training group and the control group on transfer test scores ($p = 0.403$), which did not support hypothesis 2b.

We conclude that verbal training improved students' recall and transfer test performance, in line with hypotheses 1a and 1b, and visual training improved students' recall but not their transfer test performance, in line with hypothesis 2a but not 2b.

## Eye movements

The third line in Table 1 reports the mean fixation time on text and the standard deviation for the three groups. An ANCOVA showed there was a significant difference in the fixation time on text, $F(2, 61) = 5.56$, $p = 0.006$, $\eta^2_p = 0.15$. In line with hypothesis 3a, a post-hoc test showed that the verbal training group ($p = 0.044$) had a shorter fixation time on text than the control group. Results showed that the visual training group had significantly lower fixation time on the text than the control group ($p = 0.006$) as per hypothesis 4a.

The fourth line in Table 1 shows the mean fixation time on pictures and standard deviation for the three groups. An ANCOVA showed that the three groups were significantly different on this measure, $F(2, 61) = 7.47$, $p = 0.001$, $\eta^2_p = 0.20$. In line with hypothesis 3b, a post-hoc test showed that the verbal training group spent more time looking at the pictures than the control group ($p = 0.006$). In line with hypothesis 4b, there was significantly longer fixation time on the pictures for the visual training group than the control group ($p = 0.003$).

The fifth line in Table 1 shows the mean number of transitions between text and pictures. An ANCOVA showed there was no significant difference among the three groups in the number of text-picture gaze switches, $F(2, 61) = 0.46$, $p = 0.637$. In contrast to hypothesis 3c, the verbal training group did not exhibit more text-picture gaze switches than the control group. In contrast to hypothesis 4c, the visual training group and control group did not differ significantly on transitions between text and pictures.

Overall, we conclude that there is partial support for hypothesis 3, which states that the verbal training group will allocate visual attention during learning differently than the control group. In particular, the verbal training group shifted visual attention by spending less time on the text (as per hypothesis 3a) and more time on the pictures (as per hypothesis 3b). We conclude that there is partial support for hypothesis 4. Participants in the visual training group significantly fixate less on text (as per hypothesis 4a) and fixate more on pictures (as per hypothesis 4b) than those in the control group. Contrary to hypothesis 3c and 4c, no significant differences were found on the transitions between text and pictures.

## Research question 1: Do the verbal and visual training groups differ on measures of learning outcome or visual attention processing?

In follow-up post-hoc tests to the ANCOVAs reported in foregoing sections, there were no significant differences between the verbal and visual training groups on recall test score ($p=0.661$), transfer test score ($p=0.167$), fixation time on text ($p=0.324$), fixation time on pictures ($p=0.560$) or text-picture transitions. We conclude that the two training methods did not differ in their effects.

## Research question 2: What mediates the effects of training on learning outcomes?

To test whether visual processing mediated the effect of conditions on learning outcomes, separate mediation analyses were performed comparing groups that had significant differences on learning outcomes. We used effect coding for the three experimental conditions (verbal training [1] vs. control [−1], and visual training [1] vs. control [−1]), and we used prior knowledge and age as covariates. Fixation time on pictures was used as the mediator, and z-standardized recall and transfer scores were treated as dependent variables. We did two separate mediation analyses to examine the mediation effects of group conditions on recall performance and the effect of the verbal training group versus the control group on transfer performance.

Results showed there were significant indirect effects of verbal training ($\beta=0.14$, $SE=0.08$, 95% CI=[0.02, 0.33]) and visual training ($\beta=0.19$, $SE=0.11$, 95% CI=[0.03, 0.45]) on recall performance mediated by fixation time on pictures, indicating that the higher recall scores could be explained by longer fixation time on pictures (see Figure 4). There was no significant indirect effect of verbal training on transfer performance mediated by fixation time on pictures ($\beta=0.23$, $SE=0.25$, 95% CI=[−0.43, 0.58]).

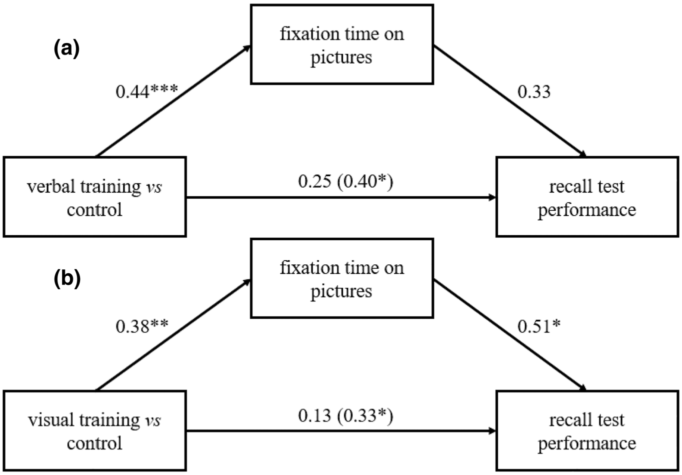

(a)
fixation time on pictures
0.44***
0.33
verbal training vs control
0.25 (0.40*)
recall test performance
(b)
fixation time on pictures
0.38**
0.51*
visual training vs control
0.13 (0.33*)
recall test performance

FIGURE 4 Mediation Model of group conditions on recall test performance (a: verbal training vs. control; b: visual training vs. control). ***$p<0.001$, ** $p<0.01$, *$p<0.05$.

# DISCUSSION

## Empirical contributions

This study demonstrated the effectiveness of verbal training and visual training methods for how to process multimedia instructional material. Both verbal and visual training methods improved learning outcomes and learning processes. In particular, training in how to process multimedia instructional material resulted in better recall test scores, better transfer test scores (for verbal training only), less visual attention for on-screen text and more visual attention for on-screen pictures. Surprisingly, we did not find improvements in text-picture transitions. There were no significant differences between the verbal and visual training methods; however, there is a suggestion favouring verbal training because it improved transfer test scores whereas visual training did not.

In the current study, only verbal training improved learning on the transfer test. This was probably because the expert in the visual training group was quite familiar with the learning content. Due to the large difference on prior knowledge, participants had difficulties in adjusting the strategy to their own learning pace. Krebs et al. (2019) found that perceived model-similarity influenced learning performance. Specifically, participants with low prior knowledge learned better when they perceived the model as a peer than when they perceived the model as an expert. Also, Scheiter et al. (2018) found that people with high prior knowledge benefited more from visual training than those with low prior knowledge. Therefore, visual training might not so beneficial to people who have relatively low prior knowledge, so it does not improve learning performance on the transfer test. Another possible explanation is that learners did not deeply process learning content of the first four slides as they may have had difficulties in following expert's eye movement. Although learning tests were only related to the last 12 slides, inadequate processing of first four slides may also influence how well learners were able to construct an overall mental model based on the whole learning phase.

Additionally, we did not find any differences among the groups on text-picture transitions. Our findings were consistent with Krebs et al. (2019), but other studies have found transition differences between groups (eg, Jian, 2021; Mason et al., 2015, 2016, 2017; Scheiter et al., 2018). This result might be because transitions between text and pictures were not always linked to integrative processing. Johnson and Mayer (2012) illustrated that learners who successfully integrated transited more between text and pictures, whereas Scheiter and Eitel (2015) did not find similar results. In addition, Arndt et al. (2015) found no relationship between the number of eye gaze shifts between text and pictures and integrative performance. For these mixed results, Schüler (2017) stated that attempts to integrate did not occur when learners looked from text to picture or vice versa. Therefore, in the current study, without shifting between text and pictures, learners might keep text information in their working memory and attempt to integrate information into a coherent mental model when watching the pictures.

Furthermore, mediation analyses showed that fixation on pictures mediated the effect of each training method on recall performance. This result is consistent with Krebs et al. (2019), indicating that increased pictorial processing was responsible for better learning performance. We also found a non-significant correlation between transitions between text and pictures and learning outcomes (recall: $r=0.11$, $p=0.409$; transfer: $r=0.13$, $p=0.299$). There are two potential explanations. Firstly, dual coding or pictorial processing was responsible for better learning outcomes, but not integrative behaviour. Secondly, the process of text-picture integration did not always happen during the eye shifts between text and pictures. Previous studies showed conflicting results. Schmidt-Weigand et al. (2010) found time spent on looking at the visualizations correlated with better learning performance, but the number of transitions did not influence learning outcomes. Holsanova et al. (2009) found that reading time predicted comprehension on newspapers, but the transition number did not. They suggested that the frequency of eye shifts might not reflect successful integrations. In contrast, Mason et al. (2017) found there were significant correlation relationships between the number of transitions and learning outcomes, and also partial mediation effect of gaze shifts. At this time, there is insufficient empirical evidence to conclude that gaze shifts between text and pictures can or cannot account for the integrative processing. Overall, we conclude that pictorial processing is important to improve learning, which is consistent to multimedia principle (Mayer, 2021).

## Theoretical implications

The present study provides preliminary empirical evidence for the pre-training principle on strategy acquisition and extends the pre-training principle to pre-training in computer-based multimedia processing strategies. According to the pre-training principle, pre-training on specific strategy, either verbally or visually, improves learners' strategy use and learning performance. Without an appropriate cognitive strategy, for example, the text-picture processing strategy, students sometimes ignore the illustrations and over-rely on the text, so they are unable to integrate text and picture information effectively (Hannus & Hÿonä, 1999; Renkl & Scheiter, 2017). After pre-training on text-picture processing strategy, learners successfully acquired the specific strategy, indicated by fixating more on pictures and less on text. Thus, by using a text-picture processing strategy helped learners to manage their processing during learning. Results also are consistent with the multimedia principle, which indicates sufficient processing of pictorial and verbal information leads to better performance.

## Practical implications

The present study suggests that both verbal and visual training are effective ways to support learners' strategy acquisition. Since verbal training can effectively improve recall and transfer performance and this form of training is convenient and less consuming, instructors should consider using verbal training for multimedia lessons.

## Limitations and future directions

The current study also has some limitations. First, paper-based instruction was provided in the verbal training group. Paper-based instruction can be inconvenient because students have to move their heads to read it when they practice the strategy. Future research can incorporate verbal training into the computer-based software, so that students can open the instruction with a click when they need to read it and close it when they do not want to read the instruction. Second, we did not measure cognitive load or perceived difficulty. Cognitive load or perceived difficulty can provide insights into the underlying mechanism of training method. Future research can use various measures to examine cognitive load or perceived difficulty. Third, the reliability of prior knowledge test was relatively low. It can be explained by several subtopics and unconnected details in the prior knowledge test. Further research can use a more internally reliable measure of people's prior knowledge. Fourth, we did not ask participants in the control group about strategy use. As they might use some other strategies in the learning phase, these strategies would influence their attention allocation on the learning content, and thus may have had an impact on learning outcomes. Future research can identify which strategy each learner used in the control group to determine whether is some influence of other strategies.

In this study, visual training did not improve the transfer test scores. The reason why might be that participants with relatively low prior knowledge had difficulties in following an expert's eye movements and acquire corresponding strategy adequately. Therefore, future research can examine the effect of people's prior knowledge on acquiring new strategies by visual training. Additionally, previous research (eg, Krebs et al., 2019; Mason et al., 2015, 2016; Scheiter et al., 2018) and the current study all examined the effect of pre-training on text-picture processing. Future research can generalize it to different strategies. Furthermore, future research can define finer AOIs of each keyword and corresponding pictorial elements to provide detailed insights into people' integrative behaviour.

# CONCLUSION

The current study aimed to examine whether visual and verbal training foster the acquisition of the text-picture processing strategy and improve learning, and compare the effectiveness of two pre-training methods on learning strategies. Moreover, this study investigated whether fixation time on pictures mediates the effects of pre-training methods on learning outcomes. Results showed both visual and verbal training could promote strategy use and facilitate learning. Verbal training improved transfer performance, whereas visual training did not. Mediation analysis showed that increased attention to pictures could be a potential explanation for better learning outcomes.

## ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China [Grant number: 62277025].

## CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data and study materials for this study are available from the corresponding author upon reasonable request.

## ETHICS STATEMENT

This research was approved by the Ethics Committee of the School of Psychology at Central China Normal University (#CCNU-IRB-202203011). The participants provided their written informed consent to participate in this study. They were informed that they had the right to withdraw from the study at any time. Confidentiality was ensured by using numbers instead of names in the research.

## REFERENCES

Ainsworth, S. (2022). The multiple representation principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 158–170). Cambridge University Press.

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education*, *125*, 413–428. https://doi.org/10.1016/j.compedu.2018.06.023

Arndt, J., Schüler, A., & Scheiter, K. (2015). Text-picture integration: How delayed testing moderates recognition of pictorial information in multimedia learning. *Applied Cognitive Psychology*, *29*, 702–712. https://doi.org/10.1002/acp.3154

Arslan-Ari, I. (2018). Learning from instructional animations: How does prior knowledge mediate the effect of visual cues? *Journal of Computer Assisted Learning*, *34*(2), 140–149. https://doi.org/10.1111/jcal.12222

Ayres, P., & Sweller, J. (2022). The split-attention principle in multimedia learning. In R. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (Cambridge Handbooks in Psychology, (pp. 199–211). Cambridge University Press. https://doi.org/10.1017/9781108894333.020

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology*, *96*, 523–535. https://doi.org/10.1037/0022-0663.96.3.523

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289–300. https://doi.org/10.2307/2346101

Bodemer, D., Plötzner, R., Bruchmüller, K., & Häcker, S. (2005). Supporting learning with interactive multimedia through active integration of representations. *Instructional Science*, *33*, 73–95. https://doi.org/10.1007/s11251-004-7685-z

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*, 1465–1477. https://doi.org/10.1016/j.cognition.2007.05.012

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology*, *98*, 182–197. https://doi.org/10.1037/0022-0663.98.1.182

Cagiltay, K., & Coskun, A. (2022). A systematic review of eye-tracking-based research on animated multimedia learning. *Journal of Computer Assisted Learning*, *38*(2), 581–598. https://doi.org/10.1111/jcal.12629

Eitel, A., Scheiter, K., & Schüler, A. (2013). How inspecting a picture affects processing of text in multimedia learning. *Applied Cognitive Psychology*, *27*, 451–461. https://doi.org/10.1002/acp.2922

Emhardt, S. N., Kok, E., van Gog, T., Brandt-Gruwel, S., van Marlen, T., & Jarodzka, H. (2023). Visualizing a task performer's gaze to foster observers' performance and learning—A systematic literature review on eye movement modeling examples. *Educational Psychology Review*, *35*(1), 23. https://doi.org/10.1007/s10648-023-09731-7

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694–724. https://doi.org/10.1037/0033-2909.133.4.694

Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology*, *67*(8), 850–857. https://doi.org/10.1016/j.jclinepi.2014.03.012

Hannus, M., & Hÿonä, J. (1999). Utilization of illustrations during learning of science textbook passages among low- and high-ability children. *Contemporary Educational Psychology*, *24*, 95–123. https://doi.org/10.1006/ceps.1998.0987

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language*, *32*, 717–742. https://doi.org/10.1006/jmla.1993.1036

Holsanova, J., Holmberg, N., & Holmqvist, K. (2009). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology*, *23*, 1215–1226. https://doi.org/10.1002/acp.1525

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction*, *20*(2), 172–176. https://doi.org/10.1016/j.learninstruc.2009.02.013

Hyönä, J., Lorch, R. F. J., & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, *94*, 44–55. https://doi.org/10.1037/0022-0663.94.1.44

Jarodzka, H. (2022). Research methods in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 41–54). Cambridge University Press.

Jarodzka, H., Van Gog, T., Dorr, M., Scheiter, K., & Gerjets, P. (2013). Learning to see: Guiding students' attention via a model's eye movements fosters learning. *Learning and Instruction*, *25*, 62–70. https://doi.org/10.1016/j.learninstruc.2012.11.004

Jian, Y. (2021). The immediate and delayed effects of text–diagram reading instruction on reading comprehension and learning processes: Evidence from eye movements. *Reading and Writing*, *34*, 727–752. https://doi.org/10.1007/s11145-020-10089-3

Jian, Y. C. (2018). Reading instructions influence cognitive processes of illustrated text reading not subject perception: An eye-tracking study. *Frontiers in Psychology*, *9*, 2263. https://doi.org/10.3389/fpsyg.2018.02263

Jian, Y. C. (2019). Reading instructions facilitate signaling effect on science text for young readers: An eye-movement study. *International Journal of Science and Mathematics Education*, *17*, 503–522. https://doi.org/10.1007/s10763-018-9878-y

Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied*, *18*, 178–191. https://doi.org/10.1037/a0026923

Kombartzky, U., Ploetzner, R., Schlag, S., & Metz, B. (2010). Developing and evaluating a strategy for learning from animations. *Learning and Instruction*, *20*, 424–433. https://doi.org/10.1016/j.learninstruc.2009.05.002

Krebs, M., Schüler, A., & Scheiter, K. (2019). Just follow my eyes: The influence of model-observer similarity on eye movement modeling examples. *Learning and Instruction*, *61*, 126–137. https://doi.org/10.1016/j.learninstruc.2018.10.005

Krebs, M., Schüler, A., & Scheiter, K. (2021). Do prior knowledge, model-observer similarity and social comparison influence the effectiveness of eye movement modeling examples for supporting multimedia learning? *Instructional Science*, *49*, 607–635. https://doi.org/10.1007/s11251-021-09552-7

Larson, C. O., Dansereau, D. F., Hythecker, V. I., O'Donnell, A., Young, M. D., Lambiotte, J. G., & Rocklin, T. R. (1986). Technical training: An application of a strategy for learning structural and functional information. *Contemporary Educational Psychology*, *11*, 217–228. https://doi.org/10.1016/0361-476X(86)90018-4

Levin, D. T., Salas, J. A., Wright, A. M., Seiffert, A. E., Carter, K. E., & Little, J. W. (2021). The incomplete tyranny of dynamic stimuli: Gaze similarity predicts response similarity in screen-captured instructional videos. *Cognitive Science*, *45*, e12984. https://doi.org/10.1111/cogs.12984

Litchfield, D., Ball, L. J., Donovan, T., Manning, D., & Crawford, T. J. (2010). Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *Journal of Experimental Psychology: Applied*, *16*(3), 251–262. https://doi.org/10.1037/a0020082

Lorch, R. F., Lemarié, J., & Chen, H. T. (2013). Signaling topic structure via headings or preview sentences. *Psicologia Educativa*, *19*, 59–66. https://doi.org/10.1016/S1135-755X(13)70011-3

Mason, L., Pluchino, P., & Tornatora, M. C. (2015). Eye-movement modeling of integrative reading of an illustrated text: Effects on processing and learning. *Contemporary Educational Psychology*, *41*, 172–187. https://doi.org/10.1016/j.cedpsych.2015.01.004

Mason, L., Pluchino, P., & Tornatora, M. C. (2016). Using eye-tracking technology as an instruction tool to improve text and picture processing and learning. *British Journal of Educational Technology*, *47*, 1083–1095. https://doi.org/10.1111/bjet.12271

Mason, L., Scheiter, K., & Tornatora, C. (2017). Using eye movements to model the sequence of text-picture processing for multimedia comprehension. *Journal of Computer Assisted Learning*, *33*(5), 443–460. https://doi.org/10.1111/jcal.12191

Mason, L., Tornatora, M. C., & Pluchino, P. (2013). Do fourth graders integrate text and picture in processing and learning from an illustrated science text? Evidence from eye-movement patterns. *Computers & Education*, *60*, 95–109. https://doi.org/10.1016/j.compedu.2012/07/011

Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction*, *20*(2), 167–171. https://doi.org/10.1016/j.learninstruc.2009.02.012

Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, *33*(5), 403–423. https://doi.org/10.1111/jcal.12197

Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press.

Mayer, R. E. (2022a). The cognitive theory of multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 57–72). Cambridge University Press. https://doi.org/10.1017/9781108894333.008

Mayer, R. E. (2022b). The multimedia principle. In R. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (pp. 145–157). Cambridge University Press. https://doi.org/10.1017/9781108894333.015

Mayer, R. E., & Fiorella, L. (2022). Principles for managing essential processing in multimedia learning: Segmenting, pre-training, and modality principles. In L. Fiorella & R. E. Mayer (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 243–260). Cambridge University Press. https://doi.org/10.1017/9781108894333.025

Mayer, R. E., & Pilegard, C. (2014). Principles for managing essential processing in multimedia learning: Segmenting, pre-training, and modality principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 316–344). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.016

Meier, C., Frank, C., Gröben, B., & Schack, T. (2020). Verbal instructions and motor learning: How analogy and explicit instructions influence the development of mental representations and tennis serve performance. *Frontiers in Psychology*, *11*, 2. https://doi.org/10.3389/fpsyg.2020.00002

Meyer, O. A., Omdahl, M. K., & Makransky, G. (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. *Computers & Education*, *140*, 103603. https://doi.org/10.1016/j.compedu.2019.103603

Paas, F., Renkl, A., & Sweller, J. (2010). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria. https://www.R-project.org/

Renkl, A., & Scheiter, K. (2017). Studying visual displays: How to instructionally support learning. *Educational Psychology Review*, *29*(3), 599–621. https://doi.org/10.1007/s10648-015-9340-4

Sanchez, R. P., Lorch, E. P., & Lorch, R. F. J. (2001). Effects of headings on test processing strategies. *Contemporary Educational Psychology*, *26*, 418–428. https://doi.org/10.1006/ceps.2000.1056

Scheiter, K., & Eitel, A. (2015). Signals foster multimedia learning by supporting integration of highlighted text and diagram elements. *Learning and Instruction*, *36*, 11–26. https://doi.org/10.1016/j.learninstruc.2014.11.002

Scheiter, K., Schubert, C., & Schüler, A. (2018). Self-regulated learning from illustrated text: Eye movement modelling to support use and regulation of cognitive processes during learning from multimedia. *British Journal of Educational Psychology*, *88*, 1–15. https://doi.org/10.1111/bjep.12175

Scheiter, K., & Van Gog, T. (2009). Using eye tracking in applied research to study and stimulate the processing of information from multi-representational sources. *Applied Cognitive Psychology*, *23*(9), 1209–1214. https://doi.org/10.1002/acp.1524

Schlag, S., & Ploetzner, R. (2011). Supporting learning from illustrated texts: Conceptualizing and evaluating a learning strategy. *Instructional Science*, *39*, 921–937. https://doi.org/10.1007/s11251-010-9160-3

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system- and self-paced instruction in multimedia learning. *Learning and Instruction*, *20*(2), 100–110. https://doi.org/10.1016/j.learninstruc.2009.02.011

Schnotz, W. (2022). Integrated model of text and picture comprehension. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 82–99). Cambridge University Press. https://doi.org/10.1017/9781108894333.010

Schnotz, W. (2023). *Multimedia comprehension*. Cambridge University Press.

Scholz, A., von Helversen, B., & Rieskamp, J. (2015). Eye movements reveal memory processes during similarity- and rule-based decision making. *Cognition*, *136*, 228–246. https://doi.org/10.1016/j.cognition.2014.11.019

Schüler, A. (2017). Investigating gaze behavior during processing of inconsistent text-picture information: Evidence for text-picture integration. *Learning and Instruction*, *49*, 218–231. https://doi.org/10.1016/j.learninstruc.2017.03.001

Seufert, T. (2019). Training for coherence formation when learning from text and picture and the interplay with learners' prior knowledge. *Frontiers in Psychology*, *10*, 1–11. https://doi.org/10.3389/fpsyg.2019.00193

Stalbovs, K., Scheiter, K., & Gerjets, P. (2015). Implementation intentions during multimedia learning: Using if-then plans to facilitate cognitive processing. *Learning and Instruction*, *35*, 1–15. https://doi.org/10.1016/j.learninstruc.2014.09.002

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*, 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tabbers, H. K., Paas, F., Lankford, C., Martens, R. L., & van Merriënboer, J. J. (2008). Studying eye movements in multimedia learning. In J. F. Rouet, R. Lowe, & W. Schnotz (Eds.), *Understanding multimedia documents*. Springer. https://doi.org/10.1007/978-0-387-73337-1_9

Van Duijn, T., Hoskens, M. C., & Masters, R. S. (2019). Analogy instructions promote efficiency of cognitive processes during hockey push-pass performance. *Sport, Exercise, and Performance Psychology*, *8*(1), 7–20. https://doi.org/10.1037/spy0000142

Van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., & Paas, F. (2009). Attention guidance during example study via the model's eye movements. *Computers in Human Behavior*, *25*, 785–791. https://doi.org/10.1016/j.chb.2009.02.007

Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, *22*, 155–174. https://doi.org/10.1007/s10648-010-9134-7

Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, *20*(2), 95–99. https://doi.org/10.1016/j.learninstruc.2009.02.009

Vine, S. J., Moore, L. J., & Wilson, M. R. (2011). Quiet eye training facilitates competitive putting performance in elite golfers. *Frontiers in Psychology*, *2*, 8. https://doi.org/10.3389/fpsyg.2011.00008

Wang, F., Zhao, T., Mayer, R. E., & Wang, Y. (2020). Guiding the learner's cognitive processing of a narrated animation. *Learning and Instruction*, *69*, 101357. https://doi.org/10.1016/j.learninstruc.2020.101357

Wilson, M. R., Vine, S. J., Bright, E., Masters, R. S., Defriend, D., & McGrath, J. S. (2011). Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: A randomized, controlled study. *Surgical Endoscopy*, *25*(12), 3731–3739. https://doi.org/10.1007/s00464-011-1802-2

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Leng, X., Wang, F., Mayer, R. E., & Zhao, T. (2024). How to train students to engage in text-picture integration for multimedia lessons. *British Journal of Educational Technology*, *55*, 1167–1188. https://doi.org/10.1111/bjet.13419