Sketchtopia : Supplementary Material

Mohd Hozaifa Khan
IIIT-Hyderabad
mohd.hozaifa@research.iiit.ac.in

Ravi Kiran Sarvadevabhatla
IIIT-Hyderabad
ravi.kiran@iiit.ac.in

1. Supplementary Metrics Description

In our investigation of collaborative success in goal-oriented games, we utilize a variety of metrics that capture different dimensions of agent performance and interaction. Below, we provide an expanded description of these metrics for understanding and benchmarking the agent behaviour.

Win Rate (WR)

Definition: The Win Rate (WR) is the ratio of successful games, where success is defined as the Guesser correctly identifying the target word.

$$WR = \frac{\text{Number of successful guesses}}{\text{Total number of games}}$$

Interpretation: A higher WR indicates effective collaboration between the Drawer and Guesser. For instance, if out of 100 games, the target word is guessed correctly in 75 cases, the WR would be 0.75 (or 75%). This serves as a baseline for assessing collaborative success, as it directly reflects the agents' ability to work together towards a common goal.

Example: In a series of games where the target word is "apple," a successful guess occurs when the Guesser responds with "apple" after a series of hints from the Drawer.

1.1. Soft Win Rate (SWR)

Definition: The Soft Win Rate (SWR) extends the concept of WR by including cases where the Guesser identifies synonyms of the target word. This captures scenarios where the agents come close to success but may not guess the exact target word.

$$SWR = \frac{\text{Number of “successful" synonym guesses}}{\text{Total number of games}}$$

Interpretation: SWR provides a broader perspective on collaborative success, acknowledging that guessing near-synonyms still demonstrates effective communication and understanding. For example, if the target word is "car," and the Guesser responds with "automobile," this scenario is considered a success.

Example: If in a session involving the target word "happy", the Guesser guesses "joyful", the session would still be considered successful in terms of SWR, reflecting that the Drawer effectively communicated the essence of the word.

1.2. Feedback Responsiveness Score (FRS)

Definition: The Feedback Responsiveness Score (FRS) measures the agents' responsiveness to feedback during gameplay. It is calculated by assessing the proportion of guesses that lead to correct action following a positive ( ) or negative ( ) feedback.

$$FRS = \frac{s}{p+n}$$

where s is the count of responsive guesses, p is the number of positive feedback actions, and n is the number of negative feedback actions.

Interpretation: A higher FRS indicates that the Guesser is effectively adapting their guesses based on the Drawer's feedback, suggesting a dynamic and responsive collaboration. For instance, if the Drawer gives a after a guess of "tree," and the next guess is "oak," this would be counted as responsive. On the other hand, if the response of the guesser was "car" or "missile", it wouldn't be counted as a correct response to given feedback and lead to a lesser FRS value.

Example: If a Drawer gives feedback after three guesses—two thumbs up and one thumbs down—and the Guesser's subsequent guesses align with the feedback, it would demonstrate an effective adaptive strategy.

1.3. Asynchronous Action Overlap (AAO)

Definition: The Asynchronous Action Overlap (AAO) quantifies how much the actions of the Guesser and Drawer overlap in time. It is calculated by the ratio of overlapping actions to total Guesser actions.

$$AAO = \frac{o}{g}$$

where o is the number of overlapping actions and g is the total number of Guesser actions.

Entity	M-Guess	M-Draw	M-👍	M-👎	M-❓
Agent	0.24	0.33	0.25	0.62	0.28

Table 1. Action-level granular MATS score

Interpretation: An AAO close to 0 indicates a structured, turn-based interaction, whereas a value closer to 1 suggests a more chaotic, uncoordinated interaction. A non-zero AAO might indicate a balance where both agents are communicating in an asynchronous manner. Determining an ideal range might be difficult. However, AAO values from human games would be a good estimate.

Example: In a game session, if the Drawer is sketching while the Guesser is actively guessing, the overlapping timestamps contribute positively to the AAO. If the Guesser waits until the Drawer finishes sketching before guessing, the AAO would be lower.

1.4. Multimodal Action Timing Similarity (MATS)

Definition: The Multimodal Action Timing Similarity (MATS) assesses the temporal similarity of human actions and the agent actions. It is computed using the Jensen-Shannon divergence between the Kernel Density estimates of correspoonding action timestamps.

$$d_{action} = \text{Jensen-Shannon}\{\mathtt{KDE}(T^{H}_{action}), \mathtt{KDE}(T^{A}_{action})\}$$

where, action = [guess, draw, 👍, 👎, ❓].

MATS score represents average JS distance over all actions. Individual action divergence can be found in Tab. 1.

Interpretation: A lower MATS score indicates that the Agent's pacing closely resembles that of human pace, suggesting effective modelling of human-like action timings. Conversely, a high score may indicate that the agent's actions are poorly timed compared to typical human behavior.

1.5. Evaluating sketch quality: CMMD

We evaluated sketch quality using the CMMD score [10], a robust alternative to FID based on CLIP embeddings and MMD (Maximum Mean Discrepancy) distance. For each target word, we considered the final canvas across all games for simplicity. We calculated CMMD scores comparing DRAWBOT's final canvas to human sketches (H-A). For the human (H-H) baseline, we split human sketches into two equal partitions and computed CMMD scores between them. As a reference for high dissimilarity with sketches, we computed H-I, the CMMD between human sketches and randomly chosen color images. CMMD scores averaged across target words are presented in Table 2. Though DRAWBOT's CMMD score (2.8) is higher than the human baseline, it is more distant from the H-I score, suggesting reasonable

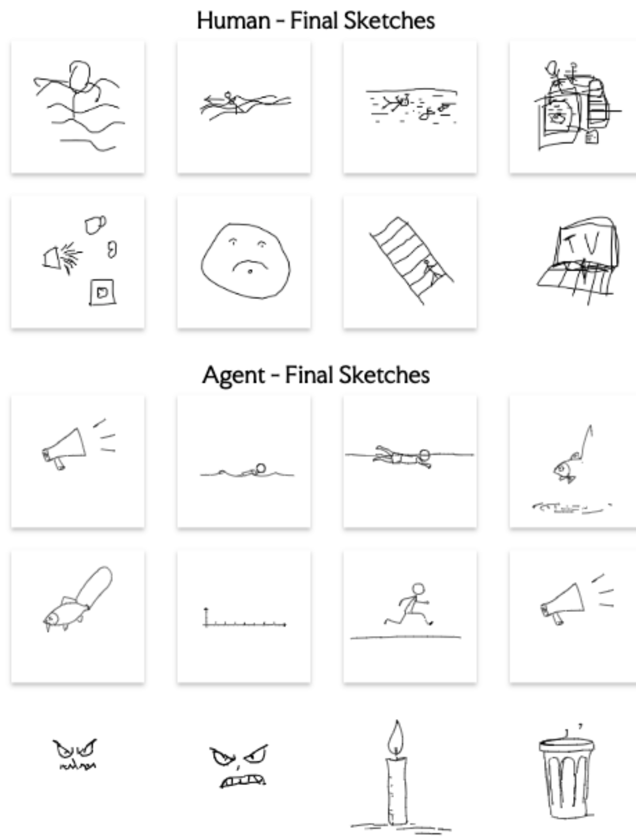


Figure 1. Sample final sketches by human players (top) and agent (bottom). Notice that the sketches by agents are relatively simpler, symmetric and have lesser stroke density.

Metric	H-H_{min}	H-H_{max}	**H-A**
CMMD	0.15	2.8	6.9

Table 2. Quantitative evaluation for comparing DRAWBOT sketches to human sketches. Lower scores indicate greater similarity.

human interpretability of DRAWBOT's sketches, supporting the qualitative observations in Figure 8.

2. VLMs

Tab. 3 shows some responses by well-known VLMs such as Cog-VLM and Mini-CPM-2.5 for the final sketch input, for the prompt: "You're playing Pictionary. Guess the target word being communicated from the last drawing, and give your guess". **It is important to note that these outputs are turn-based. Also, there is no asynchronous guessing.**

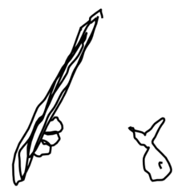
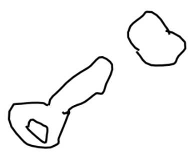
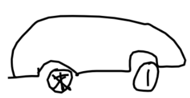
Target Word	Sample Canvas	VLM Output
fish		The drawing appears to be a simple sketch of a fish. It has a rounded head, a long, slender body, and fins on its back and tail.
key		The drawing represents a key.
wheel		Based on the drawing, the word being communicated is likely car. The drawing shows a simple representation of a car with four wheels and a basic body shape.
bullet		The drawing depicts a gun with a bullet flying out of it and a bullet casing on the ground. Therefore, the word being communicated is likely gun or shoot.
earth		The drawing seems to depict a person standing near a tree with a hole in the ground. The hole has a question mark above it. Given this context, the word being communicated could be a mystery or a puzzle.

Table 3. VLM outputs as guesser.

Word	POS	Word	POS	Word	POS
airplane	noun	ambulance	noun	angry	adjective
ant	noun	ape	noun	arrest	verb
arrow	noun	badminton	noun	baseball	noun
basket	noun	basketball	noun	bear	noun
bed	noun	bee	noun	beetle	noun
bench	noun	bend	verb	bicycle	noun
bird	noun	boat	noun	book	noun
brick	noun	brush	noun	bucket	noun
bullet	noun	burn	verb	bury	verb
butterfly	noun	buy	verb	calculator	noun
calendar	noun	call	verb	camera	noun
candle	noun	car	noun	carrot	noun
cat	noun	catch	verb	chair	noun
cigarette	noun	clap	verb	clean	verb
climb	verb	clock	noun	close	verb
comb	noun	cook	verb	cow	noun
crocodile	noun	cry	verb	cut	verb
deer	noun	dig	verb	dive	verb
door	noun	download	verb	dragonfly	noun
draw	verb	drink	verb	drive	verb
dustbin	noun	earth	noun	eat	verb
elephant	noun	email	noun	empty	adjective
enter	verb	envelope	noun	eraser	noun
exit	verb	face	noun	far	adjective
feather	noun	feet	noun	fight	verb
finger	noun	fingerprints	noun	fire	noun
fish	noun	fishing	verb	flame	noun
flute	noun	fly	verb	forest	noun
fork	noun	frog	noun	fruit	noun
giraffe	noun	goldfish	noun	grape	noun
grass	noun	gym	noun	hammer	noun
hang	verb	happy	adjective	hard	adjective
helicopter	noun	hit	verb	hollow	adjective
igloo	noun	jacket	noun	key	noun
keyboard	noun	kiss	verb	kneel	verb
knife	noun	knit	verb	knock	verb
ladder	noun	lake	noun	lantern	noun
laptop	noun	lazy	adjective	length	noun
listen	verb	lizard	noun	loud	adjective
loudspeaker	noun	love	noun	microphone	noun
microwave	noun	monkey	noun	mouse	noun
mug	noun	needle	noun	notebook	noun
oven	noun	paint	verb	paintbrush	noun
panda	noun	pant	verb	parachute	noun
passport	noun	peacock	noun	pear	noun
pencil	noun	pillow	noun	pizza	noun
plant	noun	postcard	noun	potato	noun
pray	verb	pull	verb	puppet	noun
puppy	noun	radio	noun	rain	noun
rainy	adjective	recycle	verb	rifle	noun
roof	noun	rose	noun	run	verb
sad	adjective	saturn	noun	scary	adjective
scissor	noun	scream	verb	screwdriver	noun
selfie	noun	shark	noun	shoe	noun
sing	verb	sit	verb	skate	verb
sleep	verb	slide	verb	smile	verb
snake	noun	spider	noun	spiderman	noun
spoon	noun	stairs	noun	starfish	noun
stove	noun	suitcase	noun	sun	noun
sunny	adjective	sweater	noun	swim	verb
sword	noun	table	noun	tank	noun
tattoo	noun	teaspoon	noun	telephone	noun
television	noun	tent	noun	throne	noun
toaster	noun	tomato	noun	toothbrush	noun
toothpaste	noun	tree	noun	truck	noun
van	noun	walk	verb	wheel	noun
windmill	noun	zip	verb		

Table 4. Target Words with Parts of Speech

Approaches	Shared Goal	Drawer	Guesser	Visual Content		Draw Content	Guess Content	Feedback (Multi-modal)	Content Editing	Async.
				Cont.	Grd.					
Sketchtopia (Ours)	✓	✓	✓	✓	✓	**Freehand**	**Phrase**	✓(👍, 👎, ❓)	✓	✓
Iconary [2]	✓	✓	✓	✗	✓	Icons (text)	Sent.	✗	✗	✗
Co-Draw [11]	✓	✓	✓	✗	✓	ClipArt (text)	Sent.	✗	✗	✗
Tell, Draw, Repeat [6]	✗	✓	✗	✓	✓	Image	Sent.	✗	✗	✗
Scones [9]	✗	✓	✗	✓	✓	Freehand	Sent.	✗	✗	✗
Pixelor [1]	✗	✓	✗	✓	✗	Sketches	✗	✗	✗	✗
Visual QA	✗	✗	✓	✓	✓	✗	Sent.	✗	✗	✗
SketchGuess[13]	✗	✗	✓	✓	✓	✗	Phrase	✗	✗	✗
Vis. Dialogue Agents [3]	✗	✗	✓	✓	✓	✗	Sent.	✗	✗	✗
Coop. Dialogue Agents [4]	✓	✗	✓	✓	✓	✗	Sent.	✗	✗	✗
Embodied VQA [4]	✓	✗	✓	✓	✓	✗	Sent.	✗	✗	✗

Table 5. **Agents Comparison:** A comparative summary of agent capabilities. Cont.: Continuous sketching; Grd.: Grounding of sketch content; Draw/Guess Content: Agent's output format; Async.: Asynchronous communication support; Sent.: Sentence.

Works	# Sketch	Pr. Mod.	Lang. Ground.		Form	Erase	Strokes	Cat.
			IntG.	Final				
Sketchtopia (Ours)	263K	Txt, Fdb	✓	✓	Scene	✓	✓	200
TU-Berlin [5]	20K	✗	✗	✓	Obj	✗	✓	250
Quick Draw [8]	50M	✗	✗	✓	Obj	✗	✓	345
Sketchy [12]	75K	✗	✗	✓	Obj	✗	✓	125
SketchyScene [17]	29K	Ph	✗	✓	Sc	✗	✗	-
ImageNet-Sketch [14]	50K	✗	✗	✓	Obj	✗	✗	1000
Scene Sketcher	1225	Ph	✗	✓	Obj,Sc	✗	✓	14
SketchyCOCO [7]	14K	Ph, EM	✗	✓	Obj,Sc	✗	✓	17
QUML-Shoe [15]	419	Ph	✗	✓	Obj	✗	✗	1
TU-Berlin Extended [16]	20K	Ph	✗	✓	Obj	✗	✗	250
SketchGuess-160 [13]	8960	Txt	✓	✓	Obj	✗	✓	160

Table 6. **Sketch Dataset Comparison:**; Img.: Image; Pr. Mod.: Other modalities; Lang. Ground.: Language Grounding; IntG: Intermediate Guesses;Final: Final guess;Txt: Text; Fdb: Feedback;Sc: Scene; Obj: Object; Cat: Categories; Ph: Photo; EM: Edge Map

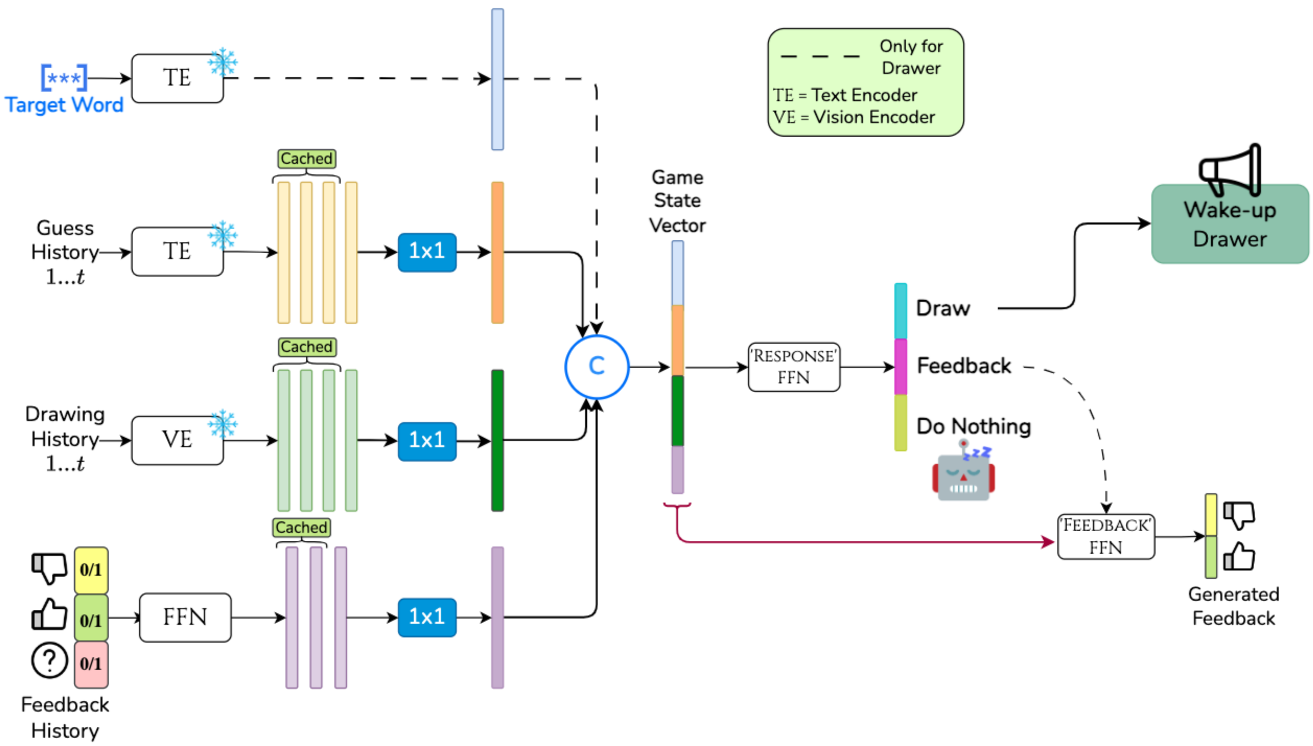


Figure 2. Schematic Diagram of the **ActionDecider Module** for decision timing and action selection. Refer **??**.

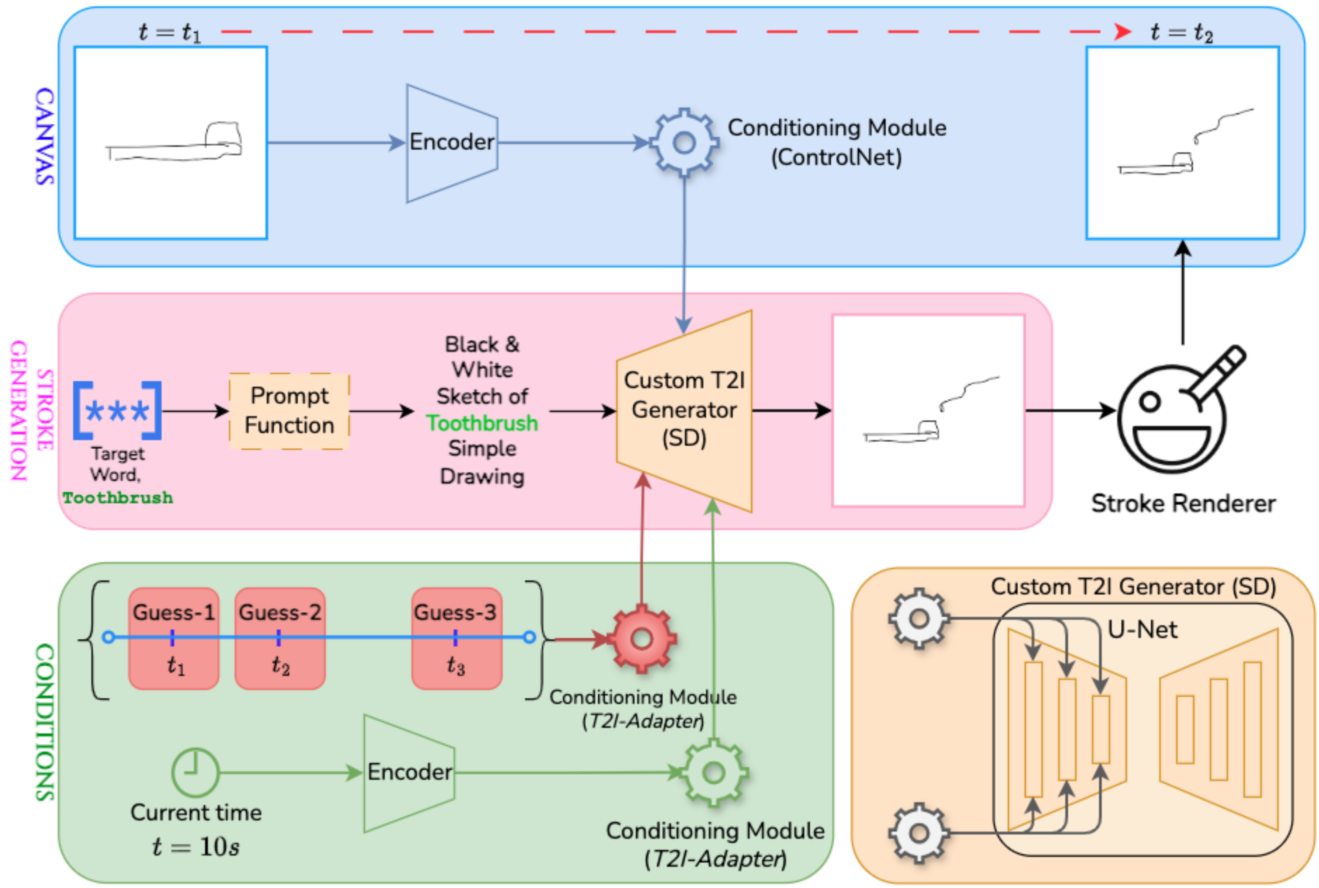


Figure 3. **Drawer Agent (DrawBot) Architecture:** The top left blue box shows the current state of the canvas. The pink module illustrates the primary Text-to-image-based generation module, which processes the target word (e.g., 'toothbrush') to produce strokes. The green module represents additional conditioning inputs — such as prior guesses and current time — that guide the generative process, enabling the agent to create iterative strokes towards a recognizable sketch. The bottom right orange box shows that conditioning outputs are passed to the encoder of the T2I generator.

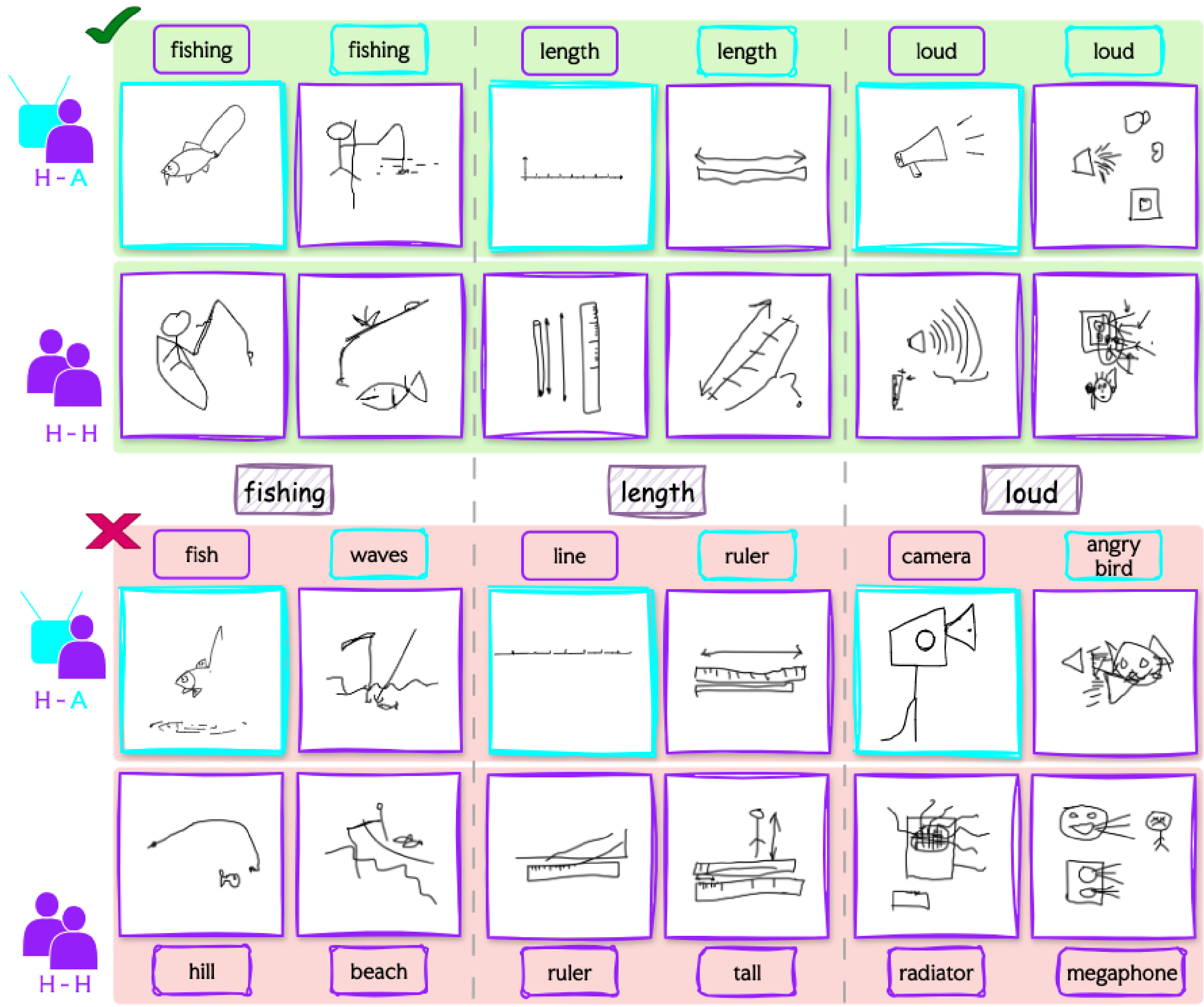


Figure 4. **Qualitative comparison between human and agent game outcomes:** The figure shows the final canvas and the final guess for select target words (`fishing`, `length`, `loud`). Human outputs are in purple and agent outputs are in blue. Successful games are with green background and unsuccessful ones in red.

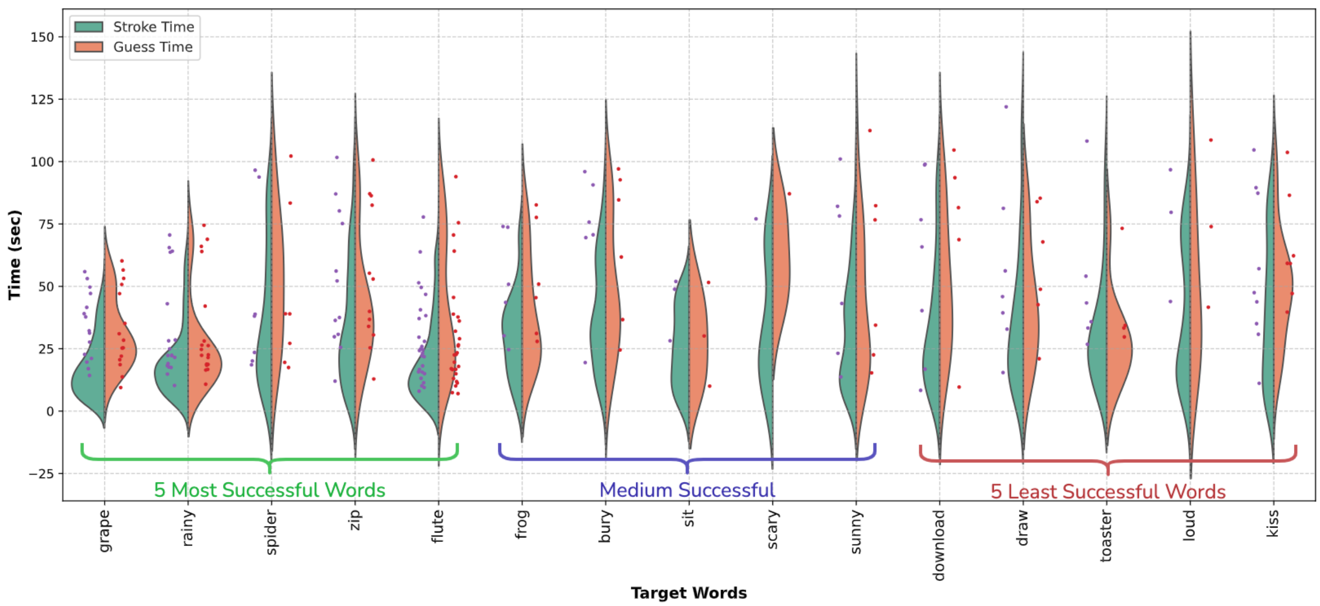


Figure 5. Game Distribution: Drawer and Guesser actions distributed across time; sorted in decreasing order of success rate. For successful games, drawer actions preceded the guesser actions and usually last a shorter duration

References

[1] Ayan Kumar Bhunia, Ayan Das, Umar Riaz Muhammad, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Pixelor: a competitive sketching ai agent. so you think you can sketch? *ACM Trans. Graph.*, 39(6), 2020. 5

[2] Christopher Clark, Jordi Salvador, Dustin Schwenk, Derrick Bonafilia, Mark Yatskar, Eric Kolve, Alvaro Herrasti, Jonghyun Choi, Sachin Mehta, Sam Skjonsberg, Carissa Schoenick, Aaron Sarnat, Hannaneh Hajishirzi, Aniruddha Kembhavi, Oren Etzioni, and Ali Farhadi. Iconary: A pictionary-based game for testing multimodal communication with drawings and text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1864–1886, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 5

[3] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[4] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[5] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31 (4):44:1–44:10, 2012. 5

[6] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W. Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[7] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5174–5183, 2020. 5

[8] David Ha and Douglas Eck. A neural representation of sketch drawings. In *ICLR 2018*, 2018. 2018. 5

[9] Forrest Huang, Eldon Schoop, David Ha, and John Canny. Scones: towards conversational authoring of sketches. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, page 313–323, New York, NY, USA, 2020. Association for Computing Machinery. 5

[10] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9307–9315, 2023. 2

[11] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy, 2019. Association for Computational Linguistics. 5

[12] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 35(4), 2016. 5

[13] Ravi Kiran Sarvadevabhatla, Shiv Surya, Trisha Mittal, and R. Venkatesh Babu. Pictionary-style word guessing on hand-drawn object sketches: Dataset, analysis and deep network models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):221–231, 2020. 5

[14] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 5

[15] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[16] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1105–1113, 2016. 5

[17] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. Sketchyscene: Richly-annotated scene sketches. In *ECCV*, pages 438–454. Springer International Publishing, 2018. 5