Multimedia Generative Script Learning for Task Planning

Qingyun Wang¹, Manling Li¹, Hou Pong Chan², Lifu Huang^{3,} Julia Hockenmaier¹, Girish Chowdhary¹, Heng Ji¹

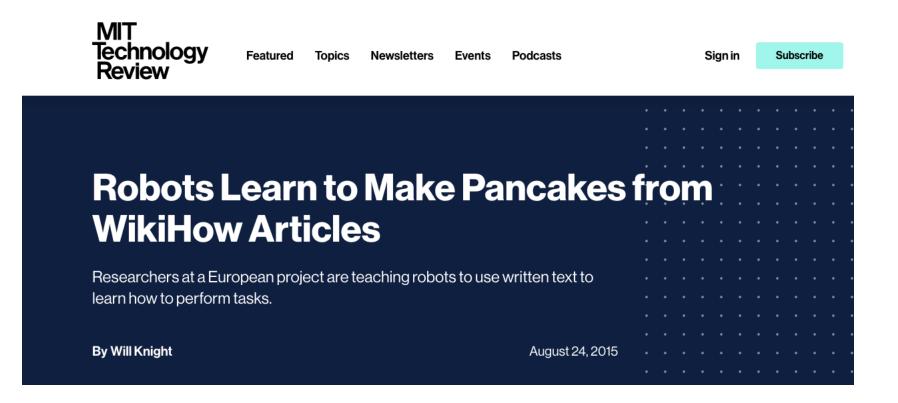
¹University of Illinois at Urbana-Champaign ²University of Macau ³Viginia Tech



BLENDER | Cross-source Information Extraction



Goal-oriented script learning helps robots perform daily activities



Previous goal-oriented generative script learning tasks

Multilingual T5^[1]

- Input
 - Obtain Travel Documents
- Output
 - 1. Authority grants approval for travel documents.
 - 2. Applicant takes a photograph if that is required to receive travel document.
 - 3. Travel documents approved and granted to traveler.
 - 4. Get and/or fill out all documents that must be submitted to apply for the travel document.
 - 5. Traveler pays fees and submits application documents.
 - 6. Applicant receives letter, passport, visa stamp, or whatever travel documents were requested.
 - 7. Item is delivered to the destination.
 - 8. Prepare required application documents.
 - 9. Pay any required application fees.

Translated GPT-3 175B^[2]

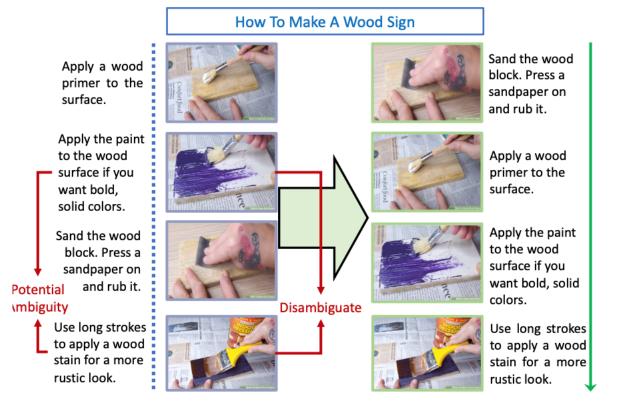
- Input
 - Throw away paper
- Output
 - 1. Walk to home office
 - 2. Walk to table
 - 3. Find table
 - 4. Turn to table
 - 5. Find paper
 - 6. Grab paper
 - 7. Walk to trashcan
 - 8. Open trashcan
 - 9. Put paper on trashcan
 - 10. Close trashcan
 - Generate responses for unseen tasks
 - Y Cannot capture visual state

[1] Lyu, Q., Zhang, L., & Callison-Burch, C. (2021). Goal-oriented script construction. INLG 2021.

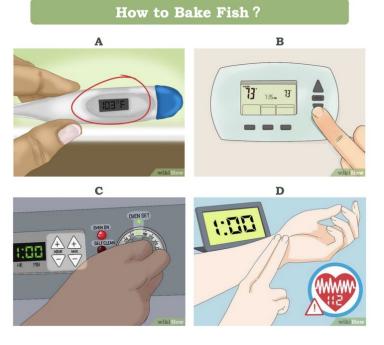
[2] Huang, W., Abbeel, P., Pathak, D., & Mordatch, I. (2022). Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. arXiv preprint arXiv:2201.07207.

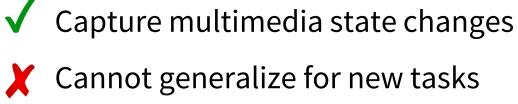
Previous multimedia script learning tasks

Multimodal procedure sequencing^[1]



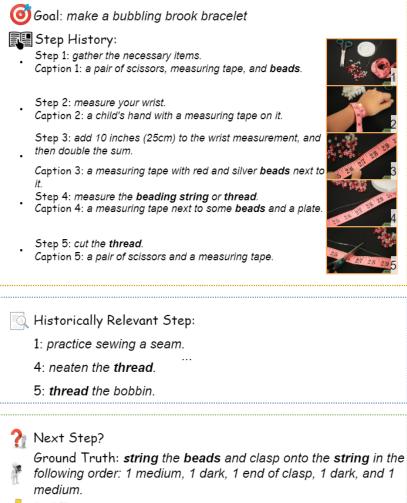
Visual Goal-Step Inference^[2]





[1] Wu, T. L., Spangher, A., Alipoormolabashi, P., Freedman, M., Weischedel, R., & Peng, N. (2022, May). Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals.. ACL 2022. [2] Yang, Y., Panagopoulou, A., Lyu, Q., Zhang, L., Yatskar, M., & Callison-Burch, C. (2021). Visual goal-step inference using wikihow. EMNLP 2021.

Multimedia Generative Script Learning



BART: measure the length of the bead.

Our Model: thread the beads onto the thread.

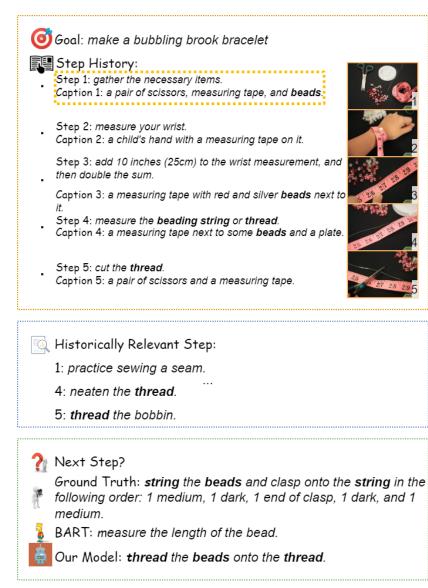


Domain	Split	# Task	# Pair	Avg # Step	Step Len
Gardening	Train	1,857	20,258	3.10	11.6
	Valid	237	2,428	3.03	10.6
	Test	238	2,684	2.88	11.2
Crafts	Train	2,654	32,082	6.06	8.98
	Valid	3,33	4,061	6.12	9.10
	Test	3,33	3,937	5.91	9.00

✓ Generate responses for unseen tasks

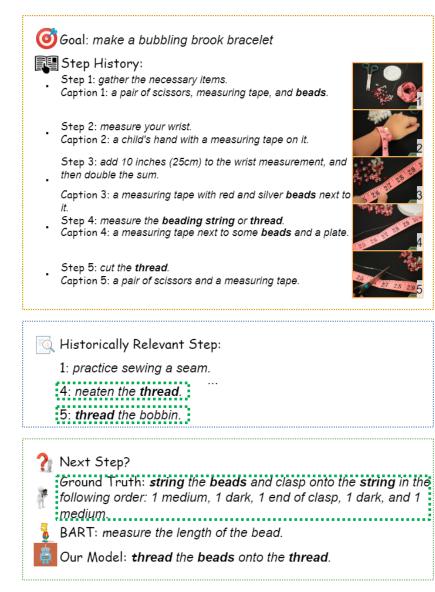
Capture multimedia state changes

A Good Script is Visual-state trackable



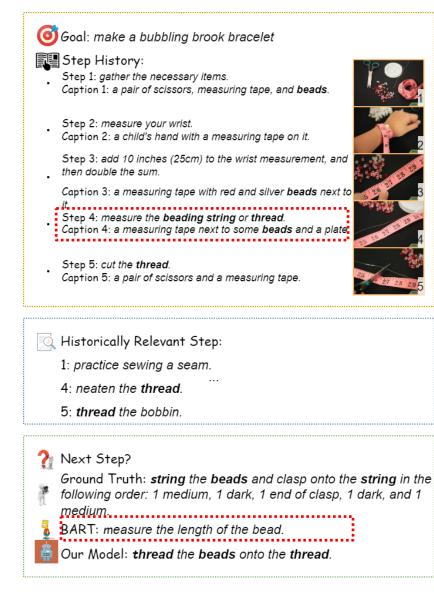
- Captions are abstract summaries of the image's salient objects
- Selection gate helps focus on selected captions and steps that are related to the future step

A Good Script is Inductive



 Previously observed similar tasks in that scenarios served as exemplars

A Good Script is **Diverse**



- Pretrained transformer-based language models tend to repeat input or generate similar future steps for different tasks
- Diversity-oriented contrastive learning objectives help the model reduce repetition

Overview

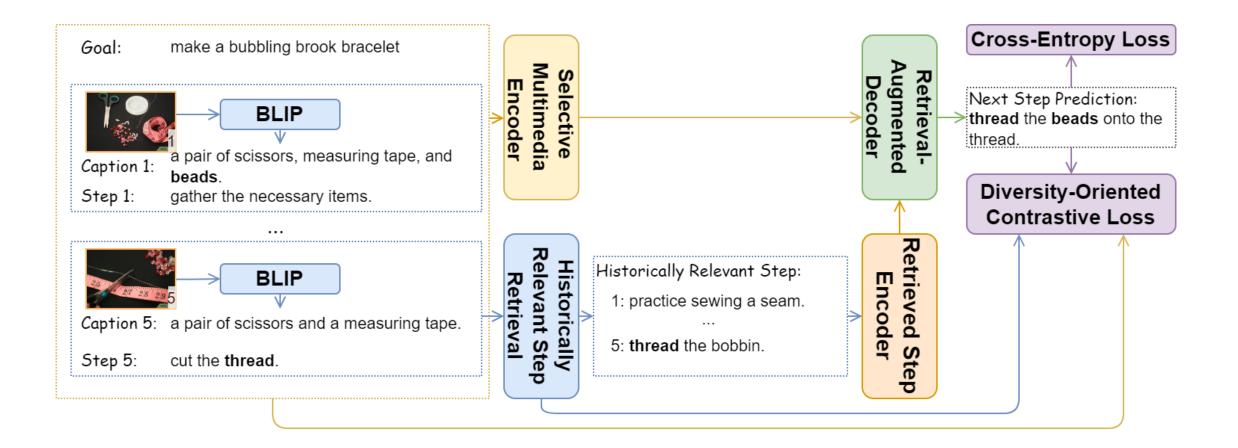
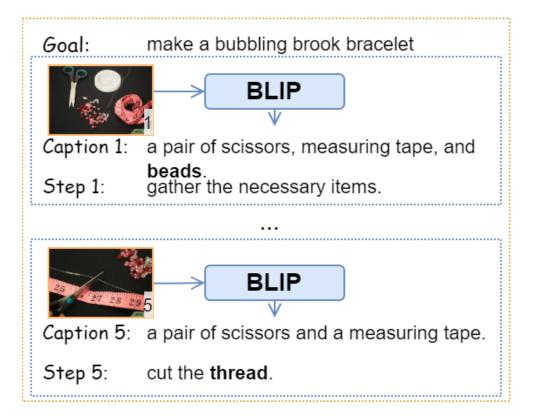
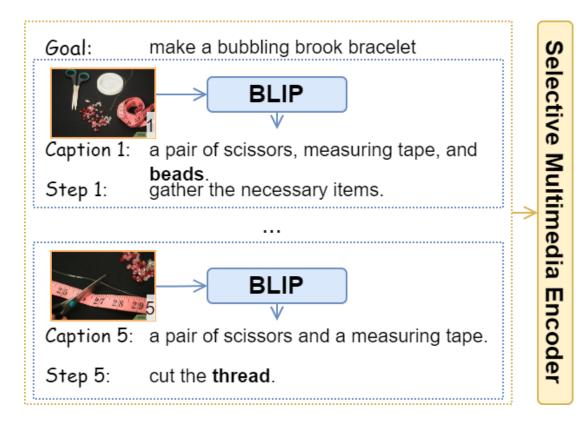


Image Encoding



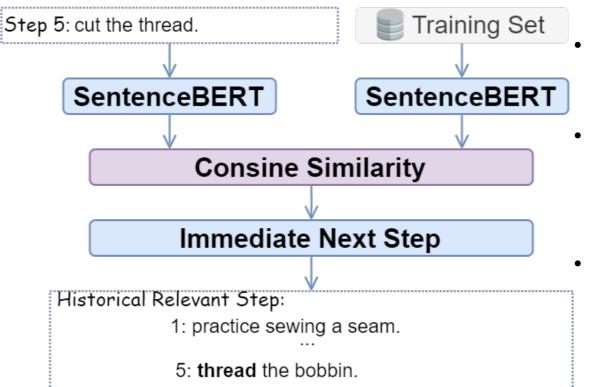
 Use BLIP^[2], a large-scale vision-andlanguage corpus with 129M images, to generate captions for each image in the step history

Selective Multimedia Encoder



• Mask out steps or captions not directly relevant to the future steps

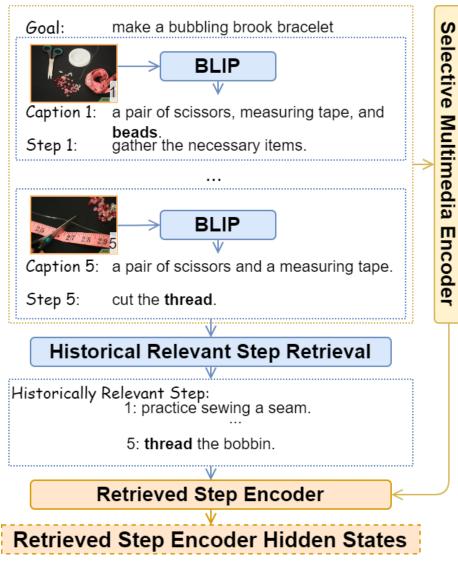
Historical Relevant Step Retrieval



Use SentenceBERT^[2] to obtain representations for each step

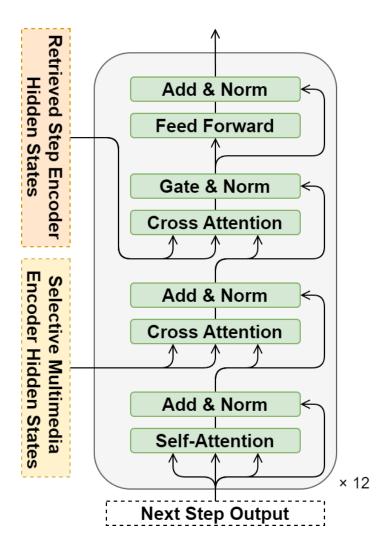
- Retrieve k steps which have the top-k similarity with the representation of the previous step from the training set
- Consider the immediate next step for those k steps as potential relevant steps

Retrieved Step Encoder



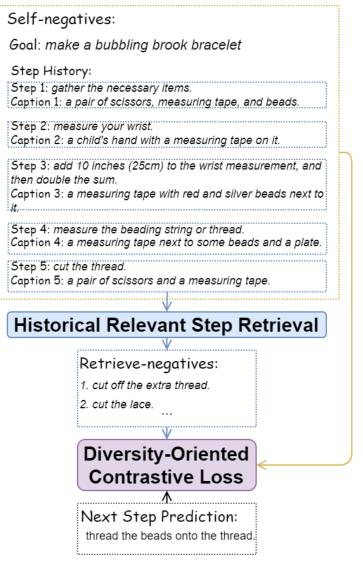
 Mask out retrieved steps not directly relevant to the future steps similar to selective multimedia encoder

Retrieval-Augmented Decoder



- Capture the most relevant historical step based on the previous decoder hidden states
 - Fuse the retrieved step encoder's hidden states with multi-head attention
 - Combine knowledge with a control gate
 - Use cross-entropy loss to optimize the next step generation

Diversity-Oriented Contrastive Learning



- Sample negatives from step history and retrieved steps from training corpus
- Given next step S₀ and K negative samples {S₁, S₂, ..., S_K}, compute InfoNCE loss

Evaluations

- Automatic evaluation
 - Generation quality evaluation
 - Inductive quality evaluation
 - Multimodal-Retrieval based metric
 - Diversity evaluation
- Human evaluation

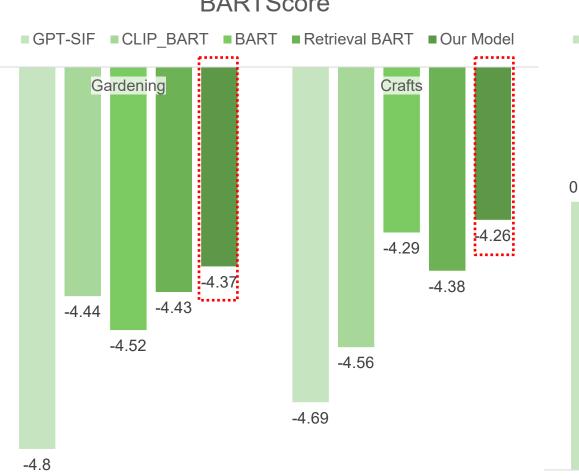
Baselines

- State-of-the-art pretrained text only generation models
- Retrieval baselines
- Multimodal generation baselines
- Ablation baselines
 - CAP is models with caption input
 - *ME* is models with multimedia selective encoder
 - *RD* is models with historical relevant step encoder and retrieval-augment decoder
 - *CL* is models with diversity-oriented contrastive learning

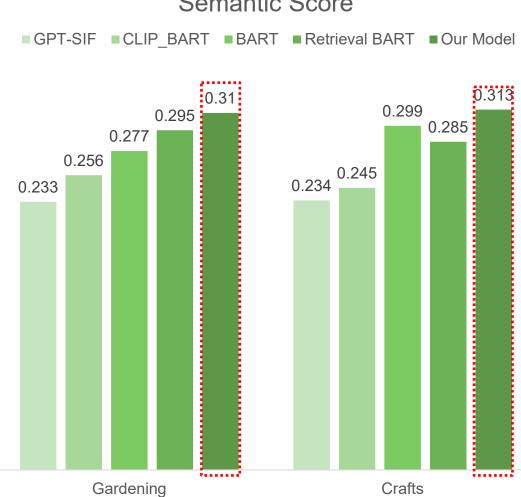
Generation Quality Results



Inductive Quality Results

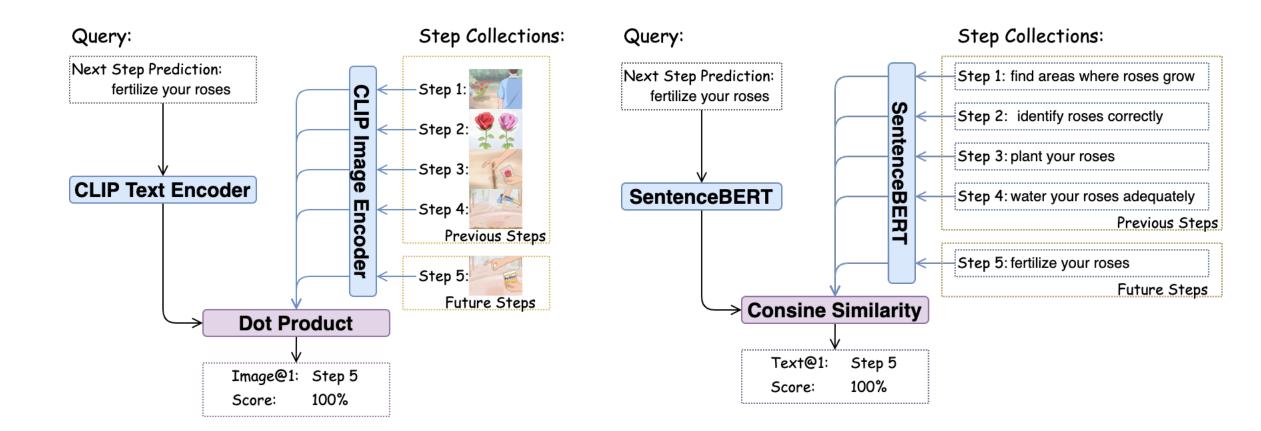


BARTScore

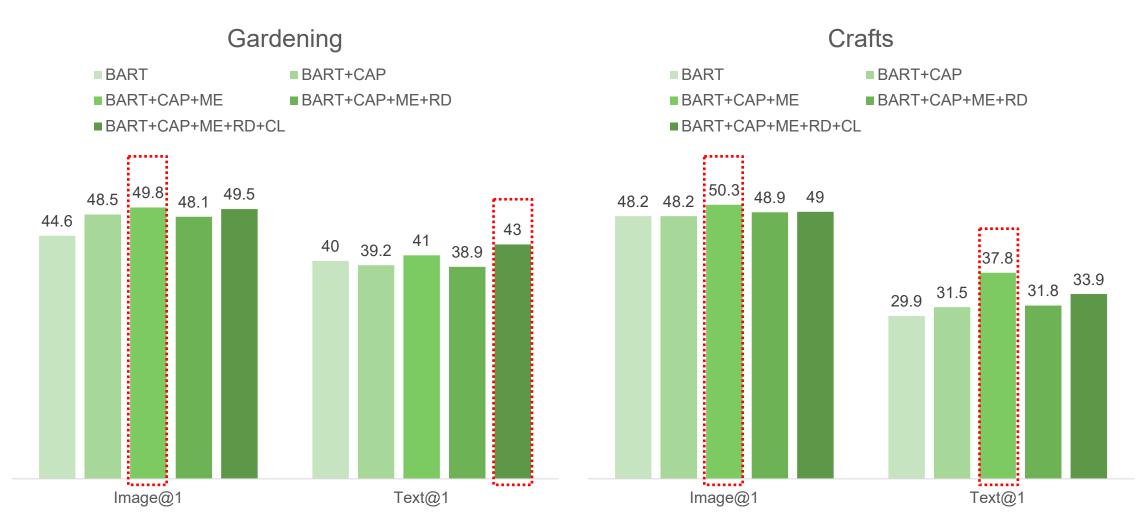


Semantic Score

Multimodal-Retrieval based Metrics



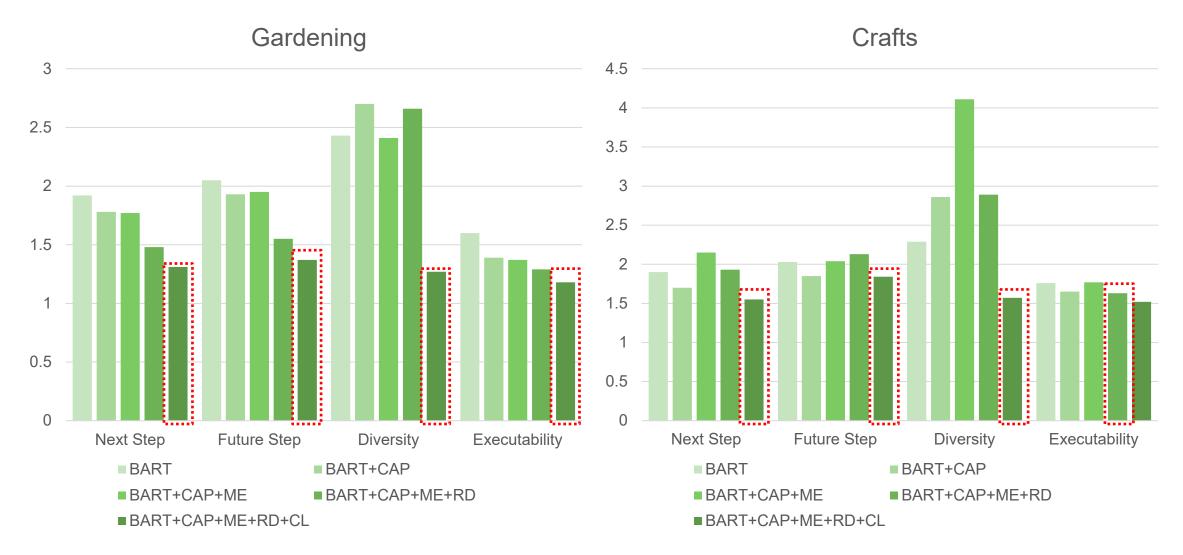
Multimodal-Retrieval results



Human Evaluation

- Four proficient English speakers to independently rank the 41 (gardening)/44 (crafts) generated steps from 1 (best) to 5 (worst)
 - *next step correctness* which measures whether the generated results match the next step
 - future steps correctness measuring whether the generated results match any of the future steps
 - *diversity* which measures the diversity of generated results under the same subgoal
 - *Executability* which checks the generated results repeat or conflict with step history

Human Evaluation



Impact of Selective Multimedia Encoder

oal: cure azaleas of leaf gall

Step History:

Step 1: identify your shrub as an azalea. Caption 1: a pink flower with green leaves on a blue background

Caption 2: a green leaf with white dots on it



Step 2: rule out other diseases

🔍 Historical Relevant Step:

1: look for signs of pests.

- 2: give your plants just the right amount of sun.
- 3: look for insect activity.
- 4: harvest spring onions after 8 weeks.
- 5: use cultural control.

?? Next Step?

Ground Truth: remove infected leaves.

Future Steps: destroy the infected pieces away from the plant.

BART: keep your shrub healthy.

BART+CAP: **remove the leaf gall**.a person holding a green leaf in their hand.

BART+CAP+ME: **remove the leaf gall** from the plant. BART +CAP+ME+RD: **remove the leaf gall**.a person cutting a plant with scissors.

Our	N

Dur Model: remove the leaf gall from the shrub.

Impact of Retrieval Augmentation

- Goal: divide gerbera daisies Subgoal: separating your flowers
- Step History:
 - . Step 1: prepare your work area.
 - Caption 1: a person pouring water into two buckets
 - Step 2: use a shovel to dig up the entire plant. Caption 2: a plant with pink flowers growing out of it



🔍 Historical Relevant Step:

- 1: trim down the stems and pick off the leaves.
- 2: cut the plant using a spade or a sharp knife.
- 3: comb through the roots to remove as much soil as you can.
- 4: store your lettuce in a bag in the fridge.
- 5: shake the plant to remove excess dirt from the roots.

- Next Step?
- Ground Truth: **separate** the individual sections with your hands.

Future Steps: *remove* all flower stems, buds, dead roots, and old or discolored leaves.

BART: dig a hole that is twice the size of the root ball.

BART+CAP: dig up the entire plant.

BART+CAP+ME: dig up the entire plant.

BART +CAP+ME+RD: separate the daisies.

🚔 Our Model: *remove* the entire plant from its container.

Impact of Contrastive Learning

- Goal: harvest creeping charlie
 - Subgoal: picking creeping charlie plants
- Step History:
 - Step 1: figure out the right time of year to harvest.



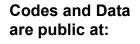
- Caption 1: a set of four different colors of leaves
- Kistorical Relevant Step:
 - 1: put on protective gear.
 - 2: harvest daily when the spears are about 6-8 inches (15.24-20.32 cm) high.
 - 3: grow larger plants.
 - 4: select only the bright red berries.
 - 5: cut the squash from the vines..

Next Step?

- Ground Truth: *pick your creeping charlie leaves*.
 - Future Steps: store the leaves in a jar or similar container.
- BART: use a sharp knife to cut the leaves.
 - BART+CAP: choose the right plant.
 - BART+CAP+ME: choose the right plant.
 - BART +CAP+ME+RD: choose the right variety of creeping charlie.
- A Our Model: *pick creeping charlie plants from the ground.*

Conclusion

- Propose the first *multimedia generative script learning task* which combines illustrative images tracking visual state changes with step history
- Propose a novel approach to produce visually trackable, inductive, and diverse scripts, through a selective multimedia encoder, a retrieval augmented decoder, and a diversity-oriented contrastive learning objective
- Propose a new *multimodal-retrieval based metric* to evaluate the cross-modal semantic similarity and the inductive ability by checking factual correctness





Thank you!

Codes and Data are public at:



BLENDER | Cross-source Information Extraction

Τ

Lab