

Can We Generate Visual Programs Without Prompting LLMs?

Michal Shlapentokh-Rothman Yu-Xiong Wang Derek Hoiem

University of Illinois at Urbana-Champaign
{michal5, yxw, dhoiem}@illinois.edu

Abstract

Visual programming prompts LLMs (large language models) to generate executable code for visual tasks like visual question answering (VQA). Prompt-based methods are difficult to improve while also being unreliable and costly in both time and money. Our goal is to develop an efficient visual programming system without 1) using prompt-based LLMs at inference time and 2) a large set of program and answer annotations. We develop a synthetic data augmentation approach and alternative program generation method based on decoupling programs into higher-level skills called templates and the corresponding arguments. Our results show that with data augmentation, prompt-free smaller LLMs ($\approx 1B$ parameters) are competitive with state-of-the-art models with the added benefit of much faster inference.

1. Introduction

Visual programming [11, 32, 33] is program generation that invokes visual models to solve tasks such as answering questions about images, typically by prompting a very large language model, such as GPT [1] or Llama [35]. Prompting LLMs has several disadvantages (see Figure 1): costly and long inference time, unreliable due to prompt sensitivity [29] and limited improvement without high training cost [31]. We investigate whether we can generate visual programs **without** prompting LLMs during inference.

Our goal is to create an efficient visual programming system with the following characteristics:

1. **Fast Inference Time and Low Cost** Our system should be able to generate thousands of programs cheaply.
2. **Prompt-Free** The only input to our system at inference time is the question.
3. **Low-Cost Improvement with Coarse Annotations** Our system should be able to improve given question/answer pairs without needing expensive annotations.

Given how data intensive LLMs are, one might think that creating such a prompt-free system would require large amounts of annotated data (either with programs or an-

swers). Surprisingly, our methods and results show that efficiency can be achieved in both data and speed.

Our key insight in achieving such capabilities lies in decoupling the skill or procedure from the question-specific concept. We call the higher-level skills *templates* and the concepts *arguments*. For example, the programs “Count the red chairs” and “Count the green bananas” have the same template but different arguments. This facilitates creating synthetic examples by replacing arguments in the question and program. We can then learn to generate programs, most simply by fine-tuning a smaller LLM, or by learning to predict the template and its arguments from a question. We show that, compared to retrieval augmented [18] prompt-based generation, this approach leads to more improvement with simple feedback and much more efficient inference, enabling larger scale testing and reduced processing costs.

To validate our approach, we evaluate our method on frequently used visual question answering (VQA) datasets and compare with prompt-based LLM methods. We analyze different aspects of our method including performance, cost (training and inference) and efficiency. In addition, we study the relative impact of different types of annotations (program and answer annotations) on final performance. Our results show that with only 500 question/answer pairs and 50 program annotations, relatively small (around 1B parameters) prompt-free language models can be competitive with state-of-the-art models while being both faster and cheaper. We hope that such improvements can both accelerate progress on visual programming systems and make them more accessible.

In summary, our contributions are:

1. Propose decoupling visual program generation into templates and arguments for prompt-free visual programming.
2. Demonstrate that template and argument decomposition can be used for both synthetic data creation and a two-step program generation.
3. Investigate the relative effect on performance of number of program annotations and answer annotations for both prompt-free and prompt-based program generation methods. Our results indicate that prompt-based meth-

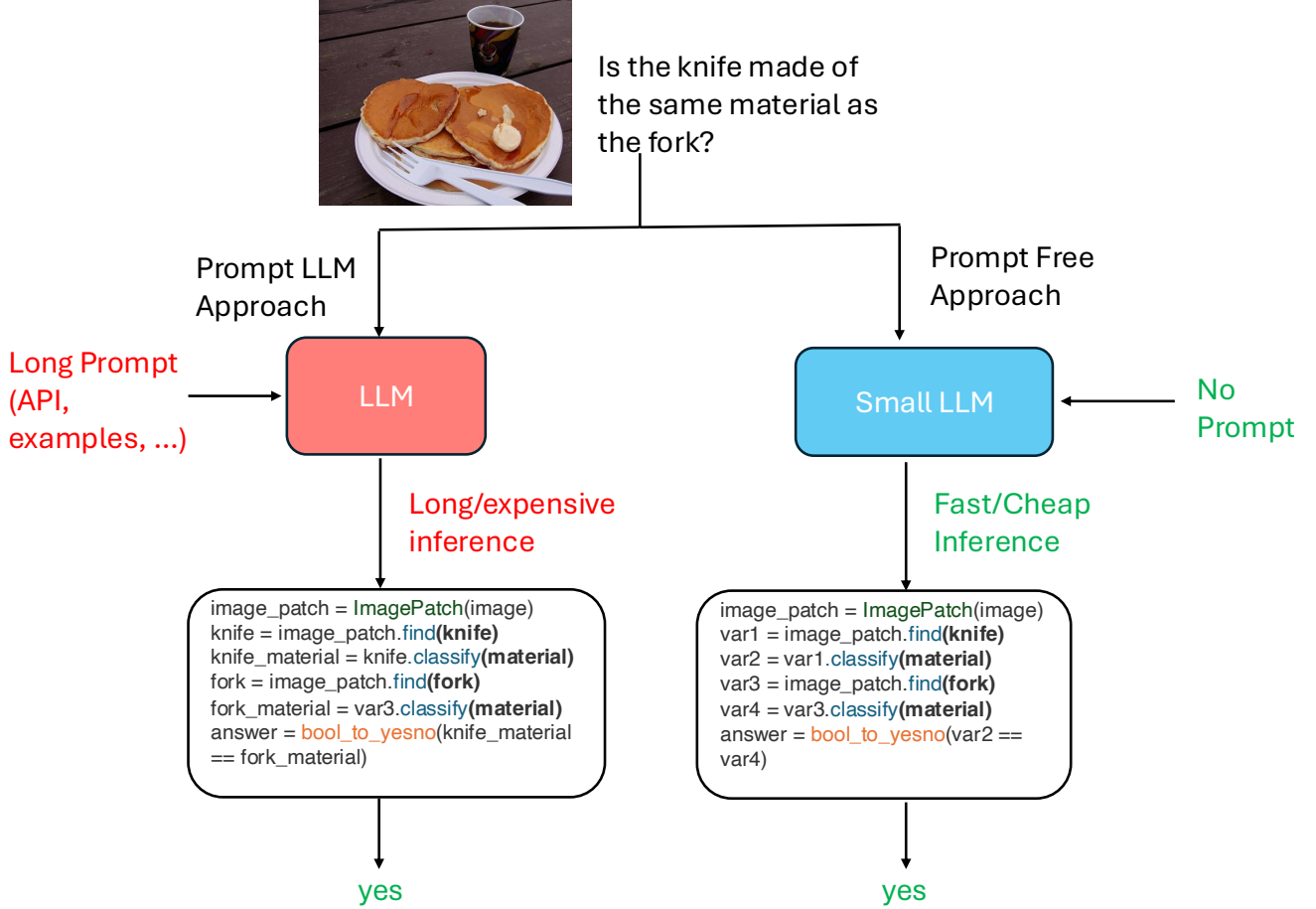


Figure 1. Prompting-based visual programming is expensive, slow and hard to improve. Decoupling visual programs into a higher level skill called templates and question-specific concepts leads to fast, cheap and prompt-free visual programming that has comparable performance to prompting-based methods.

ods benefit more from annotated programs while our prompt-free method improves more with additional answer annotations.

2. Related Works

Visual Programming A long line of work investigates generating and executing programs to perform visual tasks. Early approaches [2, 3, 13, 15] generate programs and execute the programs with learned end-to-end neural modules. Based on the impressive code generation capabilities of LLMs, visual programming frameworks [11, 32, 33] generate programs given an API and in-context examples and execute the programs with large pre-trained vision models. Visual programming has been applied to many different domains and applications including visual question answering, video question answering, text-to-image generation, and robotics [6, 19, 20, 22].

Many recent works focus on improving visual programming without modifying the underlying program generation

model or visual models. Such works can be categorized into program correction, refactoring and in-context example selection.

Program Correction A commonly used method for LLM program correction is self-refinement where execution feedback is passed (back) into the prompt of an LLM [21, 24, 30]. Stanic et al. [31] apply self-refinement to visual programming by passing in execution feedback from the entire program while Gao et al. [9] correct programs with intermediate feedback from different steps. Instead of using the same model to correct the program, Vdebugger [37] uses two separate models that are trained to identify and fix errors if needed [37]. To generate incorrect program data, VDebugger employs a mask-based data generation pipeline powered by a large language model (LLM). Similarly, our data augmentation method also utilizes a mask-based approach. While program correction can improve results, it comes with increased inference cost due to having to pass the program output into an LLM. Large prompts are often

needed for debugging. Our method focuses on efficiently generating the correct program instead of correcting an already generated program.

Refactoring In visual programming, APIs are predetermined which can be problematic for generalizing to new tasks since new tasks might not be covered in the initial API [5, 38]. CRAFT [38] prompts GPT-4 to solve questions from a training dataset, generalizes the solution through abstraction and validates the solution by prompting GPT-4 to provide the appropriate arguments to the abstracted program. Our template and argument decomposition is similar except we **learn to** both match questions to the abstractions (i.e. templates) and fill in the arguments with smaller language models [26].

In-context Example Selection Both the API and in-context examples are important for downstream performance. Designing and choosing which in-context examples to use can be time intensive. One alternative for in-context example generation, is to annotate examples from a training set and keep the annotations that return the correct answer when executed [31, 34]. Tao et al. [34] refers to this approach as auto-context generation. We use this approach to generate a pool of in-context examples for our prompt-based method. To select which in-context examples to use for a given question, we use retrieval-augmented generation (RAG) [18].

While program modification and prompt-based methods are quite popular, both suffer from long inference time and reliability. Due to the auto-regressive nature of LLMs, the longer the prompt, the longer it takes to generate a program. Prompts are also extremely sensitive to specific formats. Scalar et al. [29] show that minor changes in prompt formatting can cause performance differences up to 76%. Such sensitivity can make it difficult to reproduce results when closed models are updated or transferring results to different models. In our experiments, we show that with a small amount of training data, a prompt is not needed when using a medium-sized language model (fewer than 1 billion parameters).

Finetuning Learning Methods Compared to prompt-based methods, there are relatively few methods focused on improving program generation in visual programming or the larger field of tool-based LLMs. One of the main challenges is the lack of program annotations for input/output pairs. Language modeling has a long history designing self-supervised tasks, especially in pre-training [8, 26]. In Toolformer [28], the authors use a form of self-supervision, to create a training dataset to finetune an LLM on tool-use programs. For each question/answer pair in a pre-existing training dataset, an LLM is prompted to generate a corresponding program. If the program decreases the training loss (of the same LLM) then the program annotation is

Template	<pre> image_patch = ImagePatch(image) var1 = image_patch.find('arg.0') var2 = var2.classify('arg.1') var3 = image_patch.find('arg.2') var4 = var3.classify('arg.3') answer = bool.to.yesno(var2 == var4) </pre>
Are the cat and the tshirt the same color ?	<pre> arg.0 = cat arg.1 = color arg.2 = tshirt arg.3 = color </pre>
Is the sofa made from the same material as the chair ?	<pre> arg.0 = sofa arg.1 = material arg.2 = chair arg.3 = material </pre>
Is the vase the same shape as the table ?	<pre> arg.0 = vase arg.1 = shape arg.2 = table arg.3 = shape </pre>

Table 1. A template is a particular ordering of operations. The questions above all share the same template since they only differ in the arguments. We want to answer similar questions the same way and easily generate synthetic data.

added to a new dataset. Then the same LLM is finetuned on the generated dataset. A similar approach is used in both program correction methods such as VDebugger [37] and in finetuning based methods [16] such as VisRep. LoRA [12] is frequently used when finetuning LLMs on the generated data. In VisRep, which is most closely related to our work, Khan et al. [16] use a self-training approach similar to toolformer for visual programming. Self-annotated programs are kept if the executed program returns the correct answer. A key difference is that in our work, we use a much smaller number of question/answer pairs than in Toolformer or VisRep and use a data augmentation method similar to masked language modeling to generate a training dataset. In addition, an unspecified number of program annotations are manually corrected in VisRep.

3. Method

We present three different methods for program generation. In Section 3.1 we discuss prompt-based program generation. In Section 3.2, we introduce our template-based method and data augmentation procedure and in Section 3.3, we present a direct prediction approach that still benefits from the template-based data augmentation.

Preliminaries Following the notation in ViperGPT [33], our goal is to generate a program $z = \pi(q)$ with a program generator, π and input query q such that when executed with execution engine ϕ and corresponding visual input x , $\phi(x, z)$ returns the correct answer. The input query can contain a prompt p with an API and in-context examples depending on the program generation approach.

We refer to annotations containing (questions, answer, program) triplets as program annotations and annotations containing only (question, answer) pairs as answer annotations. Program annotations are clearly more costly to ob-

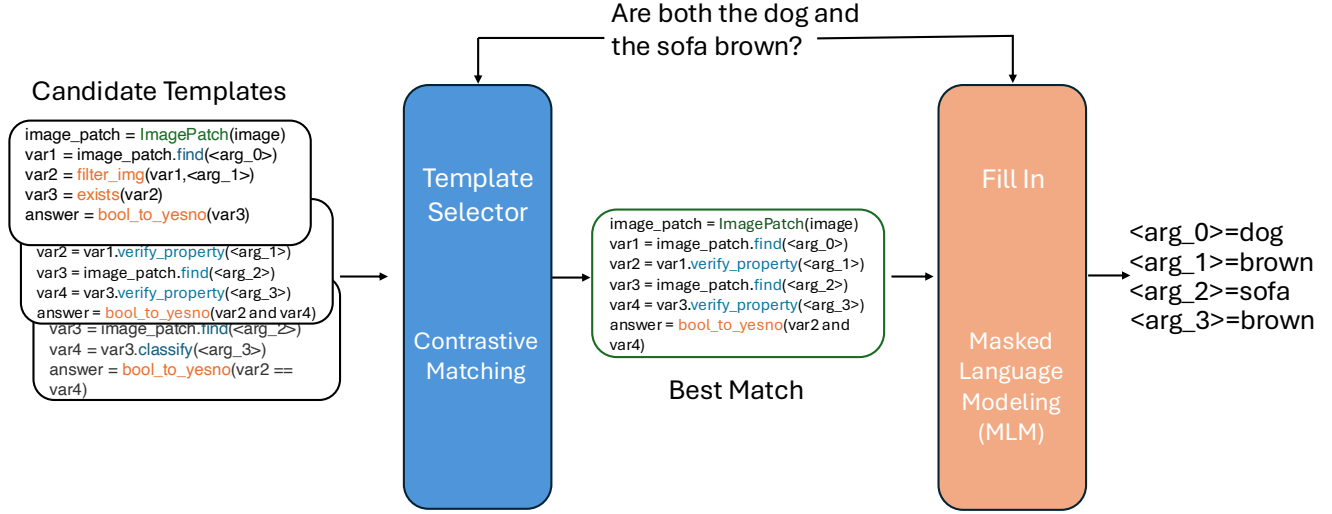


Figure 2. An overview of our main method with a matcher and infiller model. Given a question, the matcher pairs the question with the most likely template. The infiller, then fills in the template with the correct arguments.

tain. Most works [16, 28], assume that there are plenty of answer annotations. However, we assume that we only have 500 examples (for the GQA [14] dataset, 500 is .06 % of the 943,000 question/answer pairs in the train balanced split).

3.1. Prompt-Based Generation

Prompt-Based Generation uses a prompt with an API and a set of program annotations for every input. For our experiments, we use GPT-4o-mini as the LLM, a modified version of the Viper-GPT API (see next section for details), and 50 program annotations.

To improve prompt-based generation using the given 500 answer annotations, we perform a similar procedure to auto-context generation discussed in Section 2 which we refer to as auto-annotation. Programs are generated for each training example, and programs that produce the correct answer when executed are added to the set of program annotations. To limit costs and improve inference speed, we use Retrieval Augmented Generation (RAG) [18] and select the most relevant program annotations if more than 50 are available. We use finetuned MPNet-Base-v2 [27] as the RAG embedding model and for each candidate question, we compute the similarity between the candidate question embedding and the questions in the set of program annotations. The top-50 most similar questions and their corresponding programs are used in-context.

3.2. Template-Based Method

Almost all program generation methods auto-regressively generate a program based on some input. We introduce an alternative two-step approach: template matching and infilling. To understand the intuition behind our method, consider the set of questions in Table 1. All of these questions

compare properties of two objects. The general structure of each program is the same with the only difference coming from the inputs to the functions. If we answer one of these questions correctly and know that the remaining questions have the same structure, then all of the remaining questions should have that same structure or should be consistent. We will refer to the combination of template matching and infilling as the template-based method. An overview of our method can be seen in Figure 2.

The program generation objective can be rewritten as $z = \theta_{\text{infill}}(q, \theta_{\text{matcher}}(q, \mathbf{T}))$ where T is a list of candidate templates, and θ_{matcher} and θ_{infill} are matching and infilling models respectively.

Template Matching We define a template as a specific ordering of functions, where a function is an API call to a visual model or python operation. Templates are argument *independent*. For example, if the program is

```
image_patch = ImagePatch(image)
dog = image_patch.find('dog')
answer = dog.classify('color')
```

then the template would be

```
image_patch = ImagePatch(image)
var1 = image_patch.find(<arg>)
answer = var1.classify(<arg>)
```

Please see Table 2 for examples of questions and corresponding templates.

Initial Template Creation Given an initial set of program annotations, we extract templates and arguments from each program similar to the abstraction method of CRAFT [38]. Extracting is quite simple: replace specific variable names with generic ones and put in placeholders for each argument while keeping track of the ordering of the arguments (which will be filled in during infilling).

Matching The matcher, θ_{matcher} is an encoder that matches or pairs a given question x with the correct template. For a given input x , θ_{matcher} produces an embedding \mathbf{e}_x of size $m_x \times f$, where m_x is the number of tokens in the input sequence and f is the token dimension. The matcher likewise encodes each template $t \in \mathbf{T}$, where \mathbf{T} is all possible templates, creating an embedding \mathbf{e}_t of size $m_t \times f$ for each template t .

For each encoded template and input, we compute the average embedding across the tokens i.e

$$\mathbf{e}_{x_{\text{avg}}} = \frac{\sum_{i \in m_x} \mathbf{e}_{x_i}}{m_x} \quad (1)$$

and

$$\mathbf{e}_{t_{\text{avg}}} = \frac{\sum_{i \in m_t} \mathbf{e}_{t_i}}{m_t} \quad (2)$$

for each $t \in \mathbf{T}$, creating a single embedding of size f for the input and for each possible template. Our early experiments found that using the average embedding instead of the end-of-sequence token like in the CLIP text encoder, had higher performance. We use a contrastive loss similar to CLIP. For a batch of size n , we compute

$$L(\theta_m(x)) = -\frac{1}{N} \sum_{i=1}^N \left(\log \frac{\exp(\text{sim}(\mathbf{e}_{x_{i\text{avg}}}, \mathbf{e}_{t_{i\text{avg}}}))}{\sum_{j=1}^T \exp(\text{sim}(\mathbf{e}_{x_{i\text{avg}}}, \mathbf{e}_{j_{\text{avg}}}))} \right) \quad (3)$$

where $\theta_m = \theta_{\text{matcher}}$. We compute the cosine similarity between each possible template and a given input and take the most similar pair. Unlike CLIP, the loss is batch independent and the loss is not symmetric, i.e. we only require the text to match the template. While some questions could be answered correctly by multiple templates, we consider only one template to be correct.

Infilling After the matcher produces a question/template pair, the infilling θ_{infill} model generates the arguments for the template or fills in the template. We treat the problem of infilling as the span denoising or masked language modeling objective from T5 [26]. Each argument is assigned a unique mask token (referred to as sentinel tokens in T5), and the objective is to predict the values of each masked token. Unlike BERT [8], arguments that contain multiple words are represented with a single mask token. The input to the infiller is question and the template. The output is the sequence of the masked out tokens separated by the token

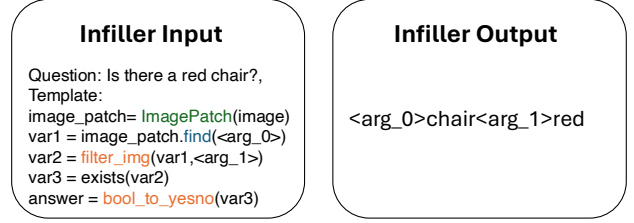


Figure 3. The input to the infiller is the question and template and the output are the values of the argument. Only a small number of tokens are generated.

ids. See Figure 3 for an example. We use the same span denoising loss as in T5:

$$L(\theta_{\text{infill}}) = -\frac{1}{S} \sum_{s=1}^S \log P(y_s | y_{<s}, x) \quad (4)$$

where S is the number of arguments or masked tokens in the input, y_s represents the input before argument s and x is the original input (question and template). Even though the infiller auto-regressively generates the output sequence, the output is quite small ($2 \times$ number of arguments), so inference speed is not affected.

Data Augmentation To improve performance with limited program and answer annotations, we perform data augmentation. We can use the decomposition of programs into templates and arguments to generate synthetic data similar to how masked language modeling is used in BERT [8]. As can be seen in Table 1, for many questions and programs, the arguments appear directly in the question. Consider the example of seen in Figure 4: “Are the dog and couch the same color?” The arguments are dog, couch, and color. Once we find similar words for each, we can simply replace them in the sentence. Since we already know the template, and the arguments, we also have the program. For GQA, the possible word replacements are from ViperGPT [33] and for arguments that are either not in the question or are equal to the entire question, we use GPT to generate similar questions. For VQAv2, we use BERT [8] when an argument is a single word and BART [17] to replace phrases. If possible, we replace arguments with arguments of the same type e.g. an attribute is replaced by an attribute.

During training, each argument in each question, has a 50% chance of being replaced. If an argument is to be replaced, we then uniformly sample among the possible replacements. For some arguments, like objects, the number of possible replacements is quite large (greater than 1500), while for arguments like directions such as left, right, etc. the number of replacements is small (fewer than 10).

Auto-Annotation Depending on a user’s needs and budget, they could be interested in different forms of annotation. Auto-annotation for the template-based method is similar to prompt-based annotation except that we use two types of annotators: self (model) annotation and LLM (GPT) annotation. For each question/answer pair, we first attempt to annotate it with the template-based model. If that is incorrect, we then attempt to annotate the pair with GPT-4o-mini. Similar to prompt-based generation, we consider a program annotation to be correct if the correct answer is returned by the executed program. Such an approach saves money and time due to the speed of our program generation.

Training Training the template-based method has several steps. First, we train the template-based method on augmented program examples. We train the matcher and infiller separately and use LoRA [12] with a batch size of 16. The second step is to perform the auto-annotation discussed above where we use both GPT-4o-mini and the template-based method to generate programs, where possible, that give answers matching the 500 answer annotations. The last step is to repeat step 1, where we train the template-based models again on augmented data (including the newly annotated data from the previous step).

3.3. Direct Generation Method

Prompt-free direct generation uses a smaller and more specialized language model [26] to generate a program. The only input to the model is the question. Since no prompt is needed, the input is much smaller. Prompt-free direct generation is trained on the same augmented dataset as the template-based method.

Questions	Template
Is the blue car the same shape as the chair ?	<pre> image.patch = ImagePatch(image) var1 = image.patch.find(<arg.0>) var2 = filter_img(<arg.1>) var3 = var2.classify(<arg.2>) var4 = image.patch.find(<arg.3>) var5 = var4.classify(<arg.4>) answer = bool.to.yesno(var3 == var5) </pre>
Is leather jacket made of the same material as the shirt ?	<pre> image.patch = ImagePatch(image) var1 = image.patch.find(<arg.0>) var2 = var1.crop.position(<arg.1>) var3 = var2.find(<arg.2>) answer = var3.classify(<arg.3>) </pre>
What type of food is near the person ?	<pre> image.patch = ImagePatch(image) var1 = image.patch.find(<arg.0>) var2 = var1.crop.position(<arg.1>) var3 = var2.find(<arg.2>) answer = var3.classify(<arg.3>) </pre>
What is the vehicle next to the animal ?	<pre> image.patch = ImagePatch(image) var1 = image.patch.find(<arg.0>) var2 = image.patch.find(<arg.1>) answer = choose_relationship(var1, var2, <arg.2>) </pre>
Is the car to the left or right of the tree ?	
Is pot above or below the pan ?	

Table 2. Some examples of questions and corresponding templates. Multi-colored words correspond to multiple arguments.

4. Experiments

Experimental Setup In all of our experiments, both the matcher and infiller are T5-Large models [26]. The matcher is the CodeT5+ 770m encoder only [36] while the infiller is (a separate) encoder-decoder. For direct program generation (referred to as ‘Direct’), we use the full CodeT5+ 770m

Method	GQA (test-dev)	VQAv2 (10K)
Template-Based Initial	37.0	50.4
Template-Based After Auto-Annotation	42.0 (+5.0)	53.6(+3.2)
Direct Initial	35.9	50.1
Direct After Auto-Annotation	42.0(+6.1)	52.9(+2.8)
GPT Initial	41.3	55.0
GPT After Auto-Annotation	42.3 (+1.0)	57.0 (+2.0)

Table 3. Results from the different methods on GQA and VQAv2. The template-based method achieves better performance than GPT initial on GQA and is close to GPT initial performance on VQAv2.

model trained for Python. The direct program generation is trained on the same dataset as the template-based method. We use LoRA [12] to update the models. GPT-4o-mini is used for prompt-based generation.

We use a slightly modified API from the original ViperGPT paper [33]. The main differences are some additional functions (to reduce program length) and removal of the use of a VQA model if earlier parts of a program fail. For visual models, we use InstructBLIP(Flan-T5 XL) [7] for general visual queries, Owl-ViT2 [23] for detection and CLIP [25] for classification.

It is difficult to compare visual programming experiments to previous works due to changes in APIs and different underlying visual models. In the following experiments, we consider GPT Initial to be representative of previous works [11, 33].

In Section 4.1 we evaluate the different methods on VQA datasets. In Section 4.2, we examine the effect of increasing and decrease the number of program annotations and answer annotations, and we perform some analysis of our results in Section 4.3

4.1. Method Comparison

Program Generation Comparison We evaluate the three methods before and after auto-annotation on two commonly used visual question answering datasets: GQA [14] and VQAv2 [10]. 500 question/answer pairs (answer annotations) are randomly sampled from respective training sets for auto-annotation. For GQA, we evaluate on the entire test-dev split while for VQAv2, we sample 10K questions from the validation set. GQA is evaluated using exact match and VQAv2 is evaluated based on annotator/answer agreement as in the original benchmark.

Table 3 shows the results of the three different approaches. GPT with Auto-Annotation has the highest performance but the other two non-prompt-based methods, which are much smaller and finetuned with limited data have close performance. Direct prediction shows the highest improvement on the GQA dataset while GPT Auto-Context Generation has the smallest increase.

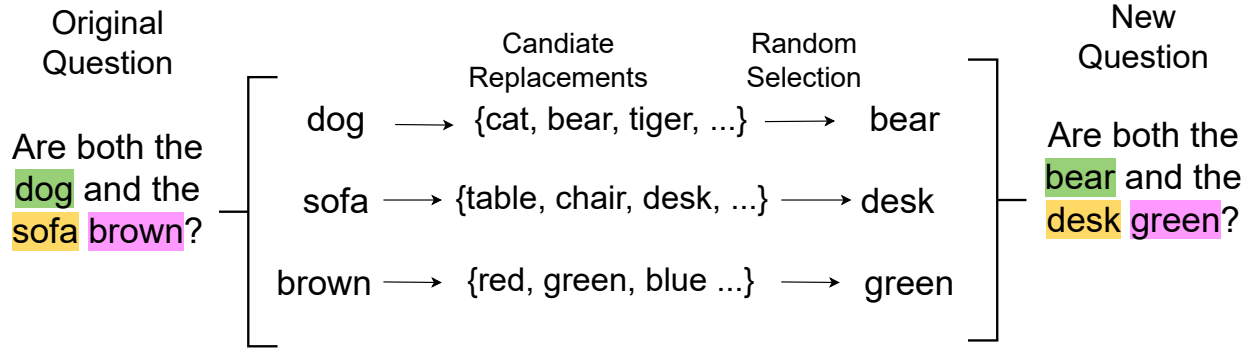


Figure 4. An example of our data augmentation approach. Both the new and old question have the same template so the template matcher output should predict the same template for both. The arguments for the new and old programs are different. But, in the arguments, (dog, sofa, brown) should be replaced with (bear, desk, green).

	GQA		VQAv2	
	Template-Based	GPT	Template-Based	GPT
Model Annotated Correctly	208	N/A	265	N/A
GPT Annotated Correctly	64	231	69	287
Total New Data Points	272	231	384	287
Cost (Training and Inference)	\$0.19	\$6.94	\$0.13	\$5.46
Inference Speed (q/s)	71.43	1.33	71.43	1.33

Table 4. Comparison of Model Annotated and GPT Annotated Data. The template-based method is much faster and cheaper.

Auto-Annotation Table 4 shows the number of annotations made from both datasets, as well as cost and inference speed. For both datasets, the template-based models successfully annotate more than half of the possible examples. Using both the template-based model and GPT results in more program annotations. When using two annotation models, GPT is given 292 partial annotations but only successfully annotates 64 of them indicating that GPT also had trouble generating programs in cases where the template-based method failed..

We can also see from Table 4, that the template-based method is both much cheaper and faster than GPT. Most of the GPT cost comes from inference (since there are many more examples). We evaluate template-based program generation inference time using a batch size of 256 on a single A100 GPU. Note that GPT inference time can vary greatly. The reported time of 1.33 questions/s is a lower bound. Our method is 53.7 times faster than GPT.

Full Dataset Evaluation Previous visual programming works either use relatively small datasets for evaluation or subsample larger datasets due to the high cost and long time it takes for evaluation. Our template-based approach can generate thousands of programs a minute, making full dataset evaluation possible. We evaluate our template-based method on the full VQAv2 validation set which consists

of 214,354 questions. The final VQA score is 53.0. Performing such an evaluation with GPT would cost \$117 and would take 44.8 hours to finish generating programs only. The main bottleneck for large-scale evaluation is program execution. It took on average 3.37 s to execute a program (and varies significantly).

4.2. Program Annotation vs Answer Annotation

To determine how helpful program annotations are compared to answer annotations, we train both the template-based models and auto-context GPT with different values for both types of annotations. The number of program annotations refers to the total number of examples given initially, not how many examples are allowed in the context window (which is still 50 as before). The results can be seen in Figure 5. The highest value for both approaches is when 100 program annotations and 500 answer annotations are used. Both values are extremely close together. GPT appears to benefit more from increasing the number of program annotations. The performance is higher for 125 answer annotations and 500 program annotations, than it is for 500 answer annotations and 25 program annotations. However, it is the opposite for the template-based method where 125 answer annotations and 100 program annotations is **lower** than using 500 answer annotations and 25 program annotations.

4.3. Analysis

Effect of Data Augmentation To measure the effect of data augmentation, we train the template-based method and direct training only on the programs annotated during auto-annotation and initial set of program annotations (322 total) only. We can see from Table 5, that data augmentation increases the performance of both methods.

Correctness of Program During Annotation Unlike GQA, VQAv2 does not have a single correct answer. Each

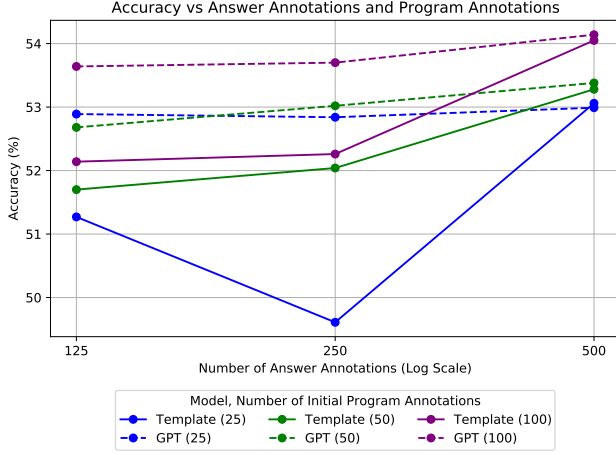


Figure 5. Performance changes on GQA after auto-annotation for the template-based methods and GPT. Program generation using the combination of answer and program annotations is more successful, leading to increased benefit for the template-based method.

Method	Template-Based	Direct
Without Data Augmentation	40.0	40.5
With Data Augmentation	42.0 (+2)	42.0 (+1.5)

Table 5. Data augmentation ablation on GQA test-dev after auto-annotation. The template-based method is more affected by use of data augmentation.

question in VQAv2 has 10 votes for the correct answer, which could lead to more than one answer considered correct. During auto-annotation, an annotated question/answer pair is added to the training set if it is correct. We experiment with ‘how’ correct a program should be by varying the threshold of the score for which we include an answer. In Table 6, we can see that using a lower threshold or accepting programs that are only partially correct improves performance for both methods. There was a larger effect on the template-based performance consistent with results from Section 4.2, that show answer annotations have more of an effect for the template-based method.

Template Characteristics To try understand the characteristics of programs generated, we evaluate the accuracy of programs based on length and template frequency. GPT on average generates longer programs with an average of 6.6 lines while the template-based method had programs with 6.0 lines. To evaluate program length accuracy, each predicted program on the GQA test-dev set is categorized based on the program length. Then the average accuracy is computed in each group for categories with more than 10 entries. The results, shown in Figure 6 show that while both methods decrease as programs become longer, the template-

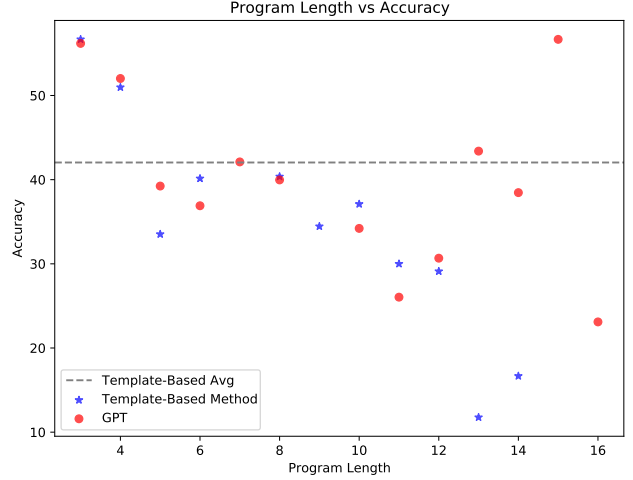


Figure 6. Accuracy on GQA (test-dev) based on program length. The template-based method has worse performance on longer programs.

based method does relatively worse. We perform a similar procedure for template frequency. Questions are grouped based on how frequently the predicted template appears in the training set (after auto-annotation). Figure 7 indicates that template frequency does not have a major affect on performance.

Method	50% Threshold	>0 Threshold
Template-Based	52.5	53.6
GPT	56.8	56.9

Table 6. VQAv2 results when using different scoring thresholds during annotation. Having a lower threshold improves performance.

5. Conclusion

High cost, unreliability and slow inference time limit the use of current LLM prompting-based visual programming methods. Our experiments and results show that prompting is not necessary for visual programming. Decoupling programs into higher level skills and arguments, leads to both fast and similarly accurate program generation and an effective data augmentation procedure. Template-based generation reaches similar performance as state-of-the-art LLMs at a fraction of the time and cost. Our method’s fast generation speed also makes large scale testing more feasible on datasets with hundreds of thousands of examples.

Template-based generation suffers from some of the same limitations as other visual programming systems including noisy evaluation. Predicting the correct answer on a VQA dataset does not mean that the program was correct, and predicting a wrong answer does not necessarily mean

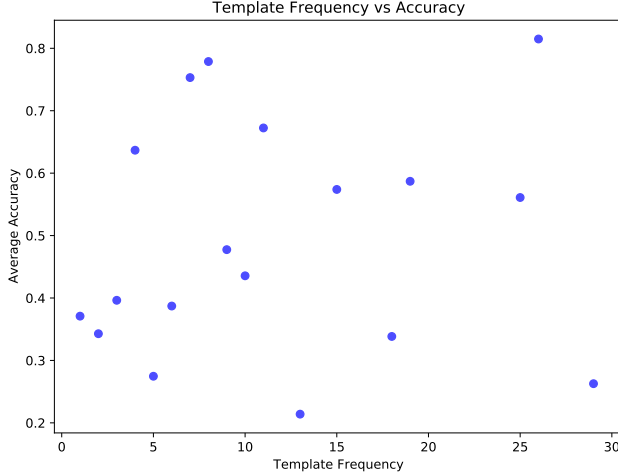


Figure 7. Accuracy on GQA (test-dev) based on predicted template frequency. Template frequency does not seem to affect performance.

a program is incorrect. Specific template-based generation limitations include only predicting entire templates at a time making it difficult to include intermediate feedback. Another limitation is program execution. While our approach generates programs quickly, executing programs still takes a long time.

In addition to addressing some of the limitations, other areas of future work based on our experiments include deeper exploration into the value of program annotations compared to answer annotations as well as how correct a program annotation should be. Our approach can also be combined with several methods described in Section 2. For example, during the auto-annotation process, an LLM could be given an opportunity to self-correct or refactor into different prompts.

In conclusion, we propose and demonstrate that template-based visual programming is cheap, effective and fast. Our experiments and results show that visual programming does not require prompting large LLMs to be successful. We anticipate that use of template-based visual programming will enable users and researchers to iterate more quickly on various visual programming systems and broaden their applications.

Acknowledgement This research is supported in part by ONR N00014-23-1-2383 and ONR N00014-21-1-2705. This work used the Delta system at the National Center for Supercomputing Applications through allocation CIS230398 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [4], which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2015. 2
- [3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554, San Diego, California, 2016. Association for Computational Linguistics. 2
- [4] Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. Access: Advancing innovation: Nsf’s advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, page 173–176, New York, NY, USA, 2023. Association for Computing Machinery. 9
- [5] Zhenfang Chen, Rui Sun, Wenjun Liu, Yining Hong, and Chuang Gan. Genome: generative neuro-symbolic visual reasoning by growing and reusing modules. *arXiv preprint arXiv:2311.04901*, 2023. 3
- [6] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, 2023. 6, 1
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 3, 5
- [9] Minghe Gao, Juncheng Li, Hao Fei, Liang Pang, Wei Ji, Guoming Wang, Zheqi Lv, Wenqiao Zhang, Siliang Tang, and Yueting Zhuang. De-fine: Decomposing and refining visual programs with auto-feedback. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 7649–7657, New York, NY, USA, 2024. Association for Computing Machinery. 2
- [10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

- [11] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 1, 2, 6
- [12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 6
- [13] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2
- [14] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 4, 6, 1
- [15] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE international conference on computer vision*, pages 2989–2998, 2017. 2
- [16] Zaid Khan, Vijay Kumar BG, Samuel Schuster, Yun Fu, and Manmohan Chandraker. Self-training large language models for improved visual program synthesis with visual reinforcement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14344–14353, 2024. 3, 4
- [17] M Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 5
- [18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 1, 3, 4
- [19] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500, 2023. 2
- [20] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [21] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [22] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13235–13245, 2024. 2
- [23] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 1
- [24] Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*, 2023. 2
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6, 1
- [26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3, 5, 6
- [27] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 4
- [28] Timo Schick, Jane Dwivedi-Yu, Roberto Dess, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4
- [29] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3
- [30] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [31] Aleksandar Stanić, Sergi Caelles, and Michael Tschannen. Towards truly zero-shot compositional visual reasoning with llms as programmers. *arXiv preprint arXiv:2401.01974*, 2024. 1, 2, 3
- [32] Sanjay Subramanian, Medhini Narasimhan, Kushal Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell, and Dan Klein. Modular visual question answering via code generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 747–761, Toronto, Canada, 2023. Association for Computational Linguistics. 1, 2
- [33] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023. 1, 2, 3, 5, 6
- [34] Heyi Tao, Sethuraman T V, Michal Shlapentokh-Rothman, Tanmay Gupta, Heng Ji, and Derek Hoiem. WebWISE: Un-

- locking web interface control for LLMs via sequential exploration. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3693–3711, Mexico City, Mexico, 2024. Association for Computational Linguistics. [3](#)
- [35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. [1](#)
- [36] Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi D.Q. Bui, Junnan Li, and Steven C. H. Hoi. Codet5+: Open code large language models for code understanding and generation. *arXiv preprint*, 2023. [6](#)
- [37] Xueqing Wu, Zongyu Lin, Songyan Zhao, Te-Lin Wu, Pan Lu, Nanyun Peng, and Kai-Wei Chang. Vdebugger: Harnessing execution feedback for debugging visual programs, 2024. [2](#), [3](#)
- [38] Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi Fung, Hao Peng, and Heng Ji. CRAFT: Customizing LLMs by creating and retrieving from specialized toolsets. In *The Twelfth International Conference on Learning Representations*, 2024. [3](#), [5](#)

Can We Generate Visual Programs Without Prompting LLMs?

Supplementary Material

We perform an error analysis in Section 6 and report details on models used and training hyper-parameters in Section 7. In Section 8, we describe the changes to the Viper-API [33] and list the full API/prompt in Section 9.

6. Error Analysis

We randomly sample 200 image/question pairs from the GQA [14] validation set and manually verify the post auto-annotation (50 program annotations, 500 answer annotations) predictions for both the template-based method and GPT. As shown in Figure 8, the different types of program generation were consistent. For 91 % of the questions, both methods were either correct or incorrect. For the remaining questions, a slightly higher number of GPT predictions are answered correctly. Some qualitative examples of images, questions and predicted programs where both methods return correct or incorrect answers can be seen in Figure 11. Examples where one method returns the correct answer and the other one does not can be seen in Figure 12.

For incorrect predictions, we examine the sources of error for each program generation method as seen in Figure 9 and Figure 10. Errors are categorized into non-overlapping categories: incorrect ground truth, ambiguous, visual models and program. Ambiguous image/question pairs either had multiple correct answers depending on the interpretation of the image and question and the predicted. Visual model errors had correct programs but produced the wrong answer and program based errors had incorrect programs. For both template-based and GPT, the highest two categories are due to ambiguity and visual model errors. Slightly more errors are due to incorrect programs for template-based generation.

7. Training and Model Details

We used the following models for executing programs:

1. CLIP ViT-L/14 [25]
2. InstructBLIP Flan-T5 XL [7]
3. OWLv2 Base Patch 16 Ensemble [23]

Program generation settings for GPT can be found in Table 7. Template-based and direct training hyper-parameters can be found in Table 8.

8. Changes to ViperGPT API

The following are major modifications made to the ViperGPT API [33].

1. Program annotations were modified not to use a vision-language model (VLM) when the program fails (see Fig-

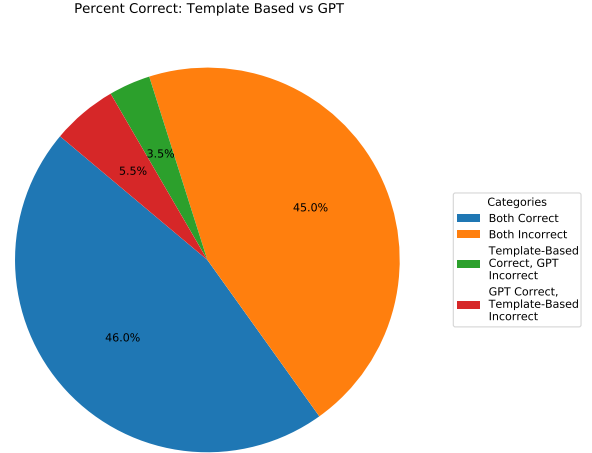


Figure 8. For the 200 validation questions sampled, 91% of the questions are either correctly answered by both types of models or incorrectly answered by both types of models.

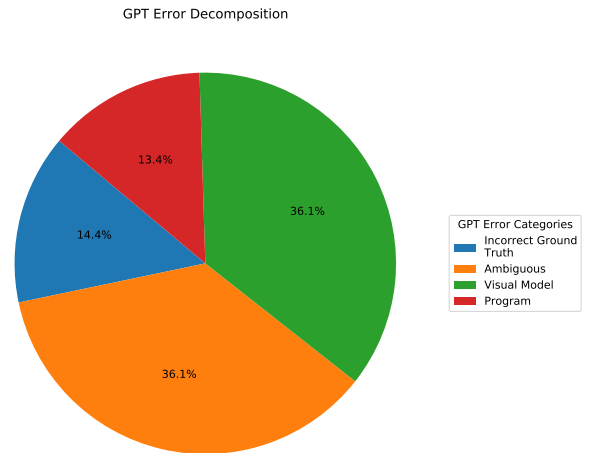


Figure 9. Different causes for error for GPT based models. Like the template-based model, the largest sources of error are due to ambiguous questions or mistakes in the visual models.

ure 13 for an example). In the original ViperGPT API, examples in the API included a line to directly query a VLM if other parts of the program failed such as when no object is found. The performance using the original ViperGPT code decreases considerably when the VLM backup lines are removed from the API as shown in Table 9.

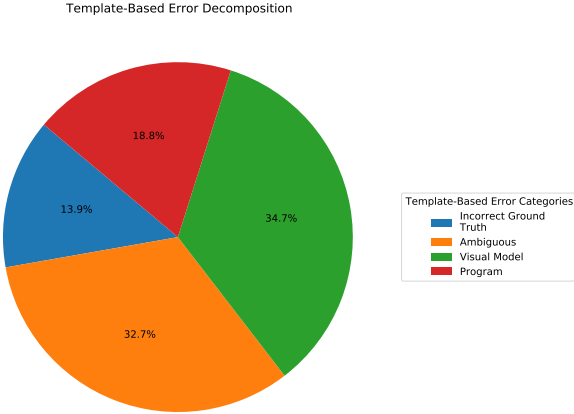


Figure 10. Different causes for error for template-based models. The percent of errors due to program generation is a bit higher compared to GPT.

Setting	Value
Temperature	0
Top_p	1.0
Frequency Penalty	0.0
Presence Penalty	0.0
Max Output Tokens	256

Table 7. GPT-4o-mini generation settings

Hyper-parameters	Value
LoRA target modules	All linear layers
LoRA rank	8
LoRA alpha	16
LoRA bias	None
LoRA dropout	0.05
LR	1e-4
Batch Size	16
Max Output Tokens	256

Table 8. Training and evaluation settings for template-based and direct training methods

2. An object is always returned by the object detector.
3. Program annotations did not include parts of the program that were shared among all examples.
4. Several new functions were added.
 - (a) Verify Relationship: Given two objects and a relation, return a boolean whether the objects satisfy that relationship.
 - (b) Choose Relationship: Given two objects, return the relationship between the two.
 - (c) Crop Position: Crop part of the image based on a

Use of VLM Backup	GQA-Test Dev
ViperGPT with VLM Backup	47.3
ViperGPT without VLM Backup	26.0

Table 9. Change in GQA test-dev accuracy using original ViperGPT API when not using a VLM when the program fails

position relative to an object



	Image	Q/A	GPT	Template Based
Both Incorrect		Q: Is the ambulance on the right? A: No	<pre>image_patch=ImagePatch(image) ambulance=image_patch.find('ambulance') answer=verify_relationship(ambulance,image_patch,'right')</pre> Prediction: yes	<pre>image_patch=ImagePatch(image) var1=image_patch.find('ambulance') answer=verify_relationship(var1,image_patch,'right')</pre> Prediction: yes
Both Correct		Q: What is the staircase made of? A: Wood	<pre>image_patch=ImagePatch(image) staircase=image_patch.find('staircase') answer=staircase[0].classify('material')</pre> Prediction: wood	<pre>image_patch=ImagePatch(image) var1=image_patch.find('staircase') answer=var1[0].classify('material')</pre> Prediction: wood

Figure 11. Image/question pairs and GPT and template-based predicted programs where answers are consistently incorrect or correct between GPT and template-based methods. Most programs generated by the two methods are similar.



	Image	Q/A	GPT	Template Based
Model Correct, GPT Incorrect		Q: What is the person that is to the right of the kid looking at? A: lake	<pre>image_patch=ImagePatch(image) kid=image_patch.find('kid') to_right_of_kid=kid[0].crop_position('right') person_looking_at=to_right_of_kid.find('person') answer=person_looking_at[0].simple_query('What is this person looking at?')</pre> Prediction: camera	<pre>image_patch=ImagePatch(image) var1=image_patch.find('kid') var2=var1[0].crop_position('right') answer=var2.simple_query('What is the person looking at?')</pre> Prediction: lake
Model Incorrect, GPT Correct		Q: Are there pedestrians on the sidewalk? A: yes	<pre>image_patch=ImagePatch(image) sidewalk=image_patch.find('sidewalk') pedestrians=sidewalk[0].find('pedestrian') pedestrians_exist=exists(pedestrians) answer=bool_to_yn(pedestrians_exist)</pre> Prediction: yes	<pre>image_patch=ImagePatch(image) var1=image_patch.find('side walk') var2=filter_img(var1,'pedestrian') var3=exists(var2) answer=bool_to_yn(var3)</pre> Prediction: no

Figure 12. Image/question pairs and GPT and template-based predicted programs where answers are different between the two methods. The GPT program that returns the wrong answer is still correct while the template-based program that returns the wrong answer is incorrect.

With VLM Backup	Without VLM Backup
<pre>image_patch= ImagePatch(image) ground_patches= image_patch.find('ground') if len(ground_patches)==0: return simple_query('Is the ground blue or brown?') return ground_patches[0].classify(['blue', 'brown'])</pre>	<pre>image_patch= ImagePatch(image) ground_patches= image_patch.find('ground') return ground_patches[0].classify(['blue', 'brown'])</pre>

Figure 13. Difference in program annotations when a VLM is used as a backup model for the question 'Is the ground blue or brown?' The highlighted portion is removed from all program annotations used.

9. Prompt

Instructions

For each question provided, generate a Python program that includes a return statement. Assume that `image_patch = ImagePatch(image)` is already defined. The final output of the program should always be a string.

ImagePatch

Attributes

1. **cropped_image**
Type: array
Description: An array representing the cropped image.
2. **left**
Type: int
Description: The left border of the crop's bounding box.
3. **lower**
Type: int
Description: The bottom border of the crop's bounding box.
4. **right**
Type: int
Description: The right border of the crop's bounding box.
5. **upper**
Type: int
Description: The top border of the crop's bounding box.
6. **name**
Type: str
Description: The name of the cropped image.
7. **confidence**
Type: float
Description: The confidence score of the prediction.

Methods

1. **find(object_name: str) -> List[ImagePatch]**
Description: Returns a list of image patches containing the specified object.
Notes: find should not be the last operation in a program.
Examples:
`image_patch.find('chair')`
`image_patch.find('table')`
2. **crop(left: int, lower: int, right: int, upper: int) -> ImagePatch**
Description: Returns a new image patch at the given coordinates.
Example:
`image_patch.crop(10, 20, 30, 40)`
3. **crop_position(direction: str, reference_patch: ImagePatch) -> ImagePatch**
Description: Returns a new image patch in the specified direction relative to the reference_patch. Directions can include 'left', 'right', 'above', 'below', 'on', 'in front', etc.
Notes: The result of crop_position should not be immediately indexed on the next line.
Examples:

```

    image_patch.crop_position('left', image_patch)
    image_patch.crop_position('above', image_patch)
4. **verify_property(property_name: str) -> bool**
    Description: Returns True if the object contains the specified property;
    otherwise, False.
    Notes: Can only be called on an image patch.
    Examples:
    image_patch.verify_property('red')
    image_patch.verify_property('running')
5. **classify(options: Union[str, List[str]]) -> str**
    Description: Given a category (e.g., 'color', 'material', 'furniture') or a
    list of options, returns the best option for the image patch.
    Notes: The input should not be 'object'.
    Examples:
    image_patch.classify(['red', 'blue'])
    image_patch.classify('color')
6. **simple_query(question: str) -> str**
    Description: Answers questions about the image, especially ambiguous ones
    (e.g., 'Who is riding?').
    Examples:
    image_patch.simple_query('Who is riding?')
7. **count(list: List[ImagePatch]) -> str(int)**
    Description: Counts the number of image patches in the list.
    Examples:
    image_patch.count(image_patch.find('trees'))

```

General Functions

```

-----
1. **filter_img(image_patches: List[ImagePatch], criteria: str) ->
List[ImagePatch]**
    Description: Filters the list of image patches based on the given criteria.
    Examples:
    filter_img(image_patches, 'red')
2. **choose_relationship(patch1: Union[ImagePatch, List[ImagePatch]], patch2:
Union[ImagePatch, List[ImagePatch]], relationships: Union[List[str], str]) ->
str**
    Description: Chooses the relationship that best matches the two patches from
    the provided options.
    Examples:
    choose_relationship(image_patch1, image_patch2, ['on top of', 'next to'])
3. **verify_relationship(patch1: Union[ImagePatch, List[ImagePatch]], patch2:
Union[ImagePatch, List[ImagePatch]], relationship: str) -> str**
    Description: Returns 'yes' or 'no' based on whether the specified relationship
    holds between the two patches.
    Examples:
    verify_relationship(image_patch1, image_patch2, 'on top of')
4. **exists(patches: Union[ImagePatch, List[ImagePatch]]) -> bool**
    Description: Checks whether any of the provided image patches exist.
    Notes: If used as the last operation, it should be followed by
    bool_to_yn().
    Examples:
    exists(image_patches)
5. **bool_to_yn(value: bool) -> str**

```

Description: Converts a boolean value to 'yes' or 'no'. Used to convert outputs of `verify_property` and `exists`.

Examples:

```
bool_to_ynno(exists(image_patches))
```

Additional Notes

Python Integration:

You may utilize standard Python functions within your programs.

Make sure to define `answer` as the last operation. Do not include comments. Only return the program.

Here are examples how to use the tools: