Information Processing and Management 60 (2023) 103347

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Visual-audio correspondence and its effect on video tipping: Evidence from Bilibili vlogs

Bu Li, Jichang Zhao *

School of Economics and Management, Beihang University, Beijing, China

ARTICLE INFO

Keywords:
Streaming videos
Visual-audio correspondence
Deep neural network
Pay What You Want
Consumer satisfaction
Tipping behavior

ABSTRACT

Video tipping takes a remarkable share in the income of online streaming platforms such as Bilibili. There are some specific mappings between the audio and visual signals that viewers can sense (e.g., congruency of pitch and size), which is generally called visual-audio correspondence (VAC). And it is believed to influence viewer satisfaction with video clips. The way to automatically measure VAC, however, still remains missing and its possible effect on video tipping is rarely examined in previous efforts. In this study, a deep neural network with two sub-networks, namely VAC-Net, is established to map both visual and audio stimuli into a shared embedding space. And the Euclidean distance between visual and audio representations in this space is accordingly presented to be the indicator of VAC. Pre-trained models of both modalities and the triplet loss are further leveraged to train the VAC-Net and it competently evaluates VAC of video clips with a test accuracy of 68.37% by outperforming alternative baselines and even exceeding humans on the similar task. Lab-experiments further show that the VAC measurement of VAC-Net conforms to human cognition. Second, considering that viewers' tipping behavior (TIP) on videos is consistent with the pricing strategy Pay What You Want (PWYW), it is hypothesized that VAC would indirectly influence TIP by reshaping viewer satisfaction (VS). Regression models are thus built to test the hypotheses and it is found that VAC can promote TIP by enhancing VS significantly. Additional tests also demonstrate the robustness of this mechanism by considering various controls and measurement errors. Our results supplement PWYW in streaming videos with a new motive of VAC for viewer tipping and provide streaming practitioners with an automatic tool to estimate the tips videos will receive.

1. Introduction

Watching streaming videos is becoming a predominant part of daily life today. In particular, there are two basic models to monetize uploaders' videos. One is adding ads to the videos to indirectly get revenue. The other is earning income directly through viewers' tipping, which is more closely linked to the video itself. Tipping videos first appeared in live-streaming (e.g., Twitch, Douyu) (Zhou, Zhou, Ding, & Wang, 2019), and later was pervasively adopted in common online streaming video platforms (e.g., Bilibili, Douyu, and AcFun) as an important source of income. So viewer tipping behavior is critical to the business model of streaming videos. According to the information-processing model in cognitive psychology, video is made up of two streams of stimuli, one audio and one vision (Lang, 2006), which is 'sensual' in nature. And there are some spontaneous mappings between the audio and visual signals, such as pitch, size, and frequency (Evans & Treisman, 2010; Spence, 2011). The compatibility effect of these mappings is named 'visual-audio correspondence' (VAC) (Spence, 2011). VAC indicates that viewers' responses to visual and

* Corresponding author.
E-mail address: jichang@buaa.edu.cn (J. Zhao).

https://doi.org/10.1016/j.ipm.2023.103347
Received 6 August 2022; Received in revised form 28 January 2023; Accepted 11 March 2023
Available online 23 March 2023

audio stimuli are interactional rather than simply aggregated. In line with this, the present study is aimed to explore how videos' VAC that viewers sense influence its income from tipping.

The first issue we attempt to address is to automatically measure the VAC. As digital products, visual and audio sensory stimuli shape the holistic customer experience (Petit, Velasco, & Spence, 2019). And Bolivar, Cohen, and Fentress (1994) already confirmed that VAC indeed exists during watching videos. In this view, VAC plays a vital role in shaping viewers' experiences. In addition, the large number of videos increases the difficulty of video screening and management, which leads to the requirement of automatically evaluating videos. Actually, unlike variable video content, the VAC is nonarbitrary (Spence, 2011), which means that it is possible to automatically measure the degree of correspondence by extracting appropriate visual and audio features. However, rare previous studies examined VAC in videos. As VAC can be broadly divided into semantic congruency (such as consistent in content or theme) and structural congruency (such as consistent in pitch and size) (Lang, 2006), previous studies that examined VAC only focused on a specific aspect of congruency, such as calm or chaotic (Becker-Olsen, 2006), eastern or western (Lalwani, Lwin, & Ling, 2009), which are further limitations in realistic applications. In the meantime, their measurements are often implemented through manual labeling that always brings about heavy labor cost (Becker-Olsen, 2006; Lalwani et al., 2009). These measurements are not only hard to obtain quantitative indicators instead of discrete variables (Demoulin, 2011; Spence, 2011) but also prone to introduce subjective bias and ignored subconscious factors which are confirmed to exist in VAC (Evans & Treisman, 2010; Spence, 2011). It is thus anticipated in this study to automatically measure the VAC through an object indicator which is more comprehensive to fill the void in earlier research.

And the second issue we seek to address is how to determine the effect of VAC on viewer tipping. Many studies about online tipping examined tipping in live-streaming, but few of them focused on streaming videos (Li, Lu, Ma, & Wang, 2021; Zhou et al., 2019). However, studies about tipping in live-streaming are mostly focused on real-time interaction (Li et al., 2021), which is never the primary feature for tipping in streaming videos. Viewers on streaming platforms, such as Bilibili, have a kind of virtual currency, called 'coin'. They can access a video free of charge, and choose to tip the video only once by giving one or two coins, i.e., a fixed price determined by viewers. Indeed, as Kim, Natter, and Spann (2009) pointed out, this behavior can be perfectly manifested by Pay What You Want (PWYW). According to the recent advances in PWYW, consumer satisfaction acts an important role in determining consumers' payment (Kim et al., 2009; Kunter, 2015), and consumer satisfaction which is defined as the consumer's post-consumption evaluation of the perceived quality. Viewer satisfaction proposed would also be based on the evaluation of the quality of the video. Accordingly, Demoulin (2011) contrasted customers' responses in restaurants with congruent and non-congruent music and found that music congruency increases consumers' evaluation of the quality and return intention. And similar results were further found and confirmed in other studies about offline products (Brengman, Willems, & De Gauquier, 2022; Krishna, Cian, & Aydınoğlu, 2017; Oakes & North, 2008). However, the study about VAC is all about offline scenes, and the results of these studies cannot be simply expanded into online scenes because of the quite difference between online and offline purchase behavior and satisfaction (Hult, Sharma, Morgeson, & Zhang, 2019). In addition, Kunter (2015) also demonstrated that different PWYW applications result in the different significance of satisfaction. It is therefore still necessary to justify the effect of VAC on viewer satisfaction, and satisfaction on tipping in streaming videos. And based on these previous advances, we posit that VAC would indirectly influence viewer tipping by improving viewer satisfaction in online streaming video platforms.

To automatically measure VAC, in this research, a deep neural network named VAC-Net is developed to capture the similarity between visual and audio stimuli in video clips. Two sub-networks and pre-trained models are leveraged to respectively extract visual and audio features, which possess the evident superiority over "hand-crafted" features (Rawat & Wang, 2017; Yosinski, Clune, Bengio, & Lipson, 2014). A triplet loss is employed to map these two modalities, i.e., visual and audio, into a shared Euclidean space where distances directly correspond to their similarity (Schroff, Kalenichenko, & Philbin, 2015), i.e., VAC. Color and temporal information are both important for VAC, but most previous studies about cross-modal learning only took one of them into account (Arandjelovic & Zisserman, 2018; Chung, Chung, & Kang, 2019). The neural network we proposed uses color images with temporal information from video clips as the input to capture more useful information. VAC-Net is proved to be a competent approach to measuring VAC, whose test accuracy achieves 68.37%, exceeding alternative baselines and even human performance (Owens & Efros, 2018) on different datasets. And the neural network can be reliably applied to videos of other domains and platforms, indicating its impressive generalization capability. An additional lab experiment also confirms that the VAC measurement derived from the proposed neural network indeed conforms to human cognition.

To determine the effect of VAC on viewer tipping, based on the theories mentioned above, we use VAC and viewer satisfaction (VS) to explain viewer tipping (TIP) with the help of multiple linear regression. A major advantage of our task is that we could collect various relevant attributes (such as numbers of likes, views, and coins, i.e., tips) directly from real scenarios so questionnaires are no longer needed. According to the definition of consumer satisfaction (Oliver, 2010), we thus chose the number of likes a video received as the measure of VS. And intuitively, we measure viewer tipping behavior in terms of the number of coins a video received. The robustness of the effect of VAC on viewer tipping is extensively tested by adding various controls and taking VAC measurement error that induced by the neural network into consideration. Furthermore, we examine the mediation effect of VS on VAC influencing viewer tipping. The result of our regression model is a complete mediation and passes both Sobel and Bootstrapping tests. As a result, we have confirmed that VS is the mediator that mediates the positive effect of VAC on viewer tipping.

One of the main contributions of this study is to design and implement a new deep neural network, VAC-Net, that integrates both audio and visual signals to learn representations that correspond to the VAC of a video clip. And the other is that we examine the effect of VAC in streaming video platforms, and demonstrate that the increase in visual-audio correspondence could enhance viewer satisfaction and indirectly lead to viewer tipping behavior. Our results extend the theory of PWYW and the application of multi-sensory marketing to the scenario of online streaming video and we also supplement a new determinant that influences

consumer satisfaction with digital products, i.e., VAC. For the first time, to our knowledge, these findings confirm the positive effect of multisensory interaction on viewer tipping in streaming video. In addition, for streaming practitioners such as platforms and uploaders, the presented deep neural network will provide an automatic tool to estimate the VAC of videos. As VAC is related to tipping, this tool could be employed to improve video quality and also screen out those that will solicit more tips after being published. It would even inspire specific instructions on how to optimize videos in terms of tuning VAC to earn more income.

2. Theoretical background and research hypotheses

2.1. Visual-audio correspondence

We receive sensation information all the time, and we interpret them, turning the sensation information into a meaningful experience, which is called perception. It is worth mentioning that, sensation information is not always consistent with our perception and this bias is caused by many different factors (Raghubir & Krishna, 1996). One of the most famous examples is Café wall illusion (Gregory & Heard, 1979), in which the lines with alternating dark and light bricks appear to be sloped, not parallel as they really are. Similarly, this bias between stimuli and perception can also be caused by the cross-modal interaction and then even affect people's behavior. In this research, we would focus on the perception of visual and audio sensations and investigate the effect of cross-modal interaction instead of the information in them.

Many psychologists did investigate the interaction between different modalities and have provided evidence for the existence of cross-modal correspondences, especially for VAC (Mondloch & Maurer, 2004; Spence, 2011; Walker et al., 2010). There are indeed some links between different sensory modalities. Walker et al. (2010) and Mondloch and Maurer (2004) even showed that the correspondence between different modalities is an innate or unlearned aspect of perception. They found that young infants could recognize the correspondence linking auditory pitch and visual sharpness. We could infer from these findings that the way we perceive VAC is similar. For instance, almost all of us hold the view that brighter visual images should induce higher-pitched sounds instead of lower (Evans & Treisman, 2010). So VAC could be seen as an objective relation between visual and audio stimuli, which is possible for us to measure automatically.

Previous studies first confirmed the existence of VAC in videos (Herget, 2021; Lipscomb & Kendall, 1994; Spence, 2011; Walker et al., 2010). They used real physical objects and pictures as the representation of visual stimuli which are already very close to the stimuli that a video provides. And as the video is only made up of audio and visual streams, the VAC viewers perceived from videos should be stronger than that from others such as pictures. Accordingly, the exploration of VAC could be expanded to the VAC of streaming videos naturally. The VAC of videos is not only subconscious (Evans & Treisman, 2010; Spence, 2011) but also related to aesthetics. Bolivar et al. (1994) showed that subjects have the ability to match the motion picture with their composer-intended musical scores, the correspondence of which is also relevant to specific subject ratings. It indicates that a musical soundtrack can change the meaning of a film presentation. Lipscomb and Kendall (1994) highlighted the important position of music in perception in terms of letting subjects select the composer-intended musical scores to match different scenes. These findings further confirm the existence of VAC in videos and the possibility to measure it. So we aim to extract VAC-related features, which include both semantic and structural information, from visual and audio streams and attempt to build an indicator of VAC. Furthermore, we will also investigate its effect on viewer behaviors like tipping.

The exploration of the effect of VAC has begun very early (Bernstein & Edelstein, 1971; Evans & Treisman, 2010). And they found that the sensation of different modalities is not processed independently and they would interact and influence people's perception (Brengman et al., 2022; Parise & Spence, 2009). As a result, VAC would positively affect people's perception and judgment (Maeda, Kanai, & Shimojo, 2004; Parise & Spence, 2009). For example, Maeda et al. (2004) asked subjects to make a two-alternative forced choice (upward or downward) depending on the visual motion of grating which is made of two superimposed, oppositely moving gratings. The result showed that the grating with ambiguous motion would more likely be seen as an upward motion when it is accompanied by ascending pitch, and those accompanied by descending pitch as a downward motion. These studies revealed that the congruency part of visual and audio stimuli can be enhanced in perception and become easier to be noticed. And its effect on perception is not only reflected in judgment but also in attitude and even the speed or accuracy of behavior (Evans & Treisman, 2010; Parise & Spence, 2009). More recently, Chen, Huang, Faber, Makransky, and Perez-Cueto (2020) found that the perceived sweetness of the beverage is significantly elevated in a sweet-congruent environment versus others in immersive virtual reality (VR).

The effect of VAC on perception could be further enhanced in the scenario of streaming video. (Bolivar et al., 1994) confirmed that the interaction between visual and audio streams of video reshapes the evaluation of video through congruency, but has not further investigated the effect on viewer behavior. VAC also contributes to the emotional aspects of films, and even enhances the effectiveness of a message in advertisement (Kellaris, Cox, & Cox, 1993). A similar study investigated how congruency between music influences attitudes and purchase intentions towards products (Lalwani et al., 2009) and showed that this congruency leads to positive attitudes and persuasion. However, to our knowledge, most previous studies only considered video as a mediator in exploiting the effect of VAC on viewer experience. Nevertheless, in streaming platforms, videos become digital products and few efforts were devoted to how VAC affects the viewers' perception of the video itself. So we accordingly aim to expand the understanding of VAC into the new scenario of the streaming scene and focus more on how it changes viewers' experience of watching videos.

Though many studies have proved the positive effect of multisensory interaction in the video or offline marketplace on human perception as mentioned above, our study on streaming videos is still necessary. Firstly, the scenario investigated in our study is quite

different from the previous studies. There are few studies that investigated how VAC affects tipping behavior on streaming video platforms where the payment object is the video itself. In addition, even the exploration of the attitudes towards advertisements on the same media is indeed distinctive (Logan, 2011). Other studies about payment focused on the VAC of advertisement (Kellaris et al., 1993), which serves for buying another commodity instead of the video itself. So investigating the effect on viewers' behavior and experience in streaming videos is indispensable and novel.

Secondly, the VAC of our study is different from previous studies coming through in distinctive correspondence definitions and video sources. Traditional VAC only focused on a specific part as mentioned above. Instead, we establish a neural network to learn a more comprehensive VAC directly from visual-audio clips. In addition, the videos we use are also quite different from previous studies. We collected a large number of videos that were filmed and post-processed spontaneously by uploaders instead of only a few fabricated samples carefully selected by researchers. These differences lead our study to more general and managerial implications.

Good sensory experience could arouse aesthetical pleasure, excitement, and satisfaction (Gentile, Spiller, & Noci, 2007). Schmitt (1999) who mentioned the sensory experiences earlier even took the cognitive consistency as a key principle of sense which could arouse satisfaction. VAC we focus on belongs to sensory experience according to the definition. Nesbitt and Hoskens (2008) found that computer games with a better multisensory display could improve players' satisfaction instead of performance, which is similar to the VAC of videos. Oliver (2010) further suggested that satisfaction is affected by pleasure including enhancement of stimulus, especially for hedonic products. As VAC may enhance viewers' additional perception, it may improve the sensory experience of watching videos, which means an improvement in viewers' satisfaction. So we hypothesize H 1 as the following:

Hypothesis 1. There is a positive relationship between visual-audio correspondence (VAC) and viewer satisfaction (VS) with video.

2.2. Pay What You Want and tipping behavior

With the development of streaming videos, more and more platforms begin to provide the feature of tipping for viewers (e.g., Bilibili, Douyu, and AcFun), which offers a path to reward videos' popularity and is an ideal proxy for videos' potential value. Before we examine how VAC influences tipping, it is important to first define this behavior. Though tipping behavior in restaurants and live-streaming has been exploited thoroughly by many efforts (Lu, Xia, Heo, & Wigdor, 2018), this kind of behavior in streaming video is really different compared to others. As mentioned before, on streaming video platforms, videos are fixed once uploaded. It is impossible for uploaders to interact with viewers in real-time. What is more, it is also impossible to know when the video would be watched by whom. So past studies on tipping behavior in restaurants and live-streaming cannot be simply extended into this domain, and the tipping behavior on streaming video is more consistent with the price strategy which is called PWYW (Kim et al., 2009).

Racherla, Babb, and Keith (2011) argued that the most special part of PWYW is that consumers would like to pay for a good that could be obtained for free (sellers do not set a threshold price). So the tipping behavior in streaming video platforms such as Bilibili is accordant with the core of PWYW, in which viewers watch streaming videos for free, but some of them still choose to tip with coins. Actually, PWYW has already been widely applied for services and digital goods (Kunter, 2015). One example is that of the rock band Radiohead, which offered its new album on the website, and fans who downloaded it could pay as much as they wanted (Kim et al., 2009).

As an increasingly popular pricing approach, many studies examined factors that reshape PWYW payments. The landmark paper of PWYW by Kim et al. (2009) revealed that satisfaction could affect payments. Kunter (2015) further categorized motivation-related factors and revealed that customer satisfaction is relatively important in certain scenarios. Social factors (e.g., fairness, loyalty, altruism, and income) have also been investigated to a sufficient extent (Gneezy, Gneezy, Riener, & Nelson, 2012; Kahsay & Samahita, 2015; Marett, Pearson, & Moore, 2012; Roy, Rabbanee, & Sharma, 2016), which could be quantified by setting reference price or referring to other social information. However, these social factors are related to the personal characteristics and social identity of customers, and cannot be directly influenced by service-providing firms, instead, satisfaction is mainly determined by the product itself. In addition, according to Kunter (2015), in the applications of the exhibition, like zoos and museums, satisfaction is relatively important. As for streaming videos, viewers would watch videos in advance, and then make the decision to tip for the watching experience, which is similar to an exhibition. So it is worth investigating whether satisfaction plays a role in PWYW of streaming video.

In addition, most extant studies on the mechanism of PWYW were performed in offline scenes (Kunter, 2015; Roy et al., 2016). Though few researchers explored motivations of payment in online PWYW settings (Lu, Yao, Chen, & Grewal, 2021; Weisstein, Kukar-Kinney, & Monroe, 2016), few of them took digital products like streaming videos and VAC into account. For example, Lu et al. (2021) only assessed the effect of popularity on viability of PWYW. The association between online digital goods and PWYW payment still deserves more in-depth explorations and the scenario of streaming video provides a proper setting. Unlike traditional offline products, customers pay for digital products, e.g., streaming videos, to get a digital experience. In this view, satisfaction is expected to be more determinant in video tipping. So our study differs from previous research on PWYW by expanding it to streaming videos and exploring the factor related to satisfaction from the service-providing side, i.e., the video itself. Accordingly, we hypothesize H 2 as follows.

Hypothesis 2. Viewer satisfaction (VS) has a positive effect on viewer tipping behavior (TIP).

Overall, VS is closely associated with both VAC and TIP according to the theoretical background. Besides, sensory marketing holds the view that different sensory stimuli would affect consumers' perception and then behaviors (Krishna, 2012). A similar framework has also been proposed by environmental psychology and has been widely adopted. It is called the S–O–R framework, which suggests stimuli (Stimuli) consumers received could affect their internal states and experience (Organism) and indirectly influence their behaviors (Response) (Koo & Ju, 2010). This theory connects stimuli and consumers' behavior naturally. Previous studies have already demonstrated the relationship between environmental VAC and consumer behavior. To be specific, a number of cross-sectional studies have proved that the congruency of different types of stimulation, including VAC, could enhance consumers' experience and even consumer behavior as mentioned before (Brengman et al., 2022; Demoulin, 2011; Oakes & North, 2008). Inspired by these studies, we expand further into streaming video and PWYW domain and pose a hypothesis H 3 as follows:

Hypothesis 3. The positive effect of VAC on TIP is mediated by VS with video.

2.3. Cross-modal learning

In this research, we aim to evaluate VAC automatically. Therefore, we have to handle data of different modalities. Many approaches in deep learning have been proposed to solve cross-modal tasks, however, principally in the form of images and text (Song & Soleymani, 2019; Zhao et al., 2019). While neither video nor audio is simple for the deep neural network to learn since both modalities are more veiled than text. Fortunately, many deep neural networks have been successfully built to capture the interaction between visual and audio and pervasively employed in tasks such as lip sync (Chung & Zisserman, 2017), sound source location (Arandjelovic & Zisserman, 2017, 2018) and action recognition (Korbar, Tran, & Torresani, 2018). One of the primary reasons for their competence is that the deep neural network could extract meaningful features from multimedia data and produce a high-level representation. And most models were implemented in terms of two-stream architecture which consists of two sub-networks and several fusion layers. Each sub-network handles one modality, and then the representations they provide will be fed into fusion layers to capture the interaction between these two different modalities. So the representations sub-networks provide and the way they are fed into fusion layers (Suris, Duarte, Salvador, Torres, & Giro-i Nieto, 2018) are both crucial to VAC measurement.

The choice of sub-network is simply determined by the form of input data. Past studies tend to focus on the visual-audio synchronization of a single component of a picture or a video frame (e.g., person, face, and lips), so the consecutive video frames are input into the network but the frames are always cropped or compressed into black and white (Chung et al., 2019; Chung & Zisserman, 2017). As for studies that took the relationship between whole visual and audio streams into account, they treated the video as a still image and only matched a single video frame with a short audio clip (Arandjelovic & Zisserman, 2017, 2018), which reluctantly ignored the temporal correspondence. However, as for VAC, to extract key visual signals from the video, the better way is to select dynamic color video clips instead of a single video frame or frames compressed into black and white. A streaming video is composed of multiple frames, so integrating temporal information of visual and audio streams is critical for capturing semantic and structural features. And color is also an important factor that affects video processing (Haber & Hershenson, 1973), color images thus are also necessary. Accordingly, unlike other studies, we would use color images with temporal information of video clips as the input of the visual sub-network and audio clips of the same length with visual clips as the input of the audio sub-network.

Data with more information and dimensions always means more sophisticated models. For example, the raw input of an image could be a matrix of pixels in three-dimensional which correspond to RGB, let alone consecutive color video frames of a video clip. The convolutional neural network (CNN) has become dominant in computer vision and its core building block is the convolutional layer. The trainable convolution kernel (or filter) in this layer, which is a small grid of parameters, slides along the input at each image position so that it could finally provide a feature map. Though achieving superior performance, CNN cannot capture both spatial and temporal dimensions in videos. Ji, Xu, Yang, and Yu (2013) developed a novel 3D CNN model which could deal with consecutive video frames. Since then, 3D CNN is proved as a rewarding approach to learning spatio-temporal video representation (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015; Varol, Laptev, & Schmid, 2018), and we attempt to use 3D CNNs to capture features from video clips.

While regarding the exact architecture of the visual sub-network, much evidence reveals that network depth is of crucial importance (Simonyan & Zisserman, 2014), but it also brings about more parameters and costs in the training. As a remarkable advance, He, Zhang, Ren, and Sun (2016) proposed ResNet based on CNN which could ease the training of deep networks by using residual learning and still retain good even better performance. Tran et al. (2018) then reconsidered 3D CNNs within the framework of residual learning. In addition, they use "(2+1)D" convolutional block as a substitute for 3D convolution, which factorizes 3D convulsion into a 2D spatial convolution and a 1D temporal convolution in a separate and successive manner. This model called R(2+1)D has better performance and is easier to optimize than those of 3D CNN. However, the model size of 3D CNNs or even (2+1)D CNNs still experiences a quadratic growth compared to 2D CNNs which results in expensive computational costs. Fortunately, using the pre-trained model could address this issue further. Because of the above-mentioned advantages, we use this R(2+1)D pre-trained model and re-train (or fine-tune) the weights of selected layers to enhance the model's performance in VAC measurement.

As for the audio sub-network, previous efforts have demonstrated that CNN also yields excellent performance on audio classification (Hershey et al., 2017), because audio can be transformed into spectrogram images. After that, some CNN models attempted to learn the audio representation with the help of cross-modal interactions. Such as SoundNet (Aytar, Vondrick, & Torralba, 2016) which takes both audio and visual signals into account. As a goal-oriented model, SoundNet performs well in audio classification, however, it cannot ensure the audio features it extracts contain sufficient information for VAC measurement. Unlike SoundNet, VGGish (Hershey et al., 2017), a pre-trained model that considers the congruency between audio and visual signals, could

be used as a feature extractor to convert the input audio clip into a semantically meaningful high-level 128-D embedding. In this research, therefore, we use VGGish as another sub-network to extract audio features.

After extracting informative representations from both visual and audio inputs, the way they are fed into fusion layers is crucial to learn VAC as well. So far, there are two ways to deal with the representations in general. One is concatenating two features (Arandjelovic & Zisserman, 2017) or converting two features into one vector by other measures, such as calculating similarity or distance (Arandjelovic & Zisserman, 2018). While the representations of audio and visual features are not aligned actually. In contrast, the second way which is also the way we use is that the visual and audio representations are trained with some special loss functions (e.g., contrastive loss or triplet loss) by building a joint embedding space the distance in which carries meanings (Hoffer & Ailon, 2015; Song & Soleymani, 2019). For instance, Arandjelovic and Zisserman (2018) trained a network to learn audio and visual embeddings which enable cross-modal retrieval with Euclidean distance between two representations, but only used a single image as the input. And Hong, Im, and Yang (2017) adopted a similar technique with more frames as input and the triplet loss for the joint embedding of visual and audio, too. However, it extracted visual features by calculating first- and second-order statistics (i.e., mean and variance). In particular, triplet loss could map different instances to a shared Euclidean space where distances directly correspond to VAC by making positive pairs attracted and negative pairs separated (Schroff et al., 2015). A major advantage of this loss function is that we could capture the interaction between visual and audio stimuli by building an appropriate training set, and VAC would be accordingly indicated by the distance in latent space.

3. VAC measurement

3.1. Background

The dataset of videos was collected from Bilibili. Bilibili was listed on the NASDAQ on March 28,2018,[1] and listed for the second time in Hong Kong.[2] It is seen as a leader and also a classic one among user-generated streaming video platforms in China. Bilibili also has a large user base, the primary user community is young people, who always get their finger on the pulse.[3] According to the report from QuestMobile, Bilibili is in the top five hottest streaming video platforms in 2021 based on Daily Active User (DAU).[4] This brings a great number of various videos to Bilibili. The primary functions Bilibili provides are similar to YouTube except for the 'coin', which is popular on many other platforms in Asia (e.g., Douyu and AcFun) but has been rarely visited in previous studies. Fig. 1 illustrates the interface of Bilibili. By tipping a video with coins, viewers express a deeper appreciation than upvoting (hitting the like button). Considering that the channel of vlogs becomes the hotspot of Bilibili and according to official statistics, it has now developed into one of the most active channels in 2021,[5] we collected videos from this channel. In particular, for vlogs, uploaders produce videos on their own and often need to add background music for their daily life images to enhance their aesthetic values. In line with this, vlogs' VAC shall be more various. As for videos from other channels, such as music channel, visual streams are deliberately fabricated according to the background music. Specifically, in this study, we collected 2164 stream videos, i.e., vlogs and their relevant attributes. However, it is worth mentioning that the performance of the proposed neural network here is still stabilizing on videos of other channels and platforms.

3.2. Generating positive and negative samples

Geng, Chen, Lam, and Zheng (2013) suggested that over short retention, people always construct their retrospective hedonic evaluation on the basis of peak and effects over a short retention interval (peak-end rules). As for streaming video, viewers could upvote or tip at any time during their watching. According to this, we suggest that the video clips with more active users could represent the whole video better and could affect viewers' behaviors such as tipping more than other parts. Bilibili offers a live chat function that enables viewers to comment while watching the video, and the number of live-comments intuitively indicates the number of active viewers (Wang et al., 2020). In this case, we select video clips from the whole video based on the number of live-comments and we are not the first study to extract representative and memorable segments based on the distribution of live-comments (Xian, Li, Zhang, & Liao, 2015). The more the video attracts viewers, the more comments are generated (Wang et al., 2020). In meanwhile, inspired by the peak-end rules (Geng et al., 2013), these representative clips are more likely to be recalled even after watching the whole video, implying that no matter what time the viewers click the like, they are the essential triggers. We thus select the only clip with the most live-comments from each video for the dataset used to train and validate the presented model (train dataset and validation dataset). And more clips for one video are selected for the dataset used to measure the effect of VAC (test dataset), which is elaborated in Section 3.3.3.

We treat the audio and visual parts of the 8-seconds clips we selected as positive pairs. To form negative pairs, we still use the audio stream of positive pairs and replace the visual part with the visual stream from the clip 8 s later. The whole procedure is illustrated in Fig. 2. The visual stream in the later 8 s clip is extracted as v_n, which is further grouped with audio stream a from the positive clip to form the negative pair. In fact, the negative sample we select is more difficult to be distinguished from

[1] https://baijiahao.baidu.com/s?id=1596189027534330826
[2] https://finance.sina.com.cn/tech/2021-03-29/doc-ikkntian0327465.shtml
[3] https://www/questmobile.com.cn/research/report-new/31
[4] https://www/questmobile.com.cn/research/report-new/222
[5] https://www/questmobile.com.cn/research/report-new/222

Fig. 1. Bilibili's interface.

the positive one than that selecting visual stream v_n from another randomly targeted video to compose the negative pair since the video stream appears very close to the audio stream (only 8 s earlier) and both streams inherently belong to the same video. According to Schroff et al. (2015), it is crucial to select hard triplets (the triplets which are hard to be differentiated) when training networks with triplet loss, because it could actually contribute to improving the model. The data set can be accordingly denoted as $D = \left\{ \left(v^{(1)}, v_n^{(1)}, a^{(1)} \right), \left(v^{(2)}, v_n^{(2)}, a^{(2)} \right), \ldots, \left(v^{(m)}, v_n^{(m)}, a^{(m)} \right) \right\}$, where $a^{(i)}$ denotes the audio sample, while $v^{(i)}$ and $v_n^{(i)}$ denotes two visual clips, which are a sequence of RGB frames, and $\left(v^{(i)}, v_n^{(i)}, a^{(i)} \right)$ makes up a triplet. More specifically, $v^{(i)}$ is the visual clip that synchronized with $a^{(i)}$, however, $v_n^{(i)}$ is a visual clip taken from the same video but 8 s later than $v^{(i)}$, i.e., to compose a negative pair with $a^{(i)}$ in terms of unsynchronized visual signals.

We use the length of 8 s considering that too short a duration would make it difficult to convey the information and too long would increase the computational overhead. And many other experiments similarly set the length of 8 s to measure video quality and other attributes (Scholler et al., 2012; Zheng, Zhang, Lv, & Yang, 2018), so it is long enough for our task, and the length of 4 s is too short. In addition, as for neural networks, the length of 8 s is also commonly used (Fan, Su, & Huang, 2022; Kitaguchi et al., 2021). As the visual sub-network encourages a 16 frames input (Tran et al., 2018), the length of 8 s is a good choice. We select 2 frames per second to capture the changing of pictures. In that case, the length of 32 s is too long to feed in. To confirm the stability of the proposed measurement, we also changed the length of the video clips into 4s and 32 s respectively to train the neural network. The result and more details are in Appendix C.

The way we build the positive and negative sample is similar to the AVC task (Arandjelovic & Zisserman, 2017), which aims to train the neural network without labels. To be specific, the correspondence between visual and audio streams is available while they appear together at the same time in the same video. So that no supervision is required to manually label the correspondence which generally results in intense labor expense, time cost, and even subjective bias. Arandjelovic and Zisserman (2018) also proposed a network with two input streams to reveal the location of the object that makes the sound. These efforts assure that these sorts of positive and negative pairs could effectively represent and particularly distinguish the correlation between visual and audio streams.

3.3. Measuring VAC

3.3.1. Data processing

The input to the audio sub-network is an 8-second sound clip converted into normalized waveform frames with mono 16 kHz samples and 0.96 s sliding windows. It is converted into a log Mel spectrogram based on the request of VGGish. Mel spectrogram

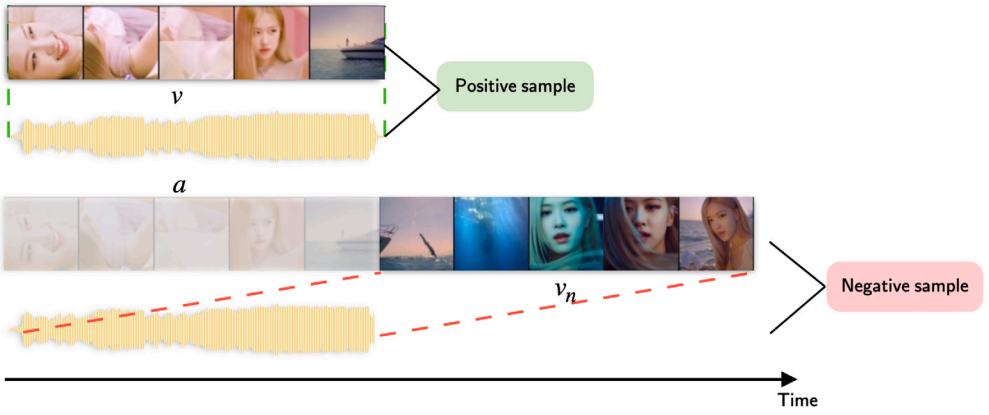

Fig. 2. Example of positive and negative sample.

has been shown to be better than other audio features like Short Time Fourier Transform (STFT) and Mel Cepstrum Coefficient (MFCC) (Choi, Fazekas, Sandler, & Cho, 2017; Murauer & Specht, 2018). And this kind of audio feature could be further extracted effectively by CNN (Hershey et al., 2017). The audio signal resampled to 16 kHz mono is transformed into a spectrogram with a window size of 25 ms and window hop of 10 ms. 64 Mel bins are used in mapping the spectrogram to the Mel spectrogram, and finally, get the stabilized log Mel spectrogram by computing the logarithm (Mel spectrogram+0.01). So the audio is represented as 96×64 for each second, and the dimension of input would be $8 \times 96 \times 64$.

The input to the visual sub-network is a set of video frames, with a resolution of 112×112 and a frame rate of 2 fps. We picked 2 frames per second to cover temporal information as long as possible. As the length of each video clip is 8 s, the sub-network ingests 16 stacked RGB frames at once, the dimension of which is $16 \times 112 \times 112 \times 3 (T \times H \times W \times 3)$, where T is the number of frames in the clip, H and W are the frame height and width, and 3 refers to the RGB channels.

3.3.2. Network architecture

Our core idea is to build a latent space where the distance between visual and audio vectors could indicate the VAC between them. To tackle this task, we propose the network structure shown in Fig. 3. As mentioned above, we use a two-stream structure to extract features from different modalities. It can be further divided into three distinct parts: the audio and visual sub-networks which extract audio and visual features respectively, and the fusion layers which map audio and visual features into a shared space. Considering the target of the presented neural network model is to measure visual-audio correspondence automatically, we named it VAC-Net. And the python code and trained parameters can be publicly available through https://github.com/LBluu/VAC-Net.

We adopt and fine-tuned the VGGish, which was pre-trained on YouTube-8M, in the audio-sub network to extract features from the input audio. Since VGGish was trained by a large number of videos, we argue that this pre-trained network is more suitable for our research than others which are trained by audio clips only. In fact, we also compared VGGSound (Chen, Xie, Vedaldi, & Zisserman, 2020) with VGGish and found that VGGish performs better in accuracy. VGGish is a variant of the VGG model, which contains four groups of convolution and maxpool layers and three fully connected layers. Because the size of the convolution kernel is 3×3 and the convolution stride and padding size are both fixed to 1, the size of each audio frame will be preserved after convolution. But its length and width will be cut in half when passing through maxpool layers performed over a 2×2 window with stride 2. So each audio frame ($1 \times 96 \times 64$) first passes through four groups of convolution and maxpool layers, and its size turns to be $512 \times 6 \times 4$. And then, it is fed into fully connected layers, and finally, transformed into 1×128 as the end of VGGish. Then we concatenated the audio features of all eight frames into a 1024-D vector and applied two fully connected layers before joint embedding to get a 256-D output of the audio representation. To further increase the pre-trained model's performance in our task, we fine-tuned the sub-network by unfreezing the four fully connected layers to re-train their weights.

With respect to the visual sub-network, we utilized a pre-trained 18-layer R(2+1)D model for extracting visual features, which captures temporal correlations between consecutive frames through (2+1)D convolutional blocks. By decomposing 3D convolution, (2+1)D contains more nonlinearities rendering the model to fit more complex functions. The 18-layer R(2+1)D model could be seen as a sequence of convolutional residual blocks followed by a fully connected layer with softmax. For instance, the output of the ith residual block x_i could be described as

$$x_i = x_{i-1} + \mathcal{F}(x_{i-1}, W_i), \tag{1}$$

where $\mathcal{F}$ implements the composition of two convolutions with parameters W_i and the application of the ReLU function. As Fig. 3 shows, the visual input passes through the first convolution with kernel size $3 \times 7 \times 7$ and stride size of $1 \times 2 \times 2$ and then is fed into a sequence of convolutional residual blocks. The first block with a solid curve which is different from others because the output size of this block is the same as the input size, so x_0 adds to x_1 directly. However, for other blocks, the output size ($\frac{T}{2} \times \frac{H}{2} \times \frac{W}{2}$) is different from input size ($T \times H \times W$). Down-sampling is performed in these blocks to change the size of x_{i-1} which is illustrated

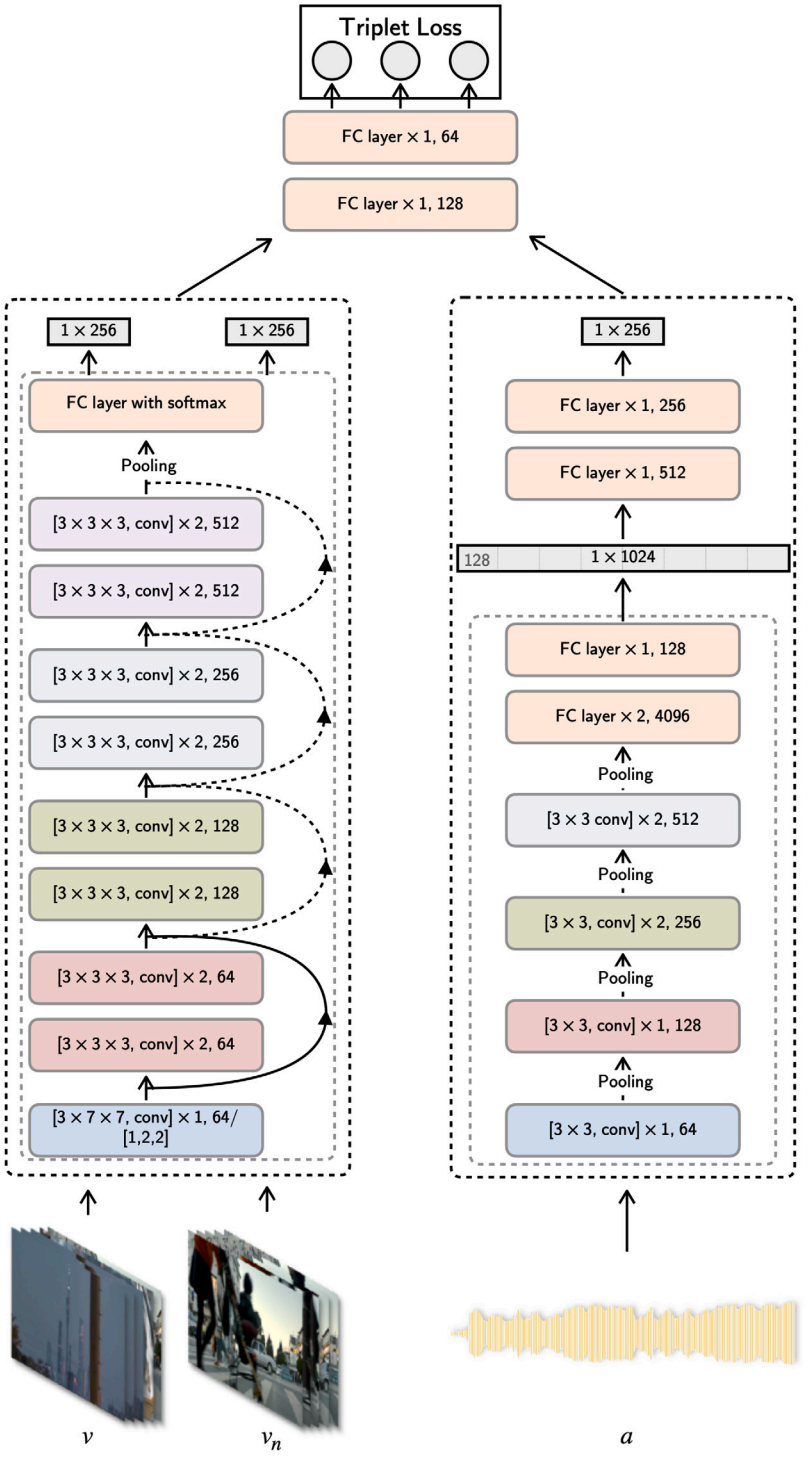

Fig. 3. The architecture of the VAC-Net.

as dotted arrows. The sub-network ends with a global spatiotemporal pooling layer and a fully-connected layer with softmax which finally yields a 256-D features vector. Similar to the audio sub-network, we unfroze the last residual block and the full-connected layer to re-train this pre-trained model.

The final step is to embed the separately extracted features of the audio and video modalities into a shared latent space. The fusion network of our model is consisted of two fully connect layers with ReLU between them. We feed a triplet into the network

each time to get two 256-D visual features and one 256-D audio feature produced by two sub-networks as the input of the fusion network. Then two fully connected layers map these three vectors into embeddings. Triplet loss trains the neural network model to make the embeddings of the positive pair close while the negative pair distant in the joint space (Schroff et al., 2015; Suris et al., 2018). And this criterion, which directly considers distance, is anticipated to ensure that the visual-audio correspondence could be continuously and effectively represented by distance. In addition, this loss function is also suitable for our unsupervised manner to differentiate negative and positive pairs grouped by a triplet. In this case, the distance in this latent space could represent VAC. The triplet loss is calculated through

$$L(a, v, v_n) = max\left\{d(a^{(i)}, v^{(i)}) - d(a_i^{(i)}, v_n^{(i)}) + \rho, 0\right\}, \tag{2}$$

where

$$d(x^{(i)}, y^{(i)}) = \|x_i - y_i\|_2, \tag{3}$$

and $a^{(i)}$ is the anchor of this triplet, $v^{(i)}$ is the positive sample which is corresponding to the anchor, the other $v_n^{(i)}$ is the negative visual sample, and ρ is a margin parameter. In our model, the negative pairs are not far from positive pairs. By using this loss function, we try to keep $d\left(v_n^{(i)}, a^{(i)}\right) > d\left(v^{(i)}, a^{(i)}\right)$. Accordingly, we can differentiate the negative sample from the positive one and get a meaningful latent space.

3.3.3. Training and result

To evaluate the performance of VAC-Net more credibly, we split our data into three parts, i.e, training dataset (train), validation dataset (val), and test dataset (test). The numbers of clips and videos for each class in the train/val/test partitions are shown in Appendix F. For the train dataset and the val dataset, as mentioned in Section 3.2, we only selected a single clip from each video, which has the most number of live-comments, to assure the quality of clips we targeted to better train and more credibly evaluate VAC-Net. While for the test dataset, the number of clips is in direct proportion to the length of each video, and the sum of the duration is approximately 1/4 of the length of the video. To be specific, we cut each video into 8 s clips as many as it could, and selected clips with the number of live-comments which is among the top 25%. In line with this, we totally selected 1943 clips from 504 videos. The test dataset is used to examine the effect of VAC, the videos of which were not involved the whole training process. In addition, we built a new dataset named 'expanded dataset', and further collected more videos without the constraint of only vlogs in both life channel (461) and music channel (452) from Bilibili. The method to select representative clips from videos in test dataset is again utilized to sample in total 3676 clips. We have also made URLs of all the videos utilized in this study publicly available, which can be accessed through https://github.com/LBluu/VAC-Net. All neural networks in experiments were trained using the Adam optimizer (Kingma & Ba, 2014) with triplet loss. The batch size was set to 10 clips per GPU and the training was generally finished in 50 epochs. The learning rate starts with 10^{-3} and was annealed according to the 1cycle policy (Smith & Topin, 2019). And then we took the val dataset to assess the performance of different networks and reported the top 1 accuracy. We calculated the Euclidean distance between positive pairs and negative pairs. If the distance between them conforms with $d\left(v_n^{(i)}, a^{(i)}\right) > d\left(v^{(i)}, a^{(i)}\right)$, VAC-Net distinguishes the triplet accurately. The results show that VAC-Net's test accuracy achieves 68.37% on the val dataset. We also compare the distance distributions of positive and negative visual-audio pairs before and after training (see Fig. 4). Because the normalized Euclidean distance monotonically decreases at the value of cosine similarity, and the Euclidean distance of 64-D vectors would be numerically large, we illustrate the comparison using cosine similarity instead, the value of which is always between -1 to 1. It can be seen that after training, the similarity between positive pairs with originally synchronized audio and visual streams is significantly smaller than that in negative pairs. The accuracy of VAC-Net on the test data set is 69.53%, and the comparison of distance distributions is also illustrated in Fig. 5. The accuracy of VAC-Net, though at first glance it may seem low, is based on a very difficult task. It is even challenging for humans. Owens and Efros (2018) has asked humans to identify the one with out-of-sync sound by providing a pair of shifted and aligned video and then calculated the accuracy. In this view, the job human subjects have to finish is very similar to VAC-Net's task, which is also trying to identify aligned or shifted video correctly. In particular, the shifted time for out-of-sync video and the length of videos in this experiment are both longer than ours, which indicates a more obvious difference and more significant temporal contexts. They found that even humans solved the task with only around 66.6% accuracy. Therefore from this perspective, the presented VAC-Net achieved an excellent performance that even outperforms humans on the similar task. Furthermore, the measurement of VAC-Net is consistent with the evaluation of human cognition based on the questionnaires' result (see Appendix A). In addition, we also use t-distributed stochastic neighbor embedding (t-SNE) to visualize high-dimensional visual and audio vectors (Hinton & Roweis, 2002) which is shown in Fig. 6. As can be seen, before training, visual and audio vectors are mapped to two separate groups, while they mingle with each other after training, implying that they are embedded into a shared space. Though we have limited our video content to vlogs in the life channel, the proposed VAC-Net is generally applicable and can be reliably extended to videos of other domains. Expanded dataset is used to verify this. The accuracy of the proposed VAC-Net in expanded dataset achieves 69.14%, suggesting its stable performance even on measuring VAC of videos from diverse domains. These results further ensure the credibility and reliability of the presented network in VAC measurement.

We compared our proposed model VAC-Net to a fused audio-visual network proposed by Owens and Efros (2018) as it was also trained in an unsupervised manner and considered visual streams with sequential information. In particular, the fused audio-visual network was trained on a larger dataset of approximately 750,000 videos, and obtained 59.9% accuracy to distinguish whether visual and audio streams are temporally synchronized. We evaluate this neural network (called Fused Model in Table 1) on the test dataset and expanded dataset in the present study. This model achieves similar accuracy with 62.23% in test dataset and 60.19%

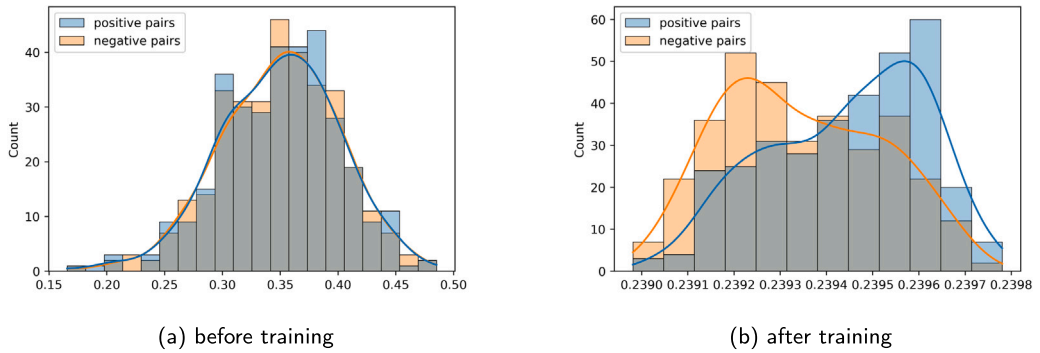

Fig. 4. Cosine similarity distribution for val dataset.

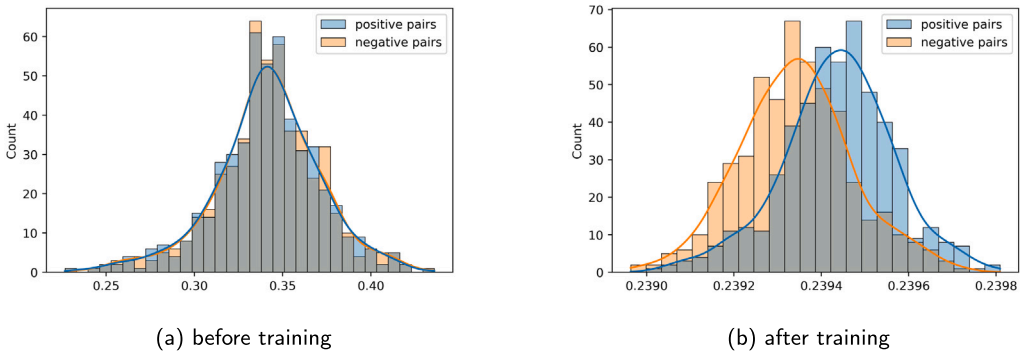

Fig. 5. Cosine similarity distribution for test dataset.

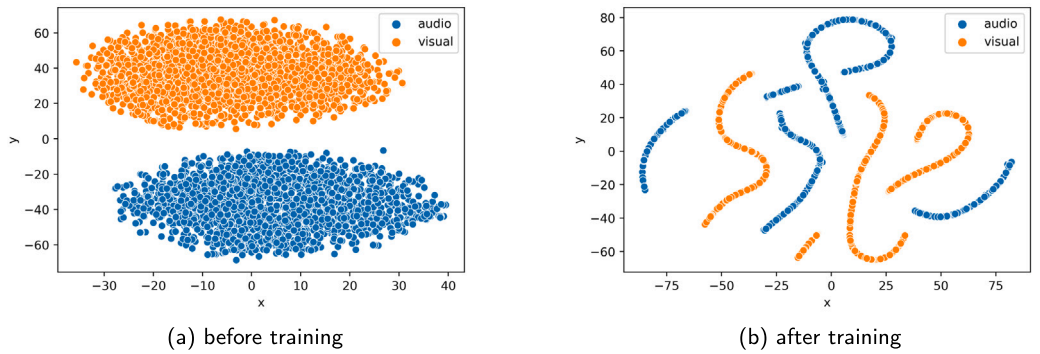

Fig. 6. t-SNE embedding for test data set.

in expanded dataset on our dataset on the classification task. We show the accuracy of the model with the same evaluation criteria with VAC-Net in Table 1, as can be seen, the Fused Model obtains lower accuracies than that of the proposed VAC-Net on both datasets.

In addition, we also replaced the audio sub-network into ResNet18 with the NetVLAD aggregation, which was pretrained on VGGSound dataset (Chen, Xie, et al., 2020). For extracting audio features, VGGish is aligned with this pretrained model to compose a competent baseline. And we trained it using an identical procedure and training data to our proposed VAC-Net. We again compared it (called Replaced Model in Table 1) with VAC-Net on test dataset and expanded dataset. As can be seen in Table 1, the baseline with ResNet18 obtains lower accuracies than the VAC-Net on both datasets. To sum up, the presented VAC-Net outperforms both baselines, implying its competence in VAC measurement.

Table 1
Accuracies of baselines and VAC-Net.

Method	test dataset	expanded dataset
Fused Model	.6782	.6852
Replaced Model	.6870	.6759
VAC-Net (proposed)	**.6953**	**.6914**

Table 2
Summary statistics of data.

	Mean	Standard deviation	Min	Max
Distance	0.00	1.00	−3.02	3.55
Coin	5.55	2.33	0.00	12.13
Like	7.09	1.95	2.20	13.04
View	10.44	1.84	5.41	15.85
Live-comment	4.58	2.27	0.00	10.68
Reply	4.67	1.60	0.00	9.31
Collect	5.84	2.11	0.00	11.55
Share	4.42	2.23	0.00	11.32
Length	5.83	1.10	2.46	8.86
Follower	9.98	2.90	1.39	14.88

Table 3
Correlation matrix after post processing.

	Distance	Coin	Like	View	Live-comment	Reply	Collect	Share	Length	Follower
Distance	1									
Coin	−0.107**	1								
Like	−0.105**	0.875***	1							
View	−0.0630	0.746***	0.895***	1						
Live-comment	−0.114**	0.858***	0.836***	0.787***	1					
Reply	−0.0460	0.832***	0.855***	0.810***	0.843***	1				
Collect	−0.100**	0.851***	0.853***	0.840***	0.762***	0.761***	1			
Share	−0.098**	0.835***	0.820***	0.801***	0.776***	0.768***	0.904***	1		
Length	−0.0240	0.156***	0.00100	−0.0170	0.259***	0.0600	0.087*	0.0620	1	
Follower	−0.141***	0.553***	0.551***	0.484***	0.596***	0.476***	0.416***	0.386***	0.185***	1

***p<0.01; **p<0.05; * $p < 0.1$.

4. Effect on tipping behavior

4.1. Data collection and processing

After successfully training the network, we measured VAC for each video in the test dataset, and collected other attributes of these videos, such as the numbers of being viewed, liked, replied, shared, and collected, to further examine the effect of VAC on tipping behavior. And the distribution of these indicators is highly skewed, so we perform a logarithm on them. The means, standard, deviations, maximum, and minimum of all variables are reported in Table 2, and Table 3 reports the correlation matrix of these variables. According to our hypotheses (H 1, H 2, and H 3), the variables that need to be measured include VAC, VS, TIP, and other control variables.

Recall that VAC in this study is defined as the correspondence between visual and audio streams in video and it is captured automatically by the deep neural network we established. In particular, the VAC of a video clip will be indicated by the similarity, e.g. distance, between visual and audio representations in the shared latent space that learned. Because when the distance increases, the VAC would decrease, the opposite of Euclidean distance is accordingly proxied to be a measure of VAC. The distance between 64-D vectors is pretty big and the hypothesis that the distribution of distance is drawn from the same distribution with normal distribution according to the K-S test ($p = 0.12 > 0.05$), so we normalized the distance by z-score normalization to achieve the same order with other variables. It is worth noting that if more than one clip are selected from a video, its VAC is simply the average over these clips.

Consumer satisfaction is usually defined as the post-usage judgments of satisfaction or the consumer's response to the evaluation of products. And for products with lengthy consumption periods (e.g., a vacation, college education, and a video), the satisfaction that would emerge before the service is finished is also called the interim judgment of satisfaction (Oliver, 2010). To be specific, viewer satisfaction refers to satisfaction with the quality of videos according to our theory basis. Oakes and North (2008) confirmed that the correspondence of multisensory could increase consumers' evaluation of the quality. And it is worth noting that the quality of videos takes video content into account. It is also inevitable to take video content into account when investigating multisensory correspondence because some parts of correspondence are based on multisensory content, such as genre congruency, semantic

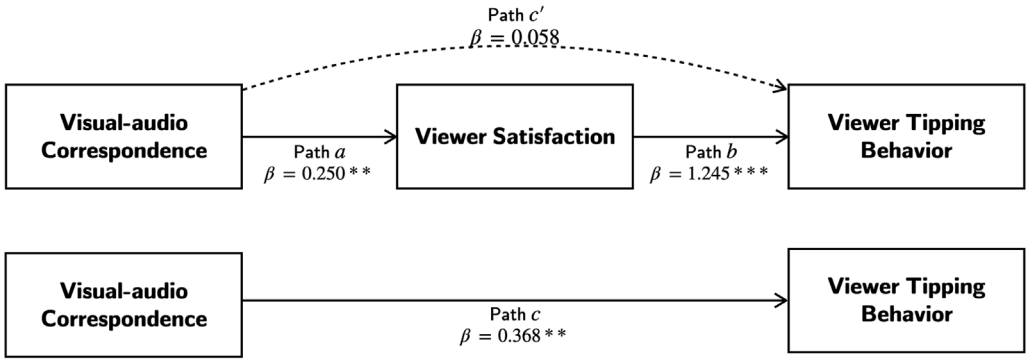

Fig. 7. Hypothesized mediation model.

congruency, and so on. And the 'quality' in previous studies indeed refers to an overall experience containing sensory content (Oakes & North, 2008). Similarly, according to the study about PWYW (Kim et al., 2009), consumer satisfaction is defined as consumers' post-consumption evaluation of the perceived quality. And it in PWYW is also not independent of content, which is a general evaluation closely related to price (Kim et al., 2009). For measurement, the methods we use are also consistent with the theory basis. The proposed neural network learns the representation of VAC in a general way with visual and audio streams directly fed in. And we use the number of like as the measurement of viewer satisfaction, which refers to viewers' overall evaluation of the video quality, in which the content is also considered inherently. As Fig. 1 shows, if viewers want to express their appreciation of the video, hitting the like button is a common reaction. Compared with other buttons, complimenting the video by replying needs more motivation than hitting the like button, and sharing indicates the video is worth circulating instead of being satisfactory. It cannot be denied that collecting videos reflects viewers' satisfaction. However, it comes with attached conditions that the video should be worth repeated viewing. So it indicates a conditional satisfaction, which is not consistent with our research purpose and would disturb our analysis. And we will not take the collecting number into account. In addition, Temme (1992) showed that there is no clear distinction between satisfaction and appreciation of art exhibits. Taken together, the number of likes a video received is used as a measure of VS and other attributes will be further tested to ensure the robustness of the effect from VAC.

Regarding the TIP, the number of coins is used to construct the measure. Though the coin is a kind of virtual currency, it is in circulation limited and could bring real revenue. Uploaders could earn coins from others by posting videos, and the number of coins the video received could influence the revenue sharing they obtain from Bilibili. And viewers only earn coins by staying active (e.g., logging in and sharing videos), so the number of coins is limited. Furthermore, the name of 'coin' and the process of paying coins all give users an experience of tipping. These characteristics make the coin a different definition with the like button, and the pricing strategy is consistent with PWYW. So we measure the viewer tipping amount in terms of the number of coins a video received.

Though it is impossible to control the content of the video, the theme of videos we collected are similar to each other because they belong to the same channel. As there are strong correlations among the variables related to video (> 0.80), which could be seen in Table 3, it is inappropriate to put them into the regression model together. We choose the number of views (View) as the control variable which is similar to Zhang, Wang, and Chen (2020). Lu et al. (2021) also proved that view times will influence users' tipping behavior, and it is more fundamental than other variables in representing the size of viewers. But we also attempt to introduce other attributes (such as the numbers of shares, collecting times, clarity, aesthetics and so on) into the regression model as control variables, which will be further illustrated in Section 4.3.

4.2. Mediation effects

According H 3, a mediation model is proposed to explore how VAC affects viewer tipping behavior as seen in Fig. 7. That is, VAC influences TIP in terms of VS. Baron and Kenny (1986) proposed four steps in establishing mediation models, which was broadly adopted. According to these four steps, we test hypotheses (H 1, H 2 and H 3) with the following regression models:

$$TIP_i = \beta_0 + \beta_1 \cdot VAC_i + \beta_2 \cdot View_i + \epsilon_i, \tag{I}$$

$$VS_i = \beta_0 + \beta_1 \cdot VAC_i + \beta_2 \cdot View_i + \epsilon_i, \tag{II}$$

$$TIP_i = \beta_0 + \beta_1 \cdot VS_i + \beta_2 \cdot View_i + \epsilon_i, \tag{III}$$

$$TIP_i = \beta_0 + \beta_1 \cdot VS_i + \beta_2 \cdot VAC_i + \beta_3 \cdot View_i + \epsilon_i. \tag{IV}$$

Specifically, in these models, i denotes a video simple, $\beta_1, \beta_2, \beta_3$ are coefficients of variables and ϵ_i is the residual of the regression. In Model I, TIP is regressed on VAC to reveal the effect of visual-audio correspondence on the tipping behavior, as Path c in Fig. 7.

Table 4
Results of regression models.

	Model I	Model II	Model III	Model IV
VAC	0.139**	0.0942**		0.0219
	(0.0691)	(0.0386)		(0.0501)
View	0.940***	0.945***	−0.235***	−0.233***
	(0.0376)	(0.0210)	(0.0606)	(0.0608)
VS			1.245***	1.242***
			(0.0572)	(0.0576)
_cons	−4.265***	−2.772***	−0.822**	−0.821**
	(0.398)	(0.223)	(0.328)	(0.328)
depvar	TIP	VS	TIP	TIP
R^2	0.561	0.804	0.772	0.772
F	319.6	1026.2	849.4	565.4
N	504	504	504	504

Standard errors in parentheses
***p<0.01; **p<0.05; * $p < 0.1$.

In Model II, VS is regressed on VAC to establish Path a in Fig. 7, that is, how visual-audio correspondence will reshape the viewer satisfaction. It is related to H 1. In Model III, VS is used to explain TIP, which is aimed to provide a test of Path b in Fig. 7, i.e., whether viewer satisfaction is related to video tips. And it is related to H 2. In Model IV, VAC is employed again to explain TIP while VS is controlled, as Path c' in Fig. 7.

If VAC is significant in Model II and VS is significant in Model III, it is possible that VAC influences TIP through VS (Frazier, Tix, & Barron, 2004; Kenny, Kashy, & Bolger, 1998). The significance of VAC in Model I is not necessary, but if so, it could strengthen our arguments. And if VS is significant while VAC is insignificant in Model IV, we can conclude that the mediation of VS is significant and complete. That is, VS influences TIP by affecting VS.

The regression results of these models can be found in Table 4. The Variance Inflation Factor (VIF) value obtained is not more than six and thus, there is no evidence of multicollinearity. Moreover, F-test was used to check the overall significance of models and confirmed that R^2 is highly significant for these models. This result further ensures that these four models are indeed significant. As can be seen in the table, VAC is statistically significant in estimating TIP (Model I, $\beta = 0.368, p = 0.045$). As we expected, the positive coefficient of VAC indicates that there is a positive relationship between VAC and TIP. As for the effect on VS, we also find that VAC is a significant explanatory variable (Model II, $\beta = 0.250, p = 0.015$) in explaining VS. Similarly, its positive coefficient also supports H 1. In Model III, VS is also statistically significant in estimating TIP (Model III, $\beta = 1.245, p = 0.000$). And the positive coefficient supports H 2. It is worth mentioning that, due to the high correlation between View and VS, the variable View's coefficients in Model III and IV change into a negative normally after introducing VS into the model. These results indicate that the increase in VAC would spur more viewer tipping and satisfaction. To be specific, when the number of views is the same, a video with better visual-audio correspondence could receive more likes and coins. To estimate the influence of measurement error VAC-Net induced on the regression model, we employed SIMEX to correct possible measurement errors (Peng, Cui, Chung, & Zheng, 2020; Yang, Adomavicius, Burtch, & Ren, 2018). The results suggest that though the accuracy of the proposed VAC-Net in automatically measuring VAC is not as high as generally expected, the effect of VAC on tipping behavior can still be credibly and sufficiently examined and profiled. The result could be seen in Appendix E.

To demonstrate that VS as a mediator for the effect from VAC on TIP, i.e., as supposed in H 3, the strength of the relation between VAC and TIP should be eliminated (indicates a complete mediator) or significantly decreased (indicates a partial mediator in the model). In particular, to compare path c' with path c in Fig. 7, according to Baron and Kenny (1986) and Kenny et al. (1998), the coefficients of these multiple regressions are worth exploring. Indirect effect (path a and path b), which is the measure of the amount of mediation, consists the total effect (path c) with direct effect (path c') together. As in this study, the coefficient of VAC in explaining TIP (path c', $\beta = 0.058$) is significantly smaller when the VS is controlled in Model IV than when VS is not controlled in Model I (path c, $\beta = 0.368$), and even close to zero. It suggests an almost complete mediation effect of VS, i.e., VS accounts for the positive effect of VAC on TIP. In addition, this mediation effect is more credible especially when VAC, which does not pass the t-test ($p = 0.663$) in Model IV, becomes insignificant when the mediating variable (VS) is introduced into the regression.

To further assess the significance of this mediated effect, Sobel test was conducted. There are three principal versions of the Sobel test (Aroian, 1947; Goodman, 1960; Sobel, 1982), we conducted all these three tests. The results in Table 5 show that the mediation effect of VS is significant ($p < 0.05$) and $VAF = 0.8429 > 20\%$. Therefore, H 3 is significantly and consistently supported, that is to say, VAC we have measured indirectly influences tips the video receives, and the mediator is viewer satisfaction. In addition, we also used bootstrapping to additionally check the mediation, which will be discussed in the test of robustness.

Table 5
Mediation test.

Sobel test			
	Test statistic	Std. Error	p-value
Sobel test	2.434	0.128	0.015
Aroian test	2.431	0.128	0.015
Goodman test	2.437	0.128	0.015

Table 6
VIF of robustness models.

	Model II'	Model III'	Model IV'
(Constant)			
VAC	1.033		1.041
Length	1.097	1.104	1.104
Share	3.535	3.811	3.823
Reply	3.44	4.163	4.199
View	4.21	**6.534**	**6.57**
Follower	1.498	1.717	1.73
Clarity	1.195	1.201	1.202
Shelf_time	1.257	1.464	1.469
Aesthetics	1.165	1.17	1.17
VS		**9.752**	**9.824**

4.3. Robustness check

Bootstrapping procedures were also applied to examine the mediation effect (Bollen & Stine, 1990). According to Preacher and Hayes (2008), the mediation effects are present when the 95% bootstrap confidence interval does not straddle a 0 between the upper and lower intervals. As for our result, the indirect effect is 0.3103, and 0 is not straddled in our confidence interval $(0.0335, 0.6021)$, which further supports H 3. The complete mediation effect of VS between VAC and TIP is therefore robust.

In addition, in Models I-IV we only introduced the number of view times as the control variable to avoid multicollinearity among explanatory variables. To verify our result we still attempt to introduce more indicators as the control variables. We control the uploaders' information using the number of followers the uploader has. It reflects the popularity of the uploader on the platform. And we control the video content by introducing other dimensions of videos such as length, the numbers of shares and replies. These attributes indirectly indicate the content of the video, such as if the topic is controversial. We also took other control variables, including video clarity, the shelf time, and aesthetics into account, which may affect the quality of videos and thus presumably contribute to the number of likes received. Streaming video platforms provide videos with several options including 360P, 480P, 720P, 1080P, and 4K or HDR. We collected the highest standard each video provides as its charity measurement. The shelf time of a video is proxied in terms of the number of days on the shelf up to October 29, 2022. As for aesthetics, SAMP-Net, provided by Zhang, Niu, and Zhang (2021), aims to assess the aesthetic score of a given image. We adopted the pre-trained model they provided to get the average aesthetic assessment for each video by calculating the aesthetic score of all frames we extracted from it. The result is shown in Table 7. Even after adding these control variables, VAC could still significantly explain VS as shown in Model II'. Serious multicollinearity is not existing in Model II' based on VIF test (see Table 6), but exists in Model III' and Model IV' as expected. Though standard errors of coefficients would be affected by multicollinearity, here the small standard errors of coefficients in the robustness regressions are resulted by the large sample size.

Besides, as a theoretical basis, Kenny et al. (1998) stated that the first step, i.e., Model I, is not necessary, and many situations could cause the absence of the relation between a predictor and an outcome. It is also suggested that the essential steps in establishing mediation are Steps 2 (Model II and Path *a*) and 3 (Model III and Path *b*) (Frazier et al., 2004). To introduce more control variables, we established Model II' based on Model II, Model III' based on Model III in Table 7 as the Step 2 and Step 3. We also built Model IV' to further test the robustness of the effect from VAC on TIP. According to the regression result shown in Table 7, we could find that VAC is still significant in explaining VS, and the coefficient is still positive, the situation of which is similar for VS in Model III. As expected, VAC becomes insignificant after introducing VS in Model IV'. In this view, when all these factors are fixed, the VAC could still positively affect VS, and then VS influences TIP.

5. Discussion

5.1. Research issues

The first object of our study is to measure the visual-audio correspondence in videos automatically. According to Arandjelovic and Zisserman (2017), we build a dataset and trained a cross-modal neural network in order to infer this correspondence. To be specific, we extract visual and audio features by using two sub-networks with the help of pre-trained models, then these audio and visual features are put into a shared layer to establish a joint embedding space the Euclidean distance within which could proxy

Table 7
Results of robustness models.

	Model II'	Model III'	Model IV'
VAC	0.0545*		−0.0128
	(0.0284)		(0.0377)
Length	−0.0000798*	0.000334***	0.000334***
	(0.0000464)	(0.0000614)	(0.0000615)
Share	0.150***	0.383***	0.384***
	(0.0236)	(0.0324)	(0.0324)
Reply	0.339***	0.333***	0.331***
	(0.0324)	(0.0471)	(0.0474)
View	0.519***	−0.497***	−0.498***
	(0.0312)	(0.0513)	(0.0515)
Follower	0.103***	0.0625***	0.0630***
	(0.0118)	(0.0167)	(0.0168)
Clarity	0.0619*	0.202***	0.202***
	(0.0368)	(0.0488)	(0.0488)
Shelf_time	−0.000779***	0.000579***	0.000581***
	(0.0000854)	(0.000122)	(0.000122)
Aesthetics	0.135	−0.0803	−0.0795
	(0.0940)	(0.124)	(0.124)
VS		0.796***	0.798***
		(0.0592)	(0.0595)
_cons	−1.599***	−0.167	−0.172
	(0.361)	(0.486)	(0.487)
depvar	VS	TIP	TIP
R^2	0.898	0.876	0.876
F	484.4	387.3	348.0
N	504	504	504

Standard errors in parentheses
***p<0.01; **p<0.05; * $p < 0.1$.

VAC after training the model with the triplet loss function. The effectiveness of our joint embedding network is demonstrated from different perspectives, VAC-Net not only does well in accuracy but also reveals the difference between positive and negative pairs (see Fig. 5). The accuracy it achieved is higher than two baselines and even better than humans on the similar task. In addition, VAC measurement obtained by the proposed VAC-Net conforms to human cognition (see Appendix A). So we defined the VAC as the opposite of the Euclidean distance between the visual vector and audio vector in the joint embedding space built by the neural network.

The second object of our study is to examine the effect of visual-audio correspondence on video tips. According to the theory about consumer experiences, sensory stimuli should bring a change of satisfaction. The result of our regression models further supports this point of view by revealing that visual-audio correspondence could improve viewer satisfaction. Furthermore, we suggest that the price strategy of Bilibili and other similar streaming video platforms is PWYW, which leads to the proposition that viewer satisfaction towards streaming video, i.e., the product, has a positive effect on tipping behavior. Integrating these theories, we posit that the effect of visual-audio correspondence on tipping behavior is mediated by satisfaction. We establish four regression models to test the mediation effect based on four multiple linear regressions and all the results consistently indicate that the mediation effect does exist. In terms of introducing bootstrapping and other control variables, the robustness of this mediation effect is further assured. So finally we demonstrate that visual-audio correspondence of video could stimulate viewer tipping behavior by increasing viewer satisfaction.

5.2. Theoretical implications

Our research not only contributes to both PWYW and streaming video literature but also fills the gap by linking tipping the video and VAC. Firstly, this study proposed an attribute VAC for streaming video to evaluate video quality which is measured automatically by a presented deep neural network VAC-Net. With the growing popularity of streaming and live-streaming, Youtube, Tiktok, Twitch, and their variants have dominated the Internet traffic globally in recent days and resulted in various business models and scenes. However, as mentioned before, the perception of VAC could be subconscious (Evans & Treisman, 2010; Spence, 2011). In addition, it is even a difficult task for humans to identify different visual-audio correspondence (Owens & Efros, 2018). Previous studies investigated viewers' behavior depending heavily on questionnaires (Alhabash, Baek, Cunningham, & Hagerstrom, 2015). The proposed VAC-Net, however, provides new opportunities to establish novel indicators of cross-modals in scenarios with

multimedia data. In terms of building two sub-networks, the proposed framework can respectively represent visual and audio streams into vectors within a shared latent space and simultaneously capture the interaction between them. In particular, audio semantics that is carried in its representations are closely associated with those that are delivered in visual representations. And supplementary experimental questionnaires (see Appendix A) also prove that the VAC obtained by VAC-Net conforms to human perception. Therefore, the presented framework could be extensively employed to construct and measure indicators of cross-modals in those streaming scenarios and inspire interesting theoretical explorations.

Secondly, The positive effect of visual-audio correspondence on tipping behavior essentially updates the extant understanding of consumer behavior in streaming videos. Previous literature noticed the importance of visual-audio correspondence in offline sales (Becker-Olsen, 2006). As for studies about online videos, previous studies only devoted efforts to visual or audio streams of the video separately and ignored the effect on the viewer as a consumer to investigate related behaviors such as tipping (Herget, 2021). To our best knowledge, visual-audio correspondence of videos has rarely been noticed before. However, visual-audio correspondence of video acts as an indispensable product attribute, which is closer to service-providing party (Becker-Olsen, 2006; Lalwani et al., 2009). And the impact of VAC is unknown. In the contrast, our findings offer a new perspective in terms of capturing the cross-modal interaction and correlating visual-audio correspondence with consumer behavior. To be specific, we confirmed the positive effect of visual-audio correspondence of streaming video on viewer tipping behavior and revealed the mediation role of viewer satisfaction between them. Empirical evidence shows that visual-audio correspondence could significantly boost the amount of tipping by enhancing viewer satisfaction. In this study, we identified and supplemented a new factor that influences viewer experience with videos. And it is suggested that even for digital products such as streaming videos, key consumer behaviors could be profoundly influenced by product-induced determinants, which cannot be simply overlooked.

The theory of PWYW is extended into the scene of streaming videos and is justified to be equally effective. In the meantime, according to PWYW, satisfaction will influence the price buyers pay. And we empirically justified this in streaming videos by demonstrating that viewer satisfaction with the video does increase the number of tips it received. It implies that even in the new scene of streaming videos, consumer satisfaction is still an important determinant in buyer pricing. In particular, the satisfaction examined here is aroused by digital products, i.e., videos and it thus supplements the PWYW further with a price factor induced by the product itself. And we also found that visual-audio correspondence, a specific attribute of the video, can positively affect viewer satisfaction, implying that critical parts of products can be targeted and leveraged with the help of PWYW.

5.3. Managerial implications

Visual-audio correspondence and its positive effect on viewer satisfaction can be instructive for streaming platforms in profitable activities such as recommendation or marketing. Screening out videos that will be popular in the future is fundamental in the business model of streaming platforms, which helps spur user activation and retention in the ecosystem. However, extant solutions might rely heavily on attributes of likes, views, or shares at the early stage and encounter the notorious issue of cold start, e.g., there are no input values of these attributes for videos just published. While with the help of the presented deep neural network, visual-audio correspondence can be directly measured from the video itself and accordingly avoid the cold start in the recommendation by offering a key input. Even more important, the visual-audio correspondence is related to the revenue of the video because higher visual-audio correspondence will result in more viewer satisfaction, which increases viewer tips. In this case, visual-audio correspondence could further be indicative in selecting and ranking videos.

The automatic inference of visual-audio correspondence would also equip uploaders with a new tool to enhance the monetization of videos. The positive effect of visual-audio correspondence on tipping behavior suggests that the income of uploaders can be improved significantly by optimizing the congruency between visual and audio streams of the video. Specifically, the proposed neural network can be employed to generate the visual representation of a video clip and audio representations of candidate audio clips; then the audio clip with the highest similarity, e.g., the smallest Euclidean distance, to the visual representation, could be selected to be the background music. By increasing the tips received by these generated videos, this automatic manner will particularly help ordinary uploaders, who occupy a dominant part in producing content for the streaming ecosystem.

5.4. Limitations

This study inevitably has limitations, which suggest promising avenues for future research. One of them is that we only sampled videos from the channel of vlogs in Bilibili. Though Bilibili is one of the most popular streaming platforms globally and vlogs channel is its hotspot, the evidence that supports the extension of our findings to other platforms is still limited. And additional evaluations from videos of other channels such as music and other platforms such as Youtube (see Appendix B) already demonstrate the generalization capability of the presented VAC-Net. In addition, the preliminary regression model result of videos from other channel indicate that VAC is still positively associated with the number of tips it received (see Appendix D). While whether the effect of VAC to tip still exists does need further efforts. In addition, because of the restriction on data access due to privacy, we only collected video-level data and were incapable of taking individual differences into account. Though many previous results implied the individual factors' effect on consumer behavior (Kunter, 2015), we failed to get more viewer attributes. What is more, viewers' satisfaction with the whole video is measured by the number of likes the video received, which could act as viewers' honest feedback. As mentioned before, one-quarter of the 8-seconds clips with more live-comments of the video are selected to represent the whole video. Though according to the previous studies (Wang et al., 2020) and peak-end rules (Geng et al., 2013), video clips with more live-comments are more memorable and representative, there is still no direct evidence to confirm that each viewer's

Table A.1
Questions of questionnaire.

questionnaire for comparing VAC	
VAC under comparison	Compared with the first video, I think the visual and audio of the second video corresponds with each other better
	Compared with the first video, I think the audio of the second video is more contributive to watching the visual stream.
	Compared with the first video, I think the visual transition of the second video is more correspondent to the audio.
	Compared with the first video, I think the audio of the second video makes visual content more vivid.
	Compared with the first video, I think the visual content of the second video is more complementary with the audio.

'like' behavior could be covered by these clips due to the complexity of the realistic scene and the lacking investigation of viewers' 'like' behavior. Last but not least, visual-audio correspondence in this study was defined and measured on the overall level without specific semantic meanings, like what kind of visual signal is drastic or which audio representation denotes joy, since the deep neural network is a black box. And the correspondence could be seen as the aggregation of various congruency. These limitations would inspire follow-up works to explore a specific congruency between visual and audio stimuli in other platforms at the individual level of viewers.

CRediT authorship contribution statement

Bu Li: Methodology, Software, Validation, Writing – original draft. **Jichang Zhao:** Conceptualization, Writing – review & editing, Supervision.

Data availability

Data will be made available on request

Acknowledgments

This work was supported by NSFC, China (Grant No. 71871006).

Appendix A. Supplementary experimental questionnaires

In the supplementary study, our objective is to validate whether the VAC measured based on VAC-Net has the same tendency as human cognition. We crafted 20 pairs of video clips with different VAC and selected them from the same video or clips with similar content such as traveling, eating, and so on. The questionnaire contains five questions which have the options of six-point Likert scale for subjects to compare the VAC of two video clips. The questions are shown in Table A.1. For each pair of video clips, we ask the subject to watch two 8 s clips and then answer the questions to get the VAC evaluation of human cognition. The example of the questionnaire is shown in Fig. A1.

We conducted this study through www.wjx.cn, a popular academic service in China, e.g. Zhang, Wang, and Wu (2021). In the end, 819 subjects joined this study, and each subject was randomly shown one of the pairs to answer the questions. That is to say, each pair of video clips have around 40 subjects on average. The reliability of the questionnaire is tested, and Cronbach's Alpha values for all twenty pairs are greater than 0.6, indicating that the questionnaires have acceptable reliability. We provide one of them as an example shown in Table A.2. And the alpha coefficient for the four items is 0.942, according to high internal consistency. In addition, the value of Cronbach's Alpha if an item is deleted from five questions is all smaller than Cronbach's Alpha based on standardized items, which means every item contributes to the total reliability. The score of each pair of video clips is shown in Fig. A2. The tendency is consistent with VAC-Net measured VAC if the score is larger than 3.5. As can be seen, the average perception score of each pair is higher than 3.5. The consistency of these samples confirms that our VAC scores conform to the cognition of viewers.

Fig. A1. An example of the questionnaire.

Table A.2
Example of reliability analysis.

Items	Cronbach's Alpha if Item Deleted	Cronbach's Alpha Based on Standardized Items
1	0.928	
2	0.932	
3	0.931	0.942
4	0.923	
5	0.918	

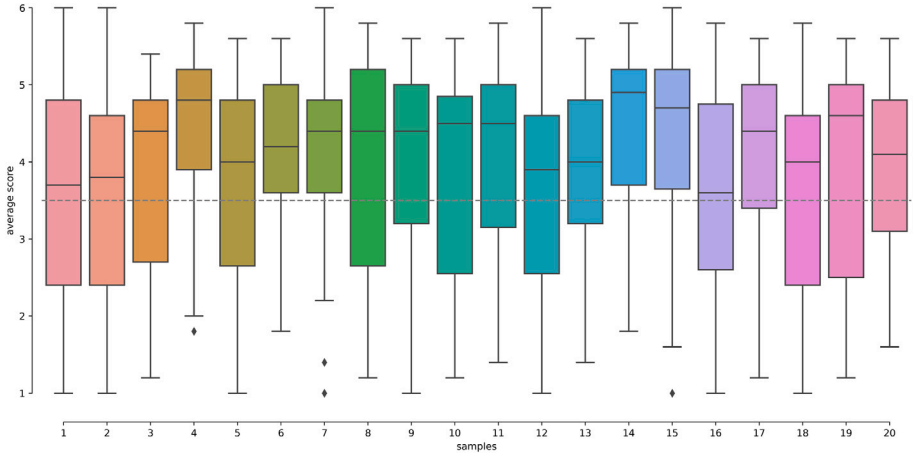

Fig. A2. Experimental results of questionnaires.

Appendix B. Performance on VGGSound dataset

We also collected top 1500 videos in VGGSound dataset the source of which is from another streaming video platform, YouTube. Because of invalid URLs contained in the csv file, we finally got 1491 videos. Due to the shorter video clips in this dataset (only 10 s), we could only displace the visual and audio streams by 2 s in generating both positive and negative samples. Curtailing the length of out-sync in fact leads the task of distinguishing the positive from the negative more difficult. However, VAC-Net still shows consistent performance, and achieves an accuracy of 67.00% in VGGSound dataset, suggesting that the VAC-Net is capable of extending videos of other domains.

Appendix C. Performance after changing the length of clips

To verify the stability of the presented neural network structure, we also tuned the length of video clips in our dataset. To be specific, we reset the length of video clips into 4s and 32 s, respectively, and trained the presented VAC-Net using an identical procedure to generate positive and negative pairs. And VAC-Net achieved an accuracy of 69.28% on val dataset with 4 s clips; and achieved an accuracy of 73.49% on val dataset with 32 s clips. Both accuracies are of the same magnitude as that on val data of 8 s clips, indicating the competency of VAC-Net is independable to clip lengths and its generalization capability is further assured.

Appendix D. Regression on videos of music channel

We selected videos of music channel from expanded dataset and the effects of VAC on TIP and VS are respectively illustrated in Table D.1. As can be seen in Table D.1, as expected, VAC of a video is still positively associated with the number of tips it receives, indicating that the primary finding of the present study can be extended to videos of other domains. In addition, putting VS into the regression, the effect of VAC on TIP is still significant. While it should also be noted that the positive association between VAC and VS is not significant in this channel. In fact, it is not that surprising. Because videos from different channels arouse different expectations. In particular, videos in music channel should have high VAC because their visual streams are deliberately fabricated according to the background music, which will result in persistently high VAC. As these videos are expected to possess high visual-audio correspondence, high VAC could not boost viewers' satisfaction towards them is reasonable. And VAC could still significantly explain TIP.

Appendix E. SIMEX correction result

The SIMEX method has been extensively employed to correct possible measurement errors induced by machine learning models (Peng et al., 2020). So we assess the measurement errors of distance based on the identification error in the presented neural network. And we use SIMEX method to estimate the influence of measurement error VAC-Net induced on the regression model. The results are shown in Table E.1. As can be found, after the SIMEX correction, the coefficient estimate of VAC shows little change compared to the original results. These results suggest that the result of our regression models about effect of VAC on tipping behavior is still credibly though the accuracy of the VAC-Net are not as high as generally expected.

Appendix F. Dataset statistics

Table F.1 shows the exact number of each dataset. The number of expanded dataset is showed as the number of videos in life channel plus the number of videos in music channel.

Table D.1
Results on videos of music channel.

	Model I	Model II	Model III
VAC	0.167**	0.0502	0.110**
	(3.16)	(1.49)	(3.00)
View	0.957***	0.976***	−0.143**
	(37.67)	(60.14)	(−2.68)
VS			1.127***
			(21.94)
_cons	−4.959***	−3.478***	−1.038***
	(−20.43)	(−22.45)	(−4.22)
depvar	TIP	VS	TIP
R^2	0.762	0.890	0.885
F	720.5	1816.1	1154.7
N	452	452	452

t statistics in parentheses
***p<0.001; **p<0.01; * $p < 0.05$.

Table E.1
SIMEX correction results.

	Model I	Model II	Model III	Model IV	Model I (SIMEX)	Model II (SIMEX)
VAC	0.139**	0.0942**		0.0219		
	(0.0691)	(0.0386)		(0.0501)		
View	0.940***	0.945***	−0.235***	−0.233***	0.940***	0.945***
	(0.0376)	(0.0210)	(0.0606)	(0.0608)	(0.0362)	(0.0198)
VS			1.245***	1.242***		
			(0.0572)	(0.0576)		
corrected_VAC					0.139*	0.0942**
					(0.0825)	(0.0415)
_cons	−4.265***	−2.772***	−0.822**	−0.821**	−4.265***	−2.772***
	(0.398)	(0.223)	(0.328)	(0.328)	(0.370)	(0.206)
depvar	TIP	VS	TIP	TIP	TIP	VS
R^2	0.561	0.804	0.772	0.772		
F	319.6	1026.2	849.4	565.4	337.4	1136.7
N	504	504	504	504	504	504

Standard errors in parentheses
***p<0.01; **p<0.05; * $p < 0.1$.

Table F.1
Dataset statistics.

	Train	Val	Test	Expanded
Clips	1328	332	1943	3676
Videos	1328	332	504	461 +452

References

Alhabash, S., Baek, J. h., Cunningham, C., & Hagerstrom, A. (2015). To comment or not to comment?: How virality, arousal level, and commenting behavior on YouTube videos affect civic behavioral intentions. *Computers in Human Behavior*, *51*, 520–531.

Arandjelovic, R., & Zisserman, A. (2017). Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision* (pp. 609–617).

Arandjelovic, R., & Zisserman, A. (2018). Objects that sound. In *Proceedings of the European conference on computer vision* (pp. 435–451).

Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics*, *18*(2), 265–271.

Aytar, Y., Vondrick, C., & Torralba, A. (2016). SoundNet: Learning sound representations from unlabeled video. In *Proceedings of the 30th international conference on neural information processing systems* (pp. 892–900). Red Hook, NY, USA: Curran Associates Inc.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Becker-Olsen, K. (2006). Music-visual congruency and its impact on two-sided message recall. *NA - Advances in Consumer Research*, *33*, 578–579.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, *87*(2), 241–247.

Bolivar, V. J., Cohen, A. J., & Fentress, J. C. (1994). Semantic and formal congruency in music and motion pictures: Effects on the interpretation of visual action. *Psychomusicology: A Journal of Research in Music Cognition*, *13*(1–2), 28–59.

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115–140.

Brengman, M., Willems, K., & De Gauquier, L. (2022). Customer engagement in multi-sensory virtual reality advertising: The effect of sound and scent congruence. *Frontiers in Psychology*, *13*, Article 747456. http://dx.doi.org/10.3389/fpsyg.2022.747456.

Chen, Y., Huang, A. X., Faber, I., Makransky, G., & Perez-Cueto, F. J. A. (2020). Assessing the influence of visual-taste congruency on perceived sweetness and product liking in immersive VR. *Foods*, *9*(4), 465. http://dx.doi.org/10.3390/foods9040465, URL https://www.mdpi.com/2304-8158/9/4/465.

Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020). Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 721–725). http://dx.doi.org/10.1109/ICASSP40776.2020.9053174.

Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 2392–2396). http://dx.doi.org/10.1109/ICASSP.2017.7952585.

Chung, S. W., Chung, J. S., & Kang, H. G. (2019). Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing* (pp. 3965–3969). http://dx.doi.org/10.1109/ICASSP.2019.8682524.

Chung, J. S., & Zisserman, A. (2017). Out of time: Automated lip sync in the wild. In *Computer vision – ACCV 2016 workshops* (pp. 251–263).

Demoulin, N. T. (2011). Music congruency in a service setting: The mediating role of emotional and cognitive responses. *Journal of Retailing and Consumer Services*, *18*(1), 10–18.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 6. http://dx.doi.org/10.1167/10.1.6.

Fan, W., Su, Y., & Huang, Y. (2022). ConchShell: A generative adversarial networks that turns pictures into piano music. http://dx.doi.org/10.48550/ARXIV.2210.05076, URL https://arxiv.org/abs/2210.05076.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, *51*(1), 115–134.

Geng, X., Chen, Z., Lam, W., & Zheng, Q. (2013). Hedonic evaluation over short and long retention intervals: The mechanism of the peak–end rule. *Journal of Behavioral Decision Making*, *26*(3), 225–236.

Gentile, C., Spiller, N., & Noci, G. (2007). How to sustain the customer experience: An overview of experience components that co-create value with the customer. *European Management Journal*, *25*(5), 395–410.

Gneezy, A., Gneezy, U., Riener, G., & Nelson, L. D. (2012). Pay-what-you-want, identity, and self-signaling in markets. *Proceedings of the National Academy of Sciences*, *109*(19), 7236–7240.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, *55*(292), 708–713.

Gregory, R. L., & Heard, P. (1979). Border locking and the Café wall illusion. *Perception*, *8*(4), 365–380.

Haber, R. N., & Hershenson, M. (1973). *The psychology of visual perception*. Holt, Rinehart & Winston.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Herget, A. K. (2021). Well-known and unknown music as an emotionalizing carrier of meaning in film. *Media Psychology*, *24*(3), 385–412.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 131–135). http://dx.doi.org/10.1109/ICASSP.2017.7952132.

Hinton, G., & Roweis, S. (2002). Stochastic neighbor embedding. In *Proceedings of the 15th international conference on neural information processing systems* (pp. 857–864). Cambridge, MA, USA: MIT Press.

Hoffer, E., & Ailon, N. (2015). Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition, vol. 9370* (pp. 84–92).

Hong, S., Im, W., & Yang, H. S. (2017). Content-based video-music retrieval using soft intra-modal structure constraint. http://dx.doi.org/10.48550/ARXIV.1704.06761, URL https://arxiv.org/abs/1704.06761.

Hult, G. T. M., Sharma, P. N., Morgeson, F. V., III, & Zhang, Y. (2019). Antecedents and consequences of customer satisfaction: Do they differ across online and offline purchases? *Journal of Retailing*, *95*(1), 10–23.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231.

Kahsay, G. A., & Samahita, M. (2015). Pay-what-you-want pricing schemes: A self-image perspective. *Journal of Behavioral and Experimental Finance*, *7*, 17–28.

Kellaris, J. J., Cox, A. D., & Cox, D. (1993). The effect of background music on ad processing: A contingency explanation. *Journal of Marketing*, *57*(4), 114–125.

Kenny, D., Kashy, D., & Bolger, N. (1998). Data analysis in social psychology. In *The handbook of social psychology, vol. 1* (pp. 233–265).

Kim, J. Y., Natter, M., & Spann, M. (2009). Pay what you want: A new participative pricing mechanism. *Journal of Marketing*, *73*(1), 44–58.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. http://dx.doi.org/10.48550/ARXIV.1412.6980, URL https://arxiv.org/abs/1412.6980.

Kitaguchi, D., Takeshita, N., Matsuzaki, H., Igaki, T., Hasegawa, H., & Ito, M. (2021). Development and validation of a 3-dimensional convolutional neural network for automatic surgical skill assessment based on spatiotemporal video analysis. *JAMA Network Open*, *4*(8), Article e2120786.

Koo, D. M., & Ju, S. H. (2010). The interactional effects of atmospherics and perceptual curiosity on emotions and online shopping intention. *Computers in Human Behavior*, *26*(3), 377–388.

Korbar, B., Tran, D., & Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 7774–7785). Red Hook, NY, USA: Curran Associates Inc.

Krishna, A. (2012). An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. *Journal of Consumer Psychology*, *22*(3), 332–351.

Krishna, A., Cian, L., & Aydınoğlu, N. Z. (2017). Sensory aspects of package design. *Journal of Retailing*, *93*(1), 43–54.

Kunter, M. (2015). Exploring the pay-what-you-want payment motivation. *Journal of Business Research*, *68*(11), 2347–2357.

Lalwani, A. K., Lwin, M. O., & Ling, P. B. (2009). Does audiovisual congruency in advertisements increase persuasion? The role of cultural music and products. *Journal of Global Marketing*, *22*(2), 139–153.

Lang, A. (2006). The limited capacity model of mediated message processing. *Journal of Communication*, *50*(1), 46–70.

Li, R., Lu, Y., Ma, J., & Wang, W. (2021). Examining gifting behavior on live streaming platforms: An identity-based motivation model. *Information & Management*, *58*(6), Article 103406.

Lipscomb, S. D., & Kendall, R. A. (1994). Perceptual judgement of the relationship between musical and visual components in film. *Psychomusicology: A Journal of Research in Music Cognition*, *13*(1–2), 60–98.

Logan, K. (2011). Hulu. com or NBC? Streaming video versus traditional TV: A study of an industry in its infancy. *Journal of Advertising Research*, *51*(1), 276–287.

Lu, Z., Xia, H., Heo, S., & Wigdor, D. (2018). You watch, you give, and you engage: A study of live streaming practices in China. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–13). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3173574.3174040.

Lu, S., Yao, D., Chen, X., & Grewal, R. (2021). Do larger audiences generate greater revenues under pay what you want? Evidence from a live streaming platform. *Marketing Science*, *40*(5), 964–984.

Maeda, F., Kanai, R., & Shimojo, S. (2004). Changing pitch induced visual motion illusion. *Current Biology*, *14*(23), R990–R991. http://dx.doi.org/10.1016/j.cub.2004.11.018, URL https://www.sciencedirect.com/science/article/pii/S0960982204008863.

Marett, K., Pearson, R., & Moore, R. S. (2012). Pay what you want: An exploratory study of social exchange and buyer-determined prices of iproducts. *Communications of the Association for Information Systems*, *30*(1), 10. http://dx.doi.org/10.17705/1CAIS.03010.

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*, *4*, 133–136.

Murauer, B., & Specht, G. (2018). Detecting music genre using extreme gradient boosting. In *Companion proceedings of the the web conference 2018* (pp. 1923–1927). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, http://dx.doi.org/10.1145/3184558.3191822.

Nesbitt, K. V., & Hoskens, I. (2008). Multi-sensory game interface improves player satisfaction but not performance. In *Proceedings of the ninth conference on Australasian user interface, vol. 76* (pp. 13–18).

Oakes, S., & North, A. C. (2008). Reviewing congruity effects in the service environment musicscape. *International Journal of Service Industry Management*, *19*(1), 63–82.

Oliver, R. (2010). *Satisfaction: A behavioral perspective on the consumer*. Routledge, http://dx.doi.org/10.4324/9781315700892.

Owens, A., & Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision* (pp. 631–648).

Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One*, *4*(5), Article e5664.

Peng, L., Cui, G., Chung, Y., & Zheng, W. (2020). The faces of success: Beauty and ugliness premiums in e-commerce platforms. *Journal of Marketing*, *84*(4), 67–85.

Petit, O., Velasco, C., & Spence, C. (2019). Digital sensory marketing: Integrating new technologies into multisensory online experience. *Journal of Interactive Marketing*, *45*, 42–61.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891.

Racherla, P., Babb, J. S., & Keith, M. J. (2011). Pay-what-you-want pricing for mobile applications: The effect of privacy assurances and social information. In *Conference for information systems applied research proceedings, vol. 4* (pp. 1–13).

Raghubir, P., & Krishna, A. (1996). As the crow flies: Bias in consumers' map-based distance judgments. *Journal of Consumer Research*, *23*(1), 26–39.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*(9), 2352–2449.

Roy, R., Rabbanee, F. K., & Sharma, P. (2016). Antecedents, outcomes, and mediating role of internal reference prices in pay-what-you-want (PWYW) pricing. *Marketing Intelligence & Planning*, *34*(1), 117–136.

Schmitt, B. (1999). Experiential marketing. *Journal of Marketing Management*, *15*(1–3), 53–67.

Scholler, S., Bosse, S., Treder, M. S., Blankertz, B., Curio, G., Muller, K. R., et al. (2012). Toward a direct measure of video quality perception using EEG. *IEEE Transactions on Image Processing*, *21*(5), 2619–2629.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. http://dx.doi.org/10.48550/ARXIV.1409.1556, URL https://arxiv.org/abs/1409.1556.

Smith, L. N., & Topin, N. (2019). Super-convergence: very fast training of neural networks using large learning rates. In T. Pham (Ed.), *Artificial intelligence and machine learning for multi-domain operations applications, vol. 11006*. International Society for Optics and Photonics, SPIE, Article 1100612. http://dx.doi.org/10.1117/12.2520589.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312.

Song, Y., & Soleymani, M. (2019). Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1979–1988).

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995.

Suris, D., Duarte, A., Salvador, A., Torres, J., & Giro-i Nieto, X. (2018). Cross-modal embeddings for video and audio retrieval. In *Proceedings of the European conference on computer vision (ECCV) workshops* (pp. 711–716). http://dx.doi.org/10.48550/arXiv.1801.02200.

Temme, J. E. V. (1992). Amount and kind of information in museums: Its effects on visitors satisfaction and appreciation of art. *Visual Arts Research*, *18*(2), 28–36.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489–4497).

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6450–6459).

Varol, G., Laptev, I., & Schmid, C. (2018). Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1510–1517.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., et al. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21–25.

Wang, Z., Zhou, J., Ma, J., Li, J., Ai, J., & Yang, Y. (2020). Discovering attractive segments in the user-generated video streams. *Information Processing & Management*, *57*(1), Article 102130.

Weisstein, F. L., Kukar-Kinney, M., & Monroe, K. B. (2016). Determinants of consumers' response to pay-what-you-want pricing strategy on the Internet. *Journal of Business Research*, *69*(10), 4313–4320.

Xian, Y., Li, J., Zhang, C., & Liao, Z. (2015). Video highlight shot extraction with time-sync comment. In *Proceedings of the 7th international workshop on hot topics in planet-scale mobile computing and online social networking* (pp. 31–36). Association for Computing Machinery, http://dx.doi.org/10.1145/2757513.2757516.

Yang, M., Adomavicius, G., Burtch, G., & Ren, Y. (2018). Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, *29*(1), 4–24.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, *27*, 3320–3328.

Zhang, B., Niu, L., & Zhang, L. (2021). Image composition assessment with saliency-augmented multi-pattern pooling. http://dx.doi.org/10.48550/ARXIV.2104.03133, URL https://arxiv.org/abs/2104.03133.

Zhang, Q., Wang, W., & Chen, Y. (2020). Frontiers: In-consumption social listening with moment-to-moment unstructured data: The case of movie appreciation and live comments. *Marketing Science*, *39*(2), 285–295.

Zhang, Z., Wang, X., & Wu, R. (2021). Is the devil in the details? Construal-level effects on perceived usefulness of online reviews for experience services. *Electronic Commerce Research and Applications*, *46*, Article 101033.

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., et al. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management*, *56*(6), Article 102097. http://dx.doi.org/10.1016/j.ipm.2019.102097.

Zheng, K., Zhang, Y., Lv, L., & Yang, C. (2018). Depth masking based binocular just-noticeable-distortion model. In *2018 IEEE international conference on multimedia & expo workshops* (pp. 1–5). http://dx.doi.org/10.1109/ICMEW.2018.8551562.

Zhou, J., Zhou, J., Ding, Y., & Wang, H. (2019). The magic of danmaku: A social interaction perspective of gift sending on live streaming platforms. *Electronic Commerce Research and Applications*, *34*, Article 100815. http://dx.doi.org/10.1016/j.elerap.2018.11.002.