STIVi: Turning Perspective Sketching Videos into Interactive Tutorials

Category: Research

Abstract

For design and art enthusiasts who seek to enhance their skills through instructional videos, following drawing instructions while practicing can be challenging.

STIVi presents perspective drawing demonstrations and commentary of prerecorded instructional videos as interactive drawing tutorials that students can navigate and explore at their own pace.

Our approach involves a semi-automatic pipeline to assist instructors in creating STIVi content by extracting pen strokes from video frames and aligning them with the accompanying audio commentary. Thanks to this structured data, students can navigate through transcript and in-video drawing, refer to provided highlights in both modalities to guide their navigation, and explore variations of the drawing demonstration to understand fundamental principles. We evaluated STIVi's interactive tutorials against a regular video player. We observed that our interface supports non-linear learning styles by providing students alternative paths for following and understanding drawing instructions.

1 Introduction

Instructional videos are a popular means for learning to draw as they offer step-by-step visual demonstrations as well as audio explanation of drawing concepts [12, 38, 44]. However, the pre-determined format of videos imposes several restrictions in an educational context [9, 47, 48]. Notably, videos play at a fixed pace, while students typically need frequent replays to understand and practice fundamental drawing concepts. Scrubbing a generic video timeline to locate these concepts can be a tedious trial-and-error process. Additionally, students must divide their attention between the video and their canvas, which can lead to missing demonstrated strokes or explanation details. Finally, videos often showcase only a few variations of a drawing concept. For example, although a cube can be used to demonstrate 2-point perspective, additional demonstrations are needed to show how this concept is applied to draw cubes of different sizes and from different viewpoints.

Our objective is to create augmented versions of instructional drawing videos that address the above limitations. In close interaction with two industrial design teachers (who are co-authors of this paper), we set three design goals: (1) to ease navigation of key concepts demonstrated in the video; (2) to mitigate attention splitting between the video and the canvas; and (3) to enable exploration of variations of the demonstrated concepts.

Our solutions draw inspiration from professional video-edited demonstrations of perspective drawing [14, 23] and previous HCI work on instructional videos for different domains [47, 48]. Typical instructional videos about drawing contain a visual demonstration of a drawing concept with an audio commentary. We leverage these complementary sources of information to identify relevant elements in the drawing (lines, planes, vanishing points), and to relate them to their descriptions in the video transcript. Combining the visual and textual representations of the drawing concepts allows us to augment the video with specialized interactions that fulfill our design goals. First, the instructor's pen strokes and their transcript descriptions serve as navigation landmarks to help students reach the parts of the video they are interested in. Second, highlighting the instructor's pen strokes when they are commented upon helps students identify the subject of the commentary, even when the video is paused. Third, transforming the instructor's pen strokes according to perspective rules helps students experience variations of the demonstration. We integrate these solutions into STIVi (**S**ketching **T**utorial from **I**nstructional **Vi**deo), an interactive tutoring system whose content is extracted from instructional videos on perspective drawing.

We describe a processing pipeline to assist instructors in converting existing instructional videos into interactive tutorials to be displayed in STIVi. We use image processing to extract individual pen strokes from the video frames, from which we deduce perspective properties of the drawing, i.e., location of the vanishing points and 3D orientation of the lines. In parallel, we apply speech processing to locate keywords about perspective drawing in the transcript. Finally, we match each visual element to its transcript description by computing a similarity score between the extracted lines and keywords. Instructors can then manually refine these matches to produce the final material.

We demonstrate this semi-automated pipeline on instructional videos of varying complexity. We also evaluate the usefulness of our interactive tutorials with a user study, where 12 participants follow drawing instructions using either a regular video player or STIVi. Subjective ratings suggest that our system helps learners navigate the tutorial, follow instructions, and understand the taught concepts. Furthermore, our analysis of interaction logs reveals that STIVi promotes non-linear navigation as users explore the tutorial with a different order than the one in the pre-recorded video.

In summary, we present the following key contributions:

1. We describe how the visual and textual elements of instructional videos on perspective drawing can be combined to ease the navigation, understanding and exploration of the taught concepts.

2. We illustrate the resulting tight coupling between visual and textual elements in the context of an interactive tutoring system that helps students follow and practice perspective drawing instructions.

3. We demonstrate how to semi-automatically extract visual and textual elements from drawing videos, and how to suggest instructors likely correspondences between the two modalities to create content for our tutoring interface.

2 Related work

Our work is at the intersection of prior research on two system types: interactive tutorials from instructional videos and intelligent tutoring systems for drawing.

Interactive tutorials from instructional videos. Most existing systems leverage domain-specific knowledge to analyze the structure of the content being taught. For example, Truong et al. [47] generate a hierarchical presentation of make-up tutorial steps based on facial parts, Wang et al. [48] provide waveform and melody visualization to support navigation and feedback in guitar instructional videos, Grossman et al. [18] and Chi et al. [9] record logs of application commands along with screen capture to generate software tutorials that highlight relevant UI components and actions. Likewise, we distill and leverage principles of perspective drawing to offer specialized features for navigation, highlighting and exploration of drawing instructions.

Our system builds on the observation that drawing instructors commonly comment on their drawing actions as they perform them. Following a similar observation, Shin et al. [43] rely on temporal

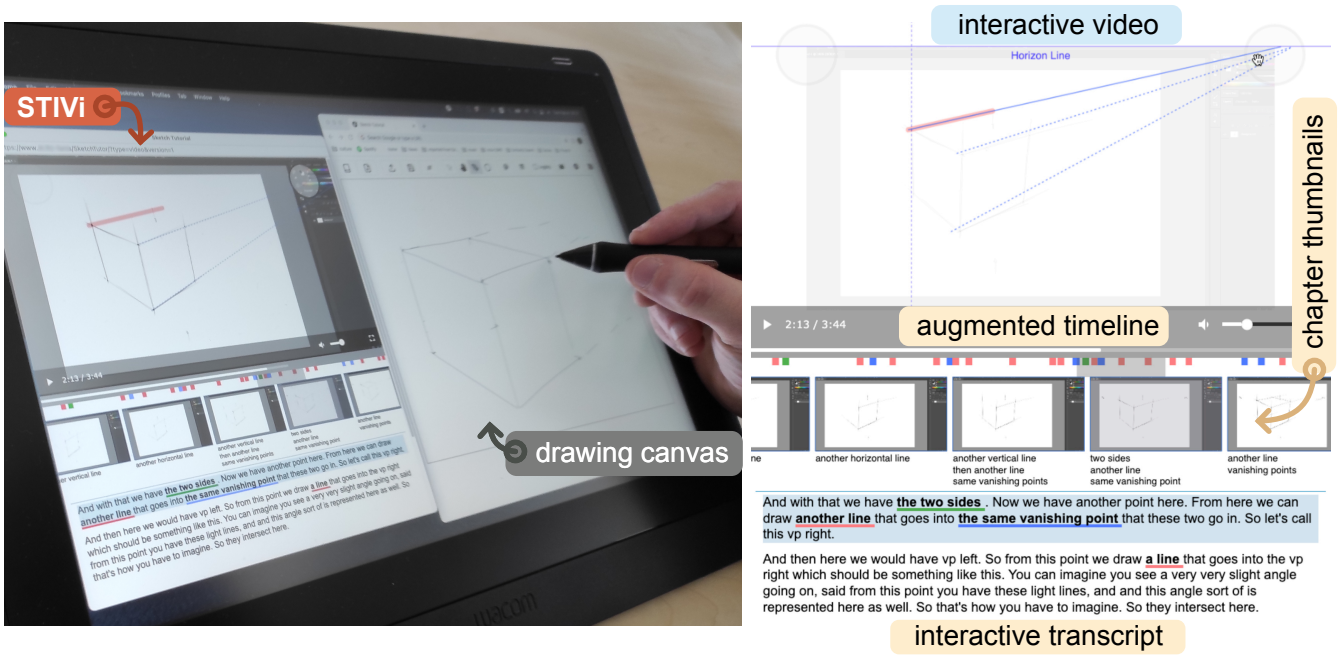

Figure 1: STIVi augments instructional sketching videos to help students follow perspective drawing instructions and practice at their own pace. STIVi's user interface links keywords from the video transcript to elements in the drawing, such as lines and planes, making it easier to navigate and familiarize with key concepts covered in the instructions. It also allows students to interact with perspective constructions (e.g., vanishing lines highlighted on the video) and understand how the geometry of depicted shapes would change if they were drawn from different viewpoints.

alignment to match transcript sentences with pen strokes traced by science instructors in blackboard-style lectures, while Jung et al. [27] and Kim et al. [30] use word embedding to match the video transcript to text elements in slide-based lectures. Similarly, we detect keywords in the transcript and put them in correspondence with lines in the drawing. Several computer vision and machine learning approaches have been proposed to automatically build correspondences between the transcript and visual content of instructional videos [25, 36, 53]. However, these methods have been trained on large datasets of natural videos rather than on drawings made of few lines. We tackle this challenge by leveraging geometric properties of the lines to identify likely correspondences with geometric terms in the transcript, even though we let the instructor decide on the final assignment, because drawing instructions often have multiple concurrent geometric interpretations.

Putting the video transcript in correspondence with the drawn elements allows us to use both modalities to navigate in the video. Our approach draws inspiration from two distinct yet complementary sources. First, we adopt principles from transcript-based interfaces [31, 40], where text acts as a navigational anchor for video exploration. Second, we are inspired by content-aware navigation techniques [13, 28, 39] that enable users to dynamically interact with video content and follow alternative navigation pathways. Furthermore, a number of systems [17, 32, 40] support navigation via video thumbnails that are linked to specific chapters. We leverage both the video transcript sentences and the drawing elements associated with these sentences to structure the tutorial into chapters.

Intelligent Tutoring Systems for drawing. While many people enjoy drawing, few feel confident in their drawing skills [8, 34]. This discrepancy has motivated the design of interactive systems that aim to assist novices in learning to draw. Many such systems rely on a pre-defined set of exercises distilled from traditional drawing lessons. In particular, Williford and colleagues developed a series of interactive tools that analyze user inputs to provide feedback on sketching accuracy as users perform custom exercises of increasing complexity, ranging from drawing straight lines, squares and circles [50] to drawing 3D primitives like cubes and cylinders [29], all the way to drawing buildings in perspective [49]. In a similar spirit, Lee et al. [35] describe a system dedicated to practicing perspective drawing of cars. In contrast, our goal is to ease the creation of new tutorials from pre-existing instructional videos, with a focus on highlighting the key steps and concepts explained by the instructor.

Closer to our goal are systems that generate drawing instructions from user-provided content, such as pictures of faces [10, 51] and objects [26], or 3D models [22]. A major part of these systems consists in analyzing the input image of 3D shape to generate step-by-step instructions. Our approach differs as we take as input videos where instructors demonstrate and explain drawing techniques one step at a time. Our challenge thus resides in helping users follow the instructions, which we achieve by augmenting the video with visual highlights and various interactions supporting navigation.

Sketch-Sketch Revolution [16] allows expert users of a drawing software to create tutorials for novices by demonstration. The authoring interface allows the creators to draw each step of their tutorials, and to add labels and additional instructions using text fields. Our system's original feature lies in its ability to extract visual and textual instructions from videos with a semi-automated approach. Additionally, it establishes meaningful connections between the key concepts articulated by instructors and the corresponding visual elements they depict, such as lines, planes, and vanishing points.

3 Challenges and Approach

Prior work has identified key limitations inherent to instructional videos [9, 47, 48], which we aim to address in the context of drawing instructions:

- **Difficulty in locating key instructions.** Users frequently find themselves scrubbing back and forth along the video timeline to spot interesting events. Moreover, their timestamp selection needs to be precise for quick events.

- **Difficulty in following instructions.** A notable drawback of conventional videos is their fixed pace, which fails to adapt to individual needs. This limitation becomes particularly pronounced when users practice the instructions while following them. In practice, users often need to pause the video to catch up, which can cause them to miss out on dynamic information about the visual content that is commented upon.

- **Difficulty in generalizing instructions beyond demonstrated cases.** Instructional videos usually showcase a methodology on a limited number of examples. Thus, users have to follow multiple videos on the same topic to grasp how the instructions apply to slightly different situations.

Our key insight is that, in the context of drawing instructions, we can address all three difficulties by identifying the drawing concepts (shapes and their relationships) that form the subject of the instructions. Once these concepts are identified, we allow users to navigate in the tutorial by selecting the concepts they are interested in. Furthermore, we highlight the visual elements associated with these concepts when they are commented upon to make them stand out even when the video is paused. Finally, we allow users to interact with the visual elements to explore how they vary under the principles they obey. We demonstrate such exploration in the context of perspective rules defined by the relative position of vanishing points on the canvas.

Figure 2 illustrates the main steps of our methodology for converting drawing instructions into interactive tutorials:

1. We first compile the key concepts in the domain of interest. As shown in Figure 2, we distinguish among *shapes*, their *properties*, and their *relationships*. In the case of perspective drawing, shapes can be *lines* or *planes*, properties can be *vertical*, *ground*, or *horizontal*, and relationships can be *parallel*, *crossing*, etc.

2. For each concept, we distill a multi-modal vocabulary (Figure 2a) that encompasses how the concept can appear both visually – as a sketch in the video frames – and textually – in the video transcript.

3. We extract these visual and textual elements using video and transcript processing (Figure 2b).

4. We put visual and textual elements in correspondence based on temporal alignment and compatibility between the two vocabularies (Figure 2c).

5. Finally, after manual correction of the extracted data, we leverage the resulting visual-textual structure to augment the video with navigation landmarks, visual highlights, and interactive variations (Figure 2d).

We next introduce the domain-specific knowledge we compiled for perspective drawing (Section 4) and describe how we integrate these concepts into our interactive tutoring system for perspective sketching (Section 5). The technical details about how we extract the concepts from instructional videos are presented in Section 6.

4 Perspective drawing concepts

While the video augmentations we propose could apply to various domains of visual art, we demonstrate them in the context of *perspective drawing*, which is ubiquitous in fine arts, architecture, industrial design, interior design. Perspective drawing has also been the focus of several prior systems for interactive tutoring [29, 35, 49, 50].

We first compiled basic principles of perspective drawing from existing videos [12, 23, 33, 38, 41, 44] and design textbooks [15, 42]. This process was guided and enriched by discussions with two industrial design teachers co-authors of this paper, who reviewed the principles and gave insights on their relevance and difficulty for students. From these observations, we chose to focus on the fundamental principles for drawing block shapes made of straight lines. Block shapes are among the first topics taught in class to develop awareness of perspective rules, and often act as preliminary scaffolds for drawing furniture, buildings, as well as more elaborate curved shapes [20]. Together with our design collaborators, we distinguished recurring concepts in drawings of block shapes and, for each of these concepts, we identified the commonly employed vocabulary used for commentary.

Line. Straight lines can be composed of one or several pen strokes. In perspective drawing, most straight lines are aligned with one of the three orthogonal axes of the 3D world that is depicted. As such, instructors often comment on the *verticality* or *horizontality* of the lines they draw, even though these lines are not necessarily vertical nor horizontal when projected on canvas, but oriented towards a vanishing point (see Figure 3).

Plane. Block shapes are formed of planar faces, many of which are drawn as quads delineated by two sets of parallel lines. Instructors refer to such surface elements as *planes*, *rectangles*, or *sides*.

Vanishing point. In perspective drawing, lines that are parallel in the 3D world project to lines that converge towards a vanishing point on the canvas. In two-point perspective, two sets of orthogonal horizontal lines converge to two vanishing points, both of which lie on a horizon line (possibly away from the boundaries of the canvas), as illustrated in Figure 3. In three-point perspective, vertical lines also converge towards a vanishing point, typically placed above or below the canvas boundary depending on viewpoint. Instructors often name the different vanishing points to explain these principles, for instance by distinguishing the *first* vanishing point being drawn from the *second* one, or the one lying on the *left* of the canvas from the one lying on the *right*. They might also point to groups of lines that *converge* towards the same vanishing point.

5 Learning from interactive videos

Our design efforts focus on exposing fundamental drawing concepts, clarifying their connections with the instructor's explanations, and providing opportunities for students to actively engage with them through interactive exploration. Figures 4-5 illustrate the functionalities of our system. We introduce them through a usage scenario:

> *Gabriela is learning how to draw in perspective. After having watched a few videos about drawing basic shapes, she now wants to move to more complex objects. She uses STIVi to follow a video tutorial that explains how to draw an armchair.*

5.1 Following the video tutorial

> *Gabriela begins by watching the full video, carefully following the teacher's instructions. As she watches, parts of the drawing get highlighted in sync with the instructor's voice. For example, as the instructor explains how to "strengthen the outline" of the inside of the chair (Figure 4a), the strokes of its outline are highlighted in red. Gabriela notices that the term "the outline" is also highlighted in the same color in the transcript, which helps her associate the terminology used in the video with the visual elements in the instructor's drawing.*

Instructors often highlight techniques by gesturing, e.g., drawing over lines, pointing at planes, emphasizing keywords in voice etc. STIVi complements such gestures by highlighting the relevant elements of the drawing when these elements are mentioned. The goal is to make the associations between instructor's speech and drawing explicit and thus help students follow instructions, even when they shortly look away from the video when a line is drawn, or if they pause the video to practice the instructions. The system uses three different types of video augmentations: (1) thick red

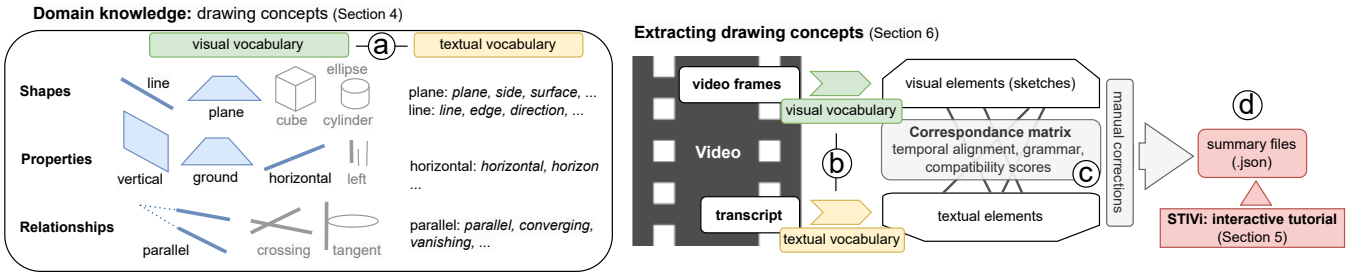

Figure 2: Overview of our tutorial creation approach. We compile key concepts of perspective drawing along with their visual and textual vocabulary (a). In this paper, we focus on a subset of such concepts (in blue). Then, we process the video frames and the transcript to extract these concepts (b). Finally, we associate visual and textual elements that coincide in time and refer to compatible concepts to form a multi-modal data structure (c), which we leverage to augment the video in STIVi (d).

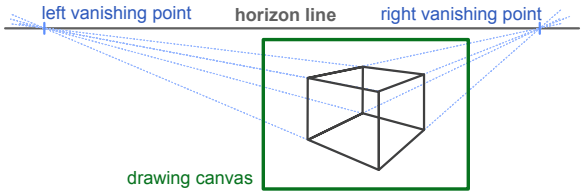

Figure 3: Illustration of the two-point perspective setting: the vanishing points lie along a horizon line that can be positioned outside of the drawing canvas. Lines converging towards the vanishing points on the canvas are horizontal in 3D, while lines that are vertical on the canvas also remain vertical in 3D.

line segments to highlight lines in the drawing (Figure 4); (2) green quadrilateral shapes to highlight planes (Figure 5); and (3) blue lines to highlight convergence relationships and vanishing points (Figure 5). We implement a cross-dissolve animation effect to display and then hide these augmentations in sync with the audio.

STIVi applies the same color-coding scheme to references of the above drawing concepts in the transcript (e.g., **the outline** refers to a group of lines, and **the sides** refers to a pair of planes). Previous studies suggest that accompanying videos with regular text does not have any positive effect on learning [46]. However, by highlighting drawing concepts in the text, we aim to direct students' attention on them [6, 37] and help them associate the concepts illustrated in the video with the terminology used by the teacher.

5.2 Navigating in the video content

After watching the video once, Gabriela opens a new project and starts drawing. To keep up with the instructions, she pauses and replays the video frequently. To review previous steps, she navigates back and forth between chapters using their thumbnails. As the video explains how to draw the inside of the armchair (Figure 4a), Gabriela decides to skip ahead in the tutorial and work on the contour outline of the chair. She presses the pen button and selects its group of strokes from a preview (Figure 4b), which advances the video to the last chapter (Figure 4c). Gabriela moves her pen to the transcript and clicks on the underlined words to navigate in the chapter.

In traditional video interfaces, viewers can navigate through the video by directly interacting with the timeline or by using fast-forward or rewind functions. STIVi augments such navigation by structuring the video into a sequence of chapters, and by offering both transcript-based and drawing-based landmarks to directly jump to relevant segments.

Timeline. To help students locate the concepts (lines, planes, and their relationships) discussed in the video, STIVi annotates the timeline with colored cues that serve as time landmarks (Figure 4). The color code corresponds again to the element type: red for lines, green for planes, blue for convergence relationships. Furthermore, the system segments the video into chapters, and adds a video thumbnail for each chapter, in a scrollable list below the timeline, together with a set of keywords that correspond to references to sketch elements in the transcript. The student can move the pen over the thumbnails list to scroll it forwards or backwards and switch between chapters.

Transcript. The area beneath the thumbnails shows the transcript for the currently active chapter (Figure 4). STIVi highlights the active paragraph and allows the student to click on the text to instantly navigate to the corresponding moment in the video when it is spoken. The colored references of sketch elements in the transcript (see Subsection 5.1) operate as hyperlinks. The student has the option to click on these links to adjust the video's playback position and activate animations that highlight the related sketch elements. When combined with the list of thumbnails, this feature provides a seamless way to navigate between different parts of the video and easily identify the relevant sketch elements.

On-video sketches. STIVi offers an alternative method of navigating the video by directly interacting with its visual content. As shown in Figure 4b, the student can press a pen button to reveal the complete set of lines and planes in the sketch. These elements are grouped together based on how they are referenced in the transcript. For instance, all the strokes associated with "all these lines" in Figure 4c are part of the same group. By moving the pen over the sketch, the student can explore the available groups and receive feedback in the form of textual tooltips and timeline pointers indicating when and where these groups are first referenced in the transcript. The student can then jump to the corresponding point in the video by releasing the pen button or selecting the group with a tap of the pen.

5.3 Exploring perspective drawing concepts

Gabriela is eager to practice drawing the armchair from a different viewpoint. She loads a video tutorial that shows how to draw the container cube for the armchair and begins sketching the lines of the front planes of the cube. However, as she chose a different perspective from the one in the video, she is unsure about how the side lines should converge. To better understand the concept, Gabriela activates the circular widgets located near the top corners of the video (Figure 5b-c). She manipulates the widgets to reposition the horizon line and the two vanishing points, which allows her to observe how the planes rotate when changing viewpoints.

As opposed to simple concepts such as lines and planes which are visible in the drawing, relationships may be hard to grasp by novice learners. Yet, such relationships are crucial for understanding higher-

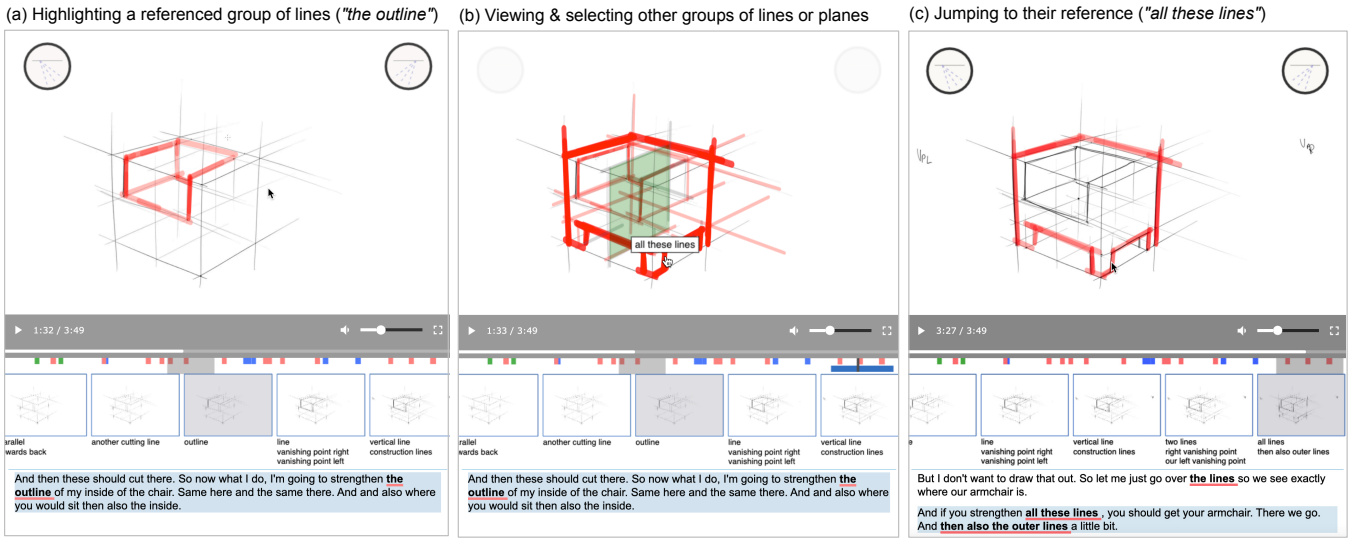

Figure 4: Navigating between different groups of sketched elements discussed in the video: (a) A group of lines that form an outline in the sketch are highlighted in red when the teacher verbally refers to them. (b) The student presses a pen button (or key on the keyboard) to view the full set of lines and planes referenced in the video. Here, the student hovers over the contour of the armchair and can preview how this group of lines are mentioned and when (see feedback below the timeline). (c) The student releases the pen button (or taps the pen on the screen) to jump to its reference in the video.

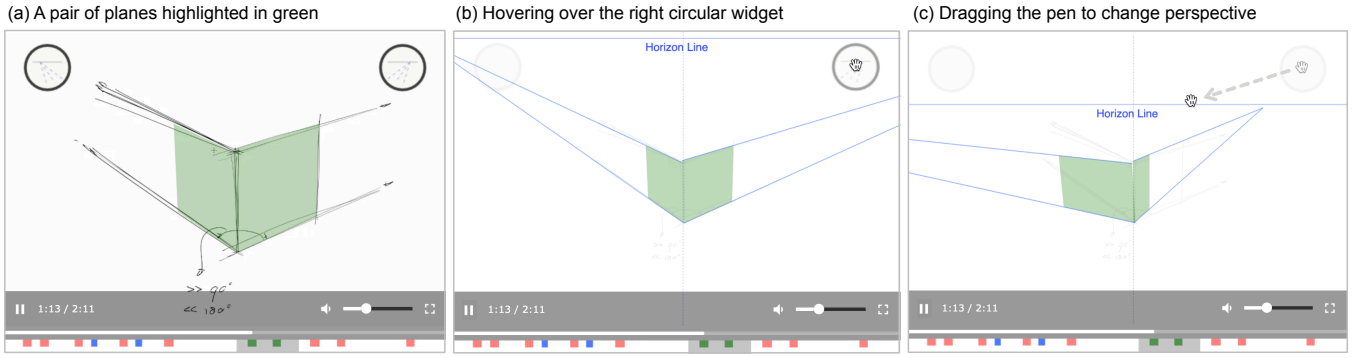

Figure 5: Exploring perspective drawing through interaction: (a) The teacher makes reference to the two front planes in the sketch (*"the two planes that are in front..."*). The two circular widgets appear and serve as interaction entries. (b) When the student hovers the pen over a widget, the scene is augmented with perspective lines and is animated to a lower scale to bring the horizon line within the window's view. (c) The student drags the pen to translate the horizon line and the two vanishing points, which rotates the two planes around their vertical intersecting line. The right vanishing point now becomes visible.

level principles of perspective drawing and for learning how the elements in a sketch should vary if viewed from a different perspective. For example, in two-point perspective, horizontal lines should converge to a left or a right vanishing point, depending on their direction, while vertical lines should appear as vertically parallel on canvas. Understanding how convergence parameters, such as the location of the horizon line or the location of a vanishing point, affect the orientation of horizontal lines is necessary to mentally rotate objects in a sketch and draw them from varying viewpoints [15].

STIVi lets students interactively explore these concepts, making the assumption that all sketches fall under the rules of one, two or three-point perspective with straight lines. To this end, we use references of basic elements (lines and planes) and their relationships in the transcript as entry points of interaction. In the example of Figure 5, the teacher refers to a pair of front planes, which are highlighted by the system. In addition, two circular interactive widgets appear and are associated with the two vanishing points along the horizon line. When the student hovers the pen over the right widget, the scene is animated to show the planes in a smaller scale, while revealing the convergence lines, the horizon line, and vertical reference axis. The user can now drag the pen to the left, which makes the planes rotate around the vertical axis and reveals the right vanishing point.

The animation coupled with this set of interactions gives users the illusion of an interactive 3D scene. Our solution relies on sketch processing techniques (see Section 6.1) and principles of 3D perspective. Since reconstructing 3D information from drawings is a difficult problem [11, 19, 21], we demonstrate our interactions in a simplified setting where we focus on specific groups of elements and their convergence relationships (i.e., the ones the teacher refers to at this particular moment), hiding other parts of the drawing (Figure 5b,c).

5.4 Implementation

STIVi is a Web application developed with Angular v14 [1], using Ngx-Videogular [2] for the video player and Paper.js [3] for video overlays. Visual and textual information extracted from videos is provided to the Web application as structured JSON files.

6 Extracting drawing concepts

Our tutoring interface requires knowledge of the individual elements that are drawn by the instructor, and of the comments that the instructor said about these elements. We now describe how we extract such visual and textual information from existing instructional videos, and how we assist the instructor in relating these two modalities to produce data suitable for our interface. We refer readers to supplemental materials for technical details.

6.1 Visual extraction

Our processing pipeline takes as input videos of digital drawing sessions captured via screen and audio recording, which we optionally crop to remove user interface elements that surround the canvas. These videos are relatively short (1 to 6 minutes long) and entail the drawing of one or a few objects. We also assume that the canvas remains static during the video sequence (no rotation, scaling, translation). Starting with the raw video frames, our algorithm extracts visual elements and their relationships in a bottom-up fashion.

Extracting strokes and lines. Observing that the pixels occluded by the mouse pointer exhibit rapid variations of intensity across frames, we filter out the pointer using a temporal median filter. We then extract pen strokes as they appear on canvas by computing the difference between each frame and its subsequent frame. To prevent false positives due to video compression artifacts, we only keep high-difference pixels that form short linear structures in the difference image. Finally, we recover long or overlapping lines by merging strokes drawn across successive frames if they share extremities.

Extracting vanishing points. We next run the vanishing point detection algorithm of Gryaditskaya et al. [19], which has been designed to handle the approximate perspective of freehand sketches. In addition to the dominant vanishing points, the algorithm classifies each line as being either horizontal and converging to the left vanishing point, horizontal and converging to the right vanishing point, vertical and optionally converging to a bottom or top vanishing point, or following any other direction.

Extracting planes. We detect planes in the drawing by searching for patterns of four intersecting lines made of two pairs of parallel lines converging to different vanishing points. We define the timestamp of a plane to be the timestamp of the last of its four lines, and its orientation to be vertical if one of the pairs of parallel lines is vertical and horizontal otherwise.

6.2 Text extraction

We first convert the instructor's audio commentary into a transcript, from which we then locate keywords referring to typical drawing elements present in perspective drawings, along with qualifiers that inform us about geometric properties of these elements.

Extracting the transcript. We use VOSK, an automatic transcription tool [5], to convert the instructor's commentary into English text. This tool produces a transcript synchronized with speech, where each word has a timestamp. We optionally manually correct the transcript in the presence of errors, which we mainly observed on technical terms pronounced by non-native speakers.

Extracting keywords. We next process the raw transcript with the NLP library spaCy [4] to obtain so-called *chunks of words*. Each chunk consists in a group of words composed of a noun – called the *root* – and its qualifiers and article. Figure 6 provides a typical example where the chunk "the left vanishing point" is composed of the root "point" qualified as "left" and "vanishing." We keep any chunk for which the root appears in our pre-defined vocabulary of perspective drawing elements. While many instructors employ the same vocabulary to comment on these elements, some also adopt unique terms, such as a *stick* to refer to a vertical line. We account for this diversity by allowing instructors to augment the pre-defined vocabulary (see textual vocabulary in Figure 2) with custom synonyms.

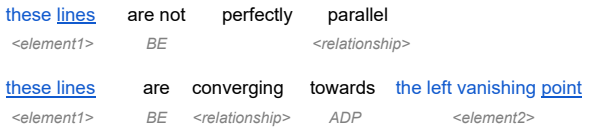

Figure 6: Extracting and relating speech elements. In this transcript, chunks of words are highlighted in blue, their roots are underlined. We keep the chunks of words whose root corresponds to a pre-defined vocabulary of perspective drawings. We then employ syntactic templates to find relationships between elements, such as parallelism and convergence.

Extracting geometric relationships. We detect geometric relationships expressed in the transcript in two ways: through syntactic patterns explicitly stating a relationship between detected elements, and through the nouns and qualifiers of chunks of words. We focus on convergence relationships between lines as they are at the core of perspective drawing of block shapes.

Figure 6 illustrates two representative syntactic patterns that convey convergence relationships. In the first example, the pattern indicates that the components of the element "these lines" share the relationship "not perfectly parallel." In the second example, the pattern also indicates towards which vanishing points the lines converge. We identify such syntactic patterns from the part-of-speech tagging and dependency graph generated by spaCy [4].

Geometric relationships can also be deduced from nouns and their qualifiers. For example, "the converging lines" or "the parallels" directly convey a convergence relationship between the lines referred to, while "the plane" implies a convergence relationships between its opposite sides. Whenever we detect a convergence relationship between lines that are present in the drawing, we highlight this relationship accordingly (blue-dotted lines overlaid on the video and blue link in the transcript as in Figure 1).

6.3 Linking speech and sketch

Processing the video frames yields visual representations of lines and planes, along with their orientation and convergence towards vanishing points. Processing the transcript yields textual representations of drawing elements, along with their geometric properties and relationships. We now relate these two modalities.

Building correspondences. To enable all of the interactions described in Section 5, it is necessary to build correspondences between textual and visual elements. We achieve this goal by computing a similarity score between all pairs of visual and textual elements. While prior work relies on similarity scores to automatically find one-to-one correspondences by solving a bipartite graph matching problem [30], we face the additional challenge that individual comments in the transcript can refer to several, ambiguous visual elements in the drawing, such as "these lines." This challenge motivated us to adopt an interactive workflow, where we use the similarity score to suggest correspondences between keywords and visual elements, which are then confirmed by the instructor.

Our similarity score combines a measure of temporal alignment with a test on the geometric compatibility between the visual and textual content. Denoting t_i the timestamp of visual element i and t_j the timestamp of textual element j, we express the score as:

$$S(i,j) = \mathbb{1}_{\text{geometry}}(i,j)\left(1 - \frac{|t_i - t_j|}{T}\right) \quad (1)$$

where $\mathbb{1}_{\text{geometry}}(i,j)$ equals one if the visual and textual elements share compatible geometric properties, and zero otherwise. Geometric properties include the type of element (line, plane) and its orientation (vertical, horizontal).

Once correspondences between lines and the transcript are established, we propagate them to the relationships of these lines. For example, when a set of lines are described as "converging" in the transcript, then the vanishing-point interaction widget for these lines is enabled.

Extracting chapters. Putting the commentary in correspondence with the drawing also enables us to segment the transcript into chapters covering a single topic.

We start with a similar approach as Truong et al. [47], although implemented with different tools. The transcript produced by the VOSK toolkit [5] is already split into parts based on timing – if the instructor stops talking for a little time, the next words are put into a new group. We implement an additional split based on Truong et al.'s observation that phrases describing different steps are often connected by conjunction words, such as "and" or "so". We split groups of words at conjunction words connecting two verbs, which we locate using the part-of-speech tagging and dependency graph generated by spaCy [4].

This initial processing tends to over-segment the transcript. We next group neighboring phrases using the information provided by our correspondences between textual and visual elements. If consecutive phrases include text that refers to the same visual element, such as the same plane, or to overlapping groups of visual elements, we assume that these phrases belong to the same chapter. We display the text corresponding to these recurring visual elements below each chapter thumbnail (e.g., see Figure 4).

6.4 Results and manual corrections

Dataset. Two industrial design teachers co-authors of this paper defined four standard beginner-level exercises on perspective drawing. We presented these exercises to two *other* design teachers and asked them to record instructional videos to explain them to an audience of learners. Both teachers were experienced in recording themselves while drawing, and one of them had prior experience creating and sharing video content about design sketching on online platforms. In addition to the eight videos produced by these designers, we tested our extraction pipeline on five online videos that we found by searching for domain-specific keywords (design sketching, perspective sketching) and by browsing dedicated channels. We selected videos compatible with our extraction pipeline, i.e. recorded within a digital drawing software, with a fixed canvas, a clear commentary, and mostly composed of straight lines.

Tutorial generation. Figure 11 shows a few steps of the interactive sketching tutorials we extracted from some of these videos. We refer readers to our supplemental materials for recordings of typical interactive sessions with these tutorials.

The whole process of generating an interactive tutorial from an existing video entails several manual correction steps. We quantified these corrections on the eight videos produced by the industrial design teachers. The first step consists in extracting individual pen strokes (Section 6.1). The number of extracted strokes varies from 50 to 300 depending on the complexity of the drawing, with an average of 120 strokes. On average, 76% of these strokes were accurate, while 9% required correction of one of the segment extremities, and 14% had to be deleted. Three of the videos also required additional strokes (around 20% of the initial number of strokes). Spurious or missing strokes are typically due to low contrast or fast movement of the pen. The second step consists in extracting the transcript, which required an average of 17 word corrections for videos up to six minutes long. A few videos required adding custom synonyms to our pre-defined vocabulary (textual vocabulary in Figure 2). The last and most involved step consists in putting the speech and sketch elements in correspondence using our similarity score as guidance (Section 6.3). This task took approximately 30 minutes per video.

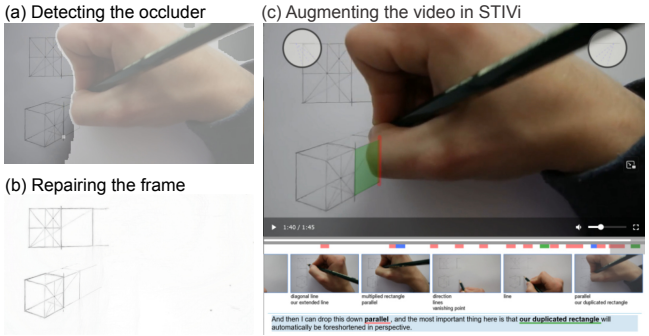

Figure 7: Processing videos with hands drawing on paper: (a) We use a skin and shadow detector, and morphological filters to segment out the hands. (b) We identify the pixels corresponding to pen strokes even in the presence of cast shadows. (c) Once pre-processed, the video can be converted into an interactive STIVi tutorial that highlights the commented lines, planes and vanishing points.

6.5 On-paper drawing instructions

We have also experimented with videos of on-paper instructions captured with a camera positioned over a drawing table. Such videos present additional challenges due to the presence of the hands of the instructor, which occasionally occlude the drawing or cast shadows over the canvas. We dealt with these challenges with additional image processing, including a skin and shadow detector to locate the hands, morphological closing to locate thin pen strokes, and temporal filtering to propagate the drawn pixels across occluded frames [45]. Figure 7 shows a representative frame of a tutorial generated using these additional steps (see supplemental material for the entire tutorial). However, this additional processing incurs dedicated thresholds and manual corrections to obtain data of sufficient quality to be displayed in STIVi. We thus defer the development of a robust pipeline for on-paper drawing sessions to future work.

7 User Evaluation

We conducted a qualitative user study[1] to assess how STIVi can help students follow the teacher's instructions, navigate in the video content, and understand key drawing concepts. STIVi integrates both novel features (video highlights, interactive exploration tools) and features found in existing video systems (clickable transcript, chapter thumbnails). To encourage participants to reflect on each of these features and their combined use in STIVi, we also asked them to complete a task with a Baseline interface. The Baseline consisted of a regular video player supporting basic navigation capabilities via a conventional timeline.

7.1 Method

Participants. 12 volunteers (8 women and 4 men) participated in our study. These participants were recruited through mailing lists provided by our local university and were mostly graduate students or worked in research. Among the participants, seven were between the ages of 18 to 25, four were between the ages of 26 to 35, and one was between the ages of 46 to 55. The participants were non-experts with varying drawing experience: eight participants practiced drawing, while four declared they drew sometimes (please refer to supplemental materials for more details on background).

Video materials. We conducted the study with two videos from the same instructor (not co-author of this work), from the dataset described in Section 6.4. The first video (Cube, Figure 11a) shows how to draw a cube in two-point perspective and has a duration of

[1]The study was approved by our Ethics Review Board and followed our institution's data protection rules.

3 minutes and 44 seconds. The second video (ARMCHAIR used in Figure 4) shows how to refine a cube into an armchair, and has a duration of 3 minutes and 49 seconds. In addition, we used a shorter video for training, lasting 2 minutes and 8 seconds. This video was extracted from an online instructional video [38] and showcases the construction of modeling planes (Figure 11c).

Apparatus and design. The participants interacted with a Wacom Cintiq 16 pen display (1920×1080 FHD resolution), using the configuration shown in Figure 1. All 12 participants tested both the BASELINE and STIVi. Six participants tested STIVi first; the other six participants tested our system second. We kept the same logical order of the two videos for all participants, where the simple CUBE was always presented before the more complex ARMCHAIR .

Procedure and task. After filling out a short background questionnaire, participants were introduced to each system configuration, one after the other. For each configuration, they spent some time to familiarize themselves with the user interface, going through the training video. We then introduced them to the main task, which had two parts. As a first step, we asked participants to watch the video tutorial (either the CUBE or the ARMCHAIR), trying to understand its content. They could use any of the video navigation tools presented to them during training. As a second step, we asked participants to reproduce the sketch of the video as closely as possible. While drawing on the canvas, they could again return to the video tutorial and browse its content. Participants were encouraged to think aloud during the task. When they were satisfied with their drawing, they answered a five-question questionnaire to report on their learning experience.

Following the testing of both system configurations, participants were asked to complete an evaluation questionnaire. The questionnaire prompted them to compare the two configurations across three evaluation criteria — navigating, following instructions, and understanding concepts — as well as to provide feedback on the usefulness of individual features of our system. Each session lasted approximately 60 to 70 minutes.

7.2 Results

Task evolution and strategies. We start by comparing the temporal strategies that participants followed to complete the task with the BASELINE and with STIVi. As seen in Figure 8-left, when using the BASELINE, participants adopted a conservative strategy of pausing and replaying the video while drawing. We also observe that several participants (e.g., P5, P7, P2, and P6) finalized their drawings after the video had ended, that is, without following instructions. In contrast, the patterns we observe for STIVi in Figure 8-right have a large number of drops and spikes, since participants could use the system's features to jump more efficiently between different moments in the video. Participant P10's trajectory is particularly noteworthy. After watching the video once, the participant did not replay it again. He relied instead on the transcript and selected video frames, which he could directly access by navigating through the chapter thumbnails or the transcript hyperlinks. In contrast, P6 followed the play-pause strategy while drawing but also used STIVi's features to reflect on and evaluate the quality of her work. She then continued with a second round to refine her drawing.

When we asked participants to describe the strategy they followed, several participants described how STIVi's features allowed them to adapt the pace of the video: *"Overview and then small jumps, skipping parts I didn't want ... there were points where I had a little more doubts, I took my time to see what was happening and go faster on other parts"*[2] (P3).

Participants relied on several of STIVi's features to navigate the tutorials. Figure 8-right shows six representative examples, where we indicate navigation events through the chapter thumbnails, the on-video sketches, or the transcript hyperlinks. In the following paragraphs, we discuss participants' feedback about individual features of STIVi.

Video highlights. Several participants reported that STIVi's highlighted elements were useful for clarifying the video material, helping them focus on the relevant parts of the drawing as they switched attention between video and canvas, *"the red [highlights] would come up to kind of show what was the last thing that he had just done. Because I do look at the other [window] and back at the video, it brought me in to where we currently were"* (P8). In contrast, highlighted sketches were sparingly used for navigation. P3 and P11 commented that their visualization was cluttered. P1, who used this feature, found it useful when she knew which group of elements she wanted to choose from.

Finally, several participants (P3, P6, P10, P11, P12) suggested that drawing assistance could be particularly useful if it was applied directly to the canvas.

Transcript and links. Links embedded within the transcript served various purposes, including helping participants *"find points of reference"* (P4), facilitating easier parsing of the text, and serving as shortcuts to make highlights appear on the video (P6): *"This highlighted text ... kind of teaches me what are the key points ... So I click on it and try to explore. It makes me understand more easily."* However, the transcript itself was deemed distracting or overwhelming by some participants (P2, P6, P9, P12), which is somewhat consistent with past results [46] on the utility of text in educational videos. P12 mentioned that she *"would only focus on underlined words, unless the sentence is really short."*

Thumbnails. Several participants frequently used the thumbnails. Participants were especially positive about their use for organizing the material into chapters, *"I find that it's easier to locate yourself with the chapters ... because you really get the visual of the drawing you have to do"* (P2). Another participant remarked that since the keywords below the thumbnails were *"part of the speech, it was much easier to remember where we are ..."* (P10).

Exploration tools. Although the perspective exploration features received positive feedback from most participants, several participants reported forgetting about their use due to the presence of other tools. P6 and P8 also remarked that reproducing the drawing closely left little room for exploring variations: *"That could be useful actually, like when you did something wrong, and when you try different perspectives"* (P6). P8 also suggested adding presets that could be dragged and dropped to create multiple views: *"presets to where I could drag it to, and ... have it stick there. ... So I kind of felt like it would be interesting if I was doing multiple views."* Finally, P5 and P7 explained that since they were already quite familiar with the concept of perspective, they did not find these tools useful at that point, but would have liked to have them when starting learning about perspective, in particular 3-point perspective (P5).

Overall assessment. Figure 9 summarizes participants' subjective evaluation. Overall, the participants found that STIVi's extended features helped them navigate the video more effectively, follow the teacher's instructions, and to some extent, understand concepts.

More specifically, participants appreciated the complementary ways of supporting learners in STIVi. P8 emphasized that information was presented *"in multiple different ways ... so it was easier to tell where we were contextually."* P3 suggested that this aspect of the system can support different learning styles: *"it's good because you can address both types of learners with the same material without having to create two different things."*

In contrast, several participants reported feeling lost when using the BASELINE with the more complex video and wishing they had the tools from STIVi (P2, P4, P6, P8). P2 said that she struggled mostly *"to understand the drawing itself as there are no planes, no*

[2]Quotes from P1, P2, P3, P4, P5, P7 and P10 have been translated from the original language to English.

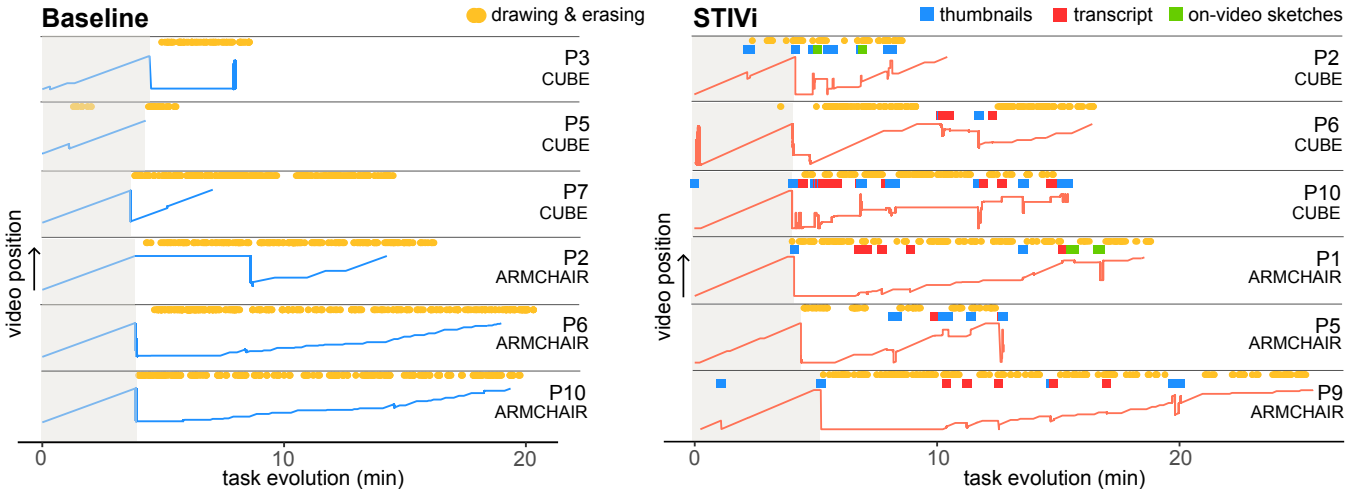

Figure 8: Evolution of the task for a subset of participants while interacting with BASELINE (left) and STIVi (right). The line trajectories show the time position on the video as a function of the time the participant spends on the task. The gray area corresponds to the initial phase of watching the video tutorial. For STIVi, we highlight user interaction events as yellow dots (drawing events) and colored squares (navigation events through thumbnails, transcript hyperlinks, and on-video sketches).

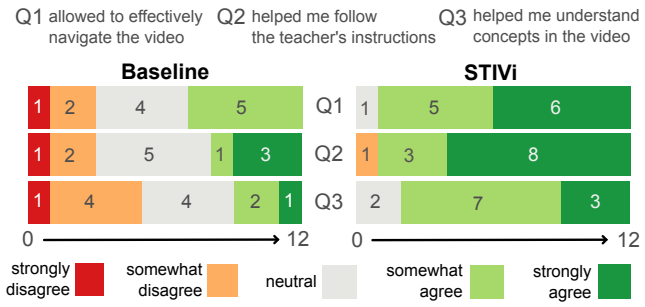

Figure 9: Subjective evaluation of the two system configurations.

different colors to better visualize what you have to draw and to what perspective they correspond." P6 also elaborated on how she felt unable to find where she made mistakes with the BASELINE: *"this one [with* BASELINE*], I'm completely lost. So I don't even know where I made mistakes."*

8 LIMITATIONS AND FUTURE WORK

We discuss limitations of our approach and explore potential avenues for future extensions.

Use of exploration features. Our study design targeted the reproduction of the instructor's demonstration, which is the first step in learning a new skill. Therefore, it did not trigger the application of the taught concepts to novel shape or viewpoint configurations. Further investigation is needed to see how our perspective manipulation features (Figure 5) would be solicited for such generalization tasks.

Feedback and drawing assistance. While STIVi augments the demonstration video, similar augmentations might also apply to the drawing canvas. For example, some participants suggested revealing vanishing lines on canvas to help draw converging edges of block shapes. Going further, registering the user pen strokes against the video demonstration could enable automatic assessment of drawing accuracy, which could be used to provide personalized feedback, or to adjust the amount of guidance within a curriculum. A significant challenge would be to distinguish unintended drawing mistakes from valid variations of the instructions.

Automation of the generation pipeline. While we studied the use of STIVi by learners, we did not evaluate how drawing instructors can benefit from our approach. Our video processing pipeline requires intervention from an expert to adapt the pre-defined vocabulary to specific videos, to correct errors in the stroke and transcript extraction, and to decide on the final assignment between visual and textual elements. Although we expect that these manual corrections require significantly less effort than preparing an interactive tutorial from scratch, we still believe that additional automation could be provided. In particular, distinguishing between singular and plural terms ("this line" vs. "these lines"), and between types of articles ("a line" vs. "the line") could help provide more accurate suggestions of correspondences. Access to a large corpus of annotated drawing videos could also enable data-driven matching.

Generalization. While we demonstrated STIVi on the domain of perspective drawing of block shapes, the five-step methodology outlined in Section 3 could generalize to other domains. Figure 10 provides a mock-up of this methodology applied to drawing the human head. The visual and textual vocabulary includes organic shapes (ellipses, spheres), along with relationships specific to facial proportions (midpoint, symmetry). Instructors often employ these shapes to construct facial regions (eyes, lips, chin), which also appear in the vocabulary (Figure 10c). Extracting such domain-specific visual elements would require more advanced image processing, such as facial landmark detection [52]. Yet, similar to block shapes, heads adhere to perspective rules, and artists make use of convergence towards vanishing points to depict heads from different viewpoints (Figure 10b). Supporting interactive exploration could help students visualize how the relationships between facial landmarks evolve with changes in viewpoint.

9 CONCLUSION

Drawing instructors often comment on their actions as they are performing them, using a vocabulary that refers to the lines they draw, their geometric properties, and their relationships. Our system leverages image and speech processing to relate the commentary of instructional drawing videos to their visual content, offering novel modes of navigation, visualization, and interaction to learners. We hope that progress in natural language processing and sketch recognition will soon allow harvesting the wealth of instructional videos available online, not only to offer more engaging educational content, but also to provide students feedback on the drawings they produce as they follow the augmented instructions.

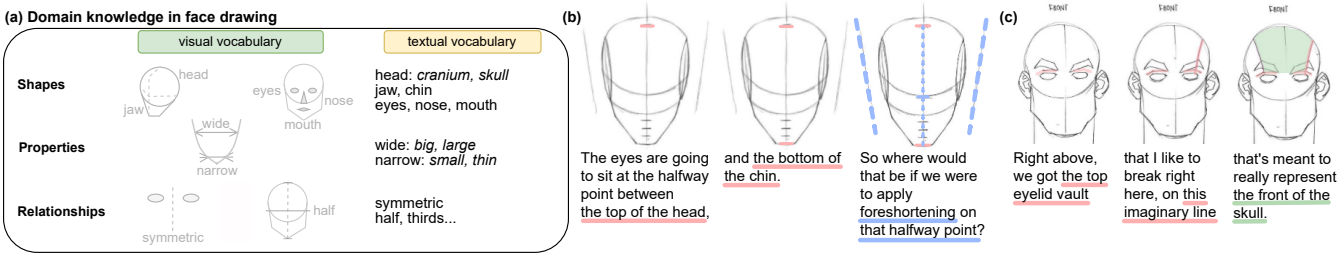

Figure 10: Mock-up illustrating how our approach could apply to face drawing. To this end, we need to define a new visual and textual vocabulary that is adapted to the domain concepts (a). We illustrate how a STIVi-like user interface would highlight these new concepts on two video excerpts (b, c) [7, 24].

References

[1] Angular. `https://angular.io`.

[2] Ngx-Videogular. `https://www.npmjs.com/package/@videogular/ngx-videogular`.

[3] Paper.js. `http://paperjs.org`.

[4] spaCy part-of-speech tagging. `https://spacy.io/usage/linguistic-features`.

[5] VOSK Offline Speech Recognition Toolkit. `https://github.com/alphacep/vosk-api`.

[6] J. Almahmoud, F. Jahanbakhsh, M. Facciotti, M. Igo, K. Sripathi, K. Gal, and D. Karger. Spotlights: Designs for directing learners' attention in a large-scale social annotation platform. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022. doi: 10.1145/3555598

[7] M. Brunet. How to draw simple faces (construction explained). `https://youtu.be/RgCgoYouwkQ`.

[8] R. Chamberlain and J. Wagemans. The genesis of errors in drawing. *Neuroscience & Biobehavioral Reviews*, 65:195–207, 2016.

[9] P.-Y. Chi, S. Ahn, A. Ren, M. Dontcheva, W. Li, and B. Hartmann. *MixT: Automatic Generation of Step-by-Step Mixed Media Tutorials*, p. 93–102. Association for Computing Machinery, New York, NY, USA, 2012.

[10] D. Dixon, M. Prasad, and T. Hammond. Icandraw: Using sketch recognition and corrective feedback to assist a user in drawing human faces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, p. 897–906. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753459

[11] J. Dorsey, S. Xu, G. Smedresman, H. Rushmeier, and L. McMillan. The mental canvas: A tool for conceptual architectural design and analysis. In *Pacific Conference on Computer Graphics and Applications*, 2007.

[12] dr. Draw (Alexander Steenhorst). Understanding Perspective Drawing like Kim Jung Gi. `https://youtu.be/Sgm1oNt7cNw?t=556`.

[13] P. Dragicevic, G. Ramos, J. Bibliowitcz, D. Nowrouzezahrai, R. Balakrishnan, and K. Singh. Video browsing by direct manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, p. 237–246. Association for Computing Machinery, New York, NY, USA, 2008. doi: 10.1145/1357054.1357096

[14] DrawlikeaSir. How to draw faces (from all angles). `https://youtu.be/Uf6mtrcUj_U?si=5iqmL_d4iIqCcu4q`.

[15] K. Eissen and R. Steur. *Sketching: The Basics*. BIS, 2011.

[16] J. Fernquist, T. Grossman, and G. Fitzmaurice. Sketch-sketch revolution: An engaging tutorial system for guided sketching and application learning. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, p. 373–382. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/2047196.2047245

[17] C. A. Fraser, J. O. Kim, H. V. Shin, J. Brandt, and M. Dontcheva. Temporal segmentation of creative live streams. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, 2020.

[18] T. Grossman, J. Matejka, and G. Fitzmaurice. Chronicle: Capture, exploration, and playback of document workflow histories. In *Proc. ACM Symposium on User Interface Software and Technology*, 2010.

[19] Y. Gryaditskaya, F. Hähnlein, C. Liu, A. Sheffer, and A. Bousseau. Lifting freehand concept sketches into 3d. *ACM Transactions on Graphics (SIGGRAPH Asia Conference Proceedings)*, 2020.

[20] Y. Gryaditskaya, M. Sypesteyn, J. W. Hoftijzer, S. Pont, F. Durand, and A. Bousseau. Opensketch: A richly-annotated dataset of product design sketches. *ACM Transactions on Graphics (SIGGRAPH Asia Conference Proceedings)*, 38(6), November 2019.

[21] F. Hähnlein, Y. Gryaditskaya, A. Sheffer, and A. Bousseau. Symmetry-driven 3d reconstruction from concept sketches. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3528233.3530723

[22] J. W. Hennessey, H. Liu, H. Winnemöller, M. Dontcheva, and N. J. Mitra. How2sketch: Generating easy-to-follow tutorials for sketching 3d objects. *Symposium on Interactive 3D Graphics and Games*, 2017.

[23] K. Henry. Folded paper sketch. `https://vimeo.com/378319866`.

[24] HowToDrawComics.NET. How to draw dynamic heads faces in perspective — drawing the top down, front view of the head. `https://youtu.be/srftwJVwdzs`.

[25] D.-A. Huang, S. Buch, L. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[26] E. Iarussi, A. Bousseau, and T. Tsandilas. The Drawing Assistant: Automated Drawing Guidance and Feedback from Photographs. In *ACM Symposium on User Interface Software and Technology (UIST)*. ACM, St Andrews, United Kingdom, Oct. 2013. doi: 10.1145/2501988.2501997

[27] H. Jung, H. V. Shin, and J. Kim. Dynamicslide: Exploring the design space of reference-based interaction techniques for slide-based lecture videos. In *Proceedings of the 2018 Workshop on Multimedia for Accessible Human Computer Interface*, MAHCI'18, p. 33–41. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3264856.3264861

[28] T. Karrer, M. Weiss, E. Lee, and J. Borchers. Dragon: A direct manipulation interface for frame-accurate in-scene video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, p. 247–250. Association for Computing Machinery, New York, NY, USA, 2008. doi: 10.1145/1357054.1357097

[29] S. Keshavabhotla, B. Williford, S. Kumar, E. Hilton, P. Taele, W. Li, J. Linsey, and T. Hammond. Conquering the cube: Learning to sketch primitives in perspective with an intelligent tutoring system. In *Proceedings of the Symposium on Sketch-Based Interfaces and Modeling*, SBIM '17. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3092907.3092911

[30] J. Kim, Y. Choi, M. Kahng, and J. Kim. Fitvid: Responsive and flexible video content adaptation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3501948

[31] J. Kim, P. J. Guo, C. J. Cai, S.-W. D. Li, K. Z. Gajos, and R. C. Miller. Data-driven interaction techniques for improving navigation of educational videos. In *Proc. ACM Symposium on User Interface Software and Technology*, 2014.

[32] J. Kim, P. T. Nguyen, S. Weir, P. J. Guo, R. C. Miller, and K. Z. Gajos.

Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[33] R. L. Kiss. How to draw anything with the help of technical drawing. `https://youtu.be/0a-FQXkJZ1k`.

[34] A. Kozbelt and J. Ostrofsky. *Expertise in Drawing*, pp. 576–596. Cambridge Handbooks in Psychology. Cambridge University Press, 2 ed., 2018. doi: 10.1017/9781316480748.030

[35] S.-J. Lee, J. H. Lee, and S.-H. Bae. An interactive car drawing system with tick'n'draw for training perceptual and perspective drawing skills. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491101.3519776

[36] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. P. Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. In *North American Chapter of the Association for Computational Linguistics*, 2015.

[37] L. McCay-Peet, M. Lalmas, and V. Navalpakkam. On saliency, affect and focused attention. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, p. 541–550. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2207676.2207751

[38] J. Murphy. Vehicle Sketching With Jeremy I: Creating Complex Forms. `https://youtu.be/AXn979hRyIs?t=162`.

[39] C. Nguyen and F. Liu. Making software tutorial video responsive. CHI '15, p. 1565–1568. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2702123.2702209

[40] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala. Video digests: A browsable, skimmable format for informational lecture videos. In *Proc. ACM Symposium on User Interface Software and Technology*, 2014.

[41] S. Robertson. How to Draw: page 084 ortho views. `https://youtu.be/N9hqkyGrNQU`.

[42] S. Robertson and T. Bertling. *How to Draw: drawing and sketching objects and environments from your imagination*. Design Studio Press, 2013.

[43] H. V. Shin, F. Berthouzoz, W. Li, and F. Durand. Visual transcripts: lecture notes from blackboard-style lecture videos. *ACM Transactions on Graphics (TOG)*, 34(6):1–10, 2015.

[44] M. Sypesteyn and J. W. Hoftijzer. Putting it into perspective. `https://vimeo.com/214849073/549bbe3361`.

[45] J. Tan, M. Dvorožňák, D. Sýkora, and Y. Gingold. Decomposing time-lapse paintings into layers. *ACM Transactions on Graphics*, 34(4), 2015.

[46] C. Tarchi, S. Zaccoletti, and L. Mason. Learning from text, video, or subtitles: A comparative analysis. *Computers Education*, 160:104034, 2021. doi: 10.1016/j.compedu.2020.104034

[47] A. Truong, P. Chi, D. Salesin, I. Essa, and M. Agrawala. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445721

[48] B. Wang, M. Y. Yang, and T. Grossman. Soloist: Generating mixed-initiative tutorials from existing guitar instructional videos through audio processing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445162

[49] B. Williford, M. Runyon, and T. Hammond. Recognizing perspective accuracy: An intelligent user interface for assisting novices. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, p. 231–242. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3377325.3377511

[50] B. Williford, P. Taele, T. Nelligan, W. Li, J. Linsey, and T. Hammond. *PerSketchTivity: An Intelligent Pen-Based Educational Application for Design Sketching Instruction*, pp. 115–127. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-31193-7_8

[51] J. Xie, A. Hertzmann, W. Li, and H. Winnemöller. Portraitsketch: Face sketching assistance for novices. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, p. 407–417. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2642918.2647399

[52] J. Yaniv, Y. Newman, and A. Shamir. The face of art: Landmark detection and geometric style in portraits. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 38(4), 2019.

[53] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019.

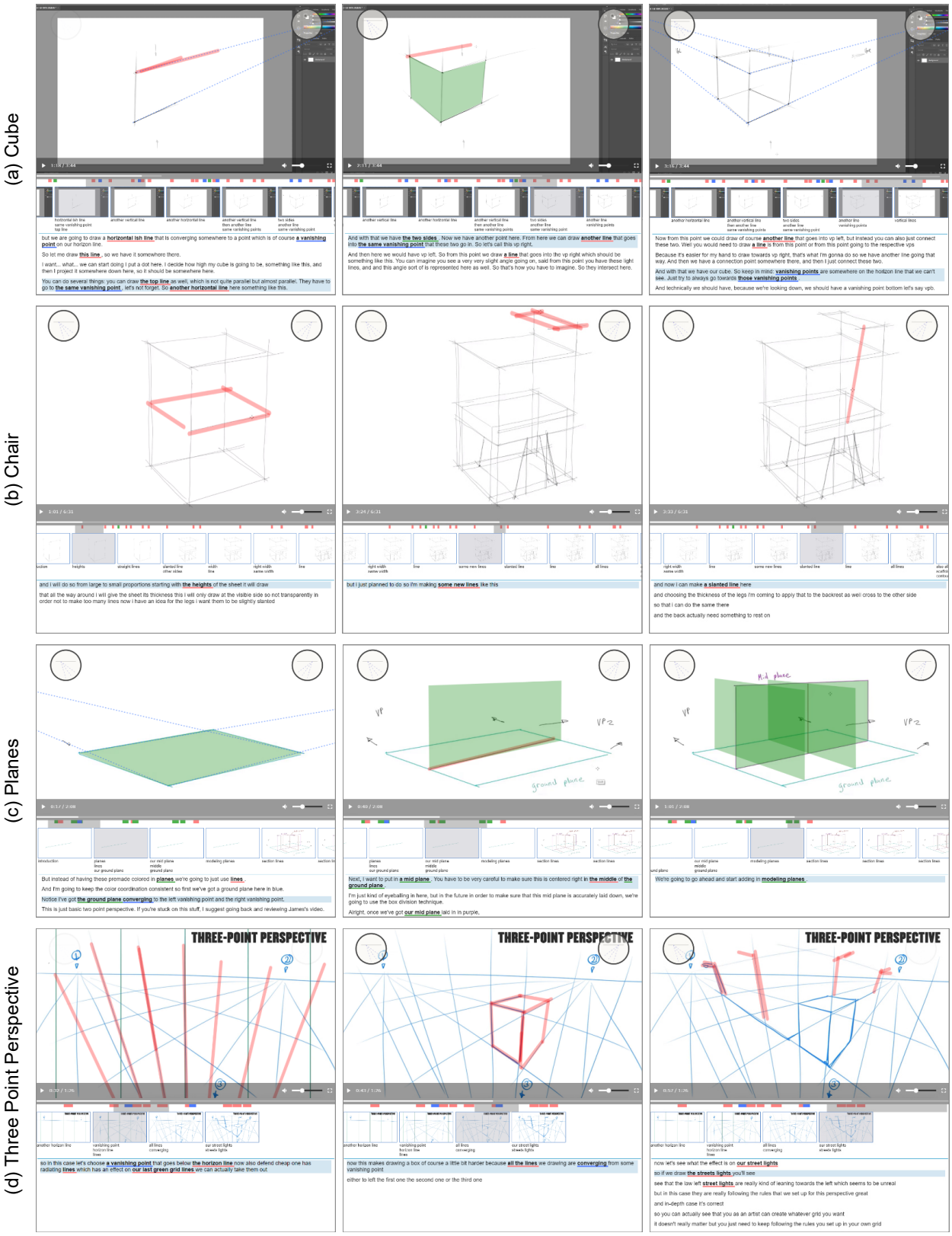

Figure 11: Video gallery. We present a few highlighted frames with transcripts from a selection of four videos. (a) and (b) have been produced by industrial design teachers on our demand. (c) and (d) are two videos which we processed from YouTube [12, 38]