

# 左右の脳を一緒に使う: ビジョンと言語計画に向けて (ICML2024)

Hiroshi Yamakawa, Matsuo Lab

# 書誌情報

- Using Left and Right Brains Together: Towards Vision and Language Planning
  - 著者は9名（主要著者：Jun Cen、Chenfei Wu、Xiao Liuが）所属は以下が含まれる
    - Southern University of Science and Technology（南方科技大学）
    - Microsoft Research Asia（マイクロソフトリサーチアジア）
    - The Hong Kong University of Science and Technology（香港科技大学）
    - Xi'an Jiaotong University（西安交通大学）
    - City University of Hong Kong（香港城市大学）
    - Peng Cheng Laboratory, Shenzhen（鵬城実験室、深セン）
  - URL
    - (arXiv) <https://arxiv.org/html/2402.10534v1>
    - (ICML2024 page) <https://icml.cc/virtual/2024/poster/33100>
    - No code implementations yet
  - 人間の脳の左右半球の働きにヒントを得て、言語処理と視覚処理を統合した新しいAI推論フレームワーク（Vision-Language Planning）を提案する研究

# ヒトの脳における左右の機能棲み分けの一般論

## How Human Thinks



著者等のビデオより：<https://www.youtube.com/watch?v=UgGuivPolaY>

Each hemisphere of the brain governs different types of thinking.

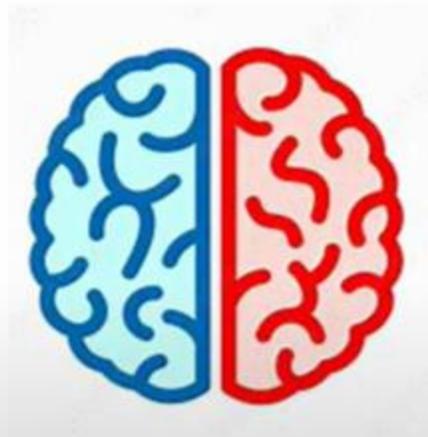
**Left hemisphere**

**Logical  
Analytical Thinking**



*Language*

**Large Language Model  
(LLM, such as GPT4)**



**Right hemisphere**

**Imagination  
Creative Thinking**



*Vision*

**Large Vision Model  
(LVM, such as Sora/SVD)**

# 研究背景と目的

## 1. 現在のAIモデルの限界

### 大規模マルチモーダルモデル（LMM）の問題点

- a. 主に言語空間での計画に依存
- b. 視覚的な推論能力が不足

## 2. 人間の脳の働きからのヒント

左半球：言語処理と論理的思考・右半球：視覚空間処理と直感的思考

## 3. 研究目的

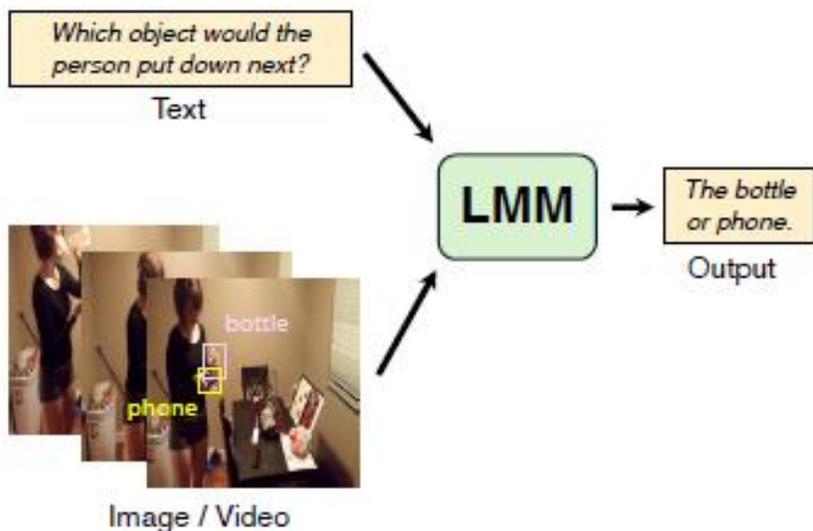
### Vision-Language Planning (VLP) フレームワークの提案

- a. 言語計画と視覚計画の統合
- b. より人間に近い認知プロセスの実現

## 4. VLPの期待される効果

- a. マルチモーダルタスクでの性能向上
- b. より柔軟で文脈に応じた推論の実現

# 従来手法vs提案手法



(a) Existing works

- 従来手法 (LMM)

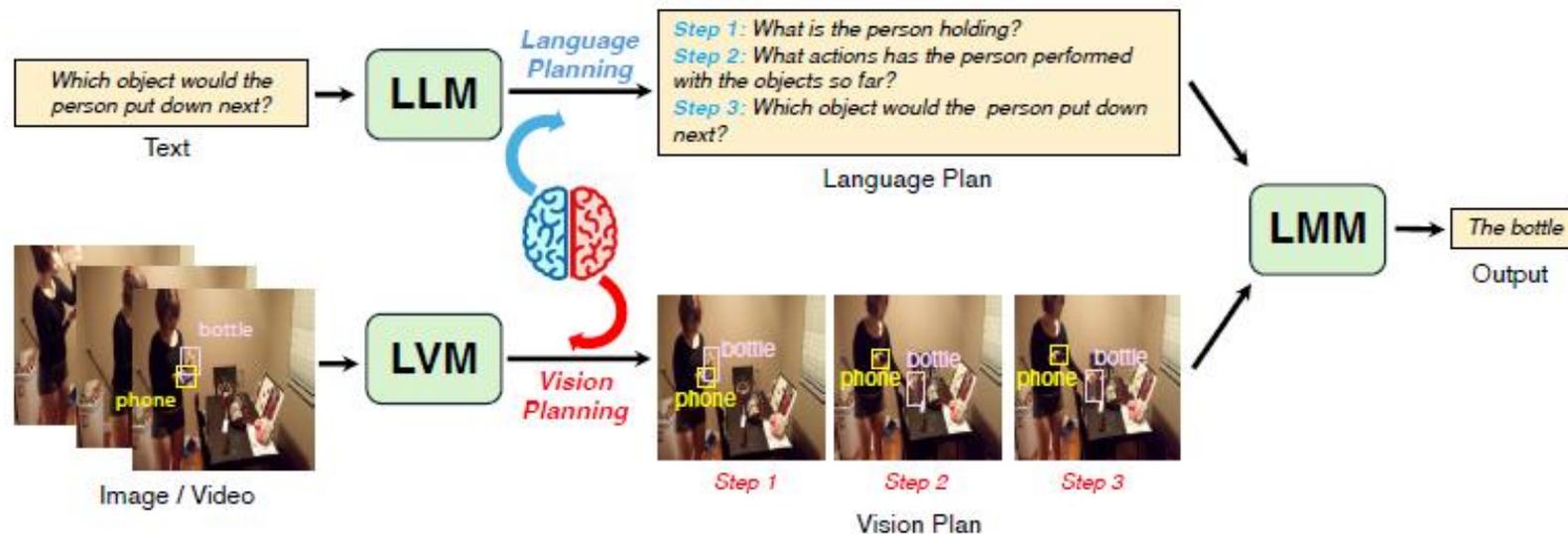
- 言語処理に重点
- 視覚情報を言語空間に変換
- 限定的な視覚的推論能力

- 提案手法 (VLP)

- 言語計画と視覚計画の統合
- 未来の視覚情報を生成・活用
- より柔軟で文脈に応じた推論

- VLPの利点

- 複雑なマルチモーダルタスクへの対応力向上
- 人間の認知プロセスに近いアプローチ
- 視覚的想像力を活用した推論



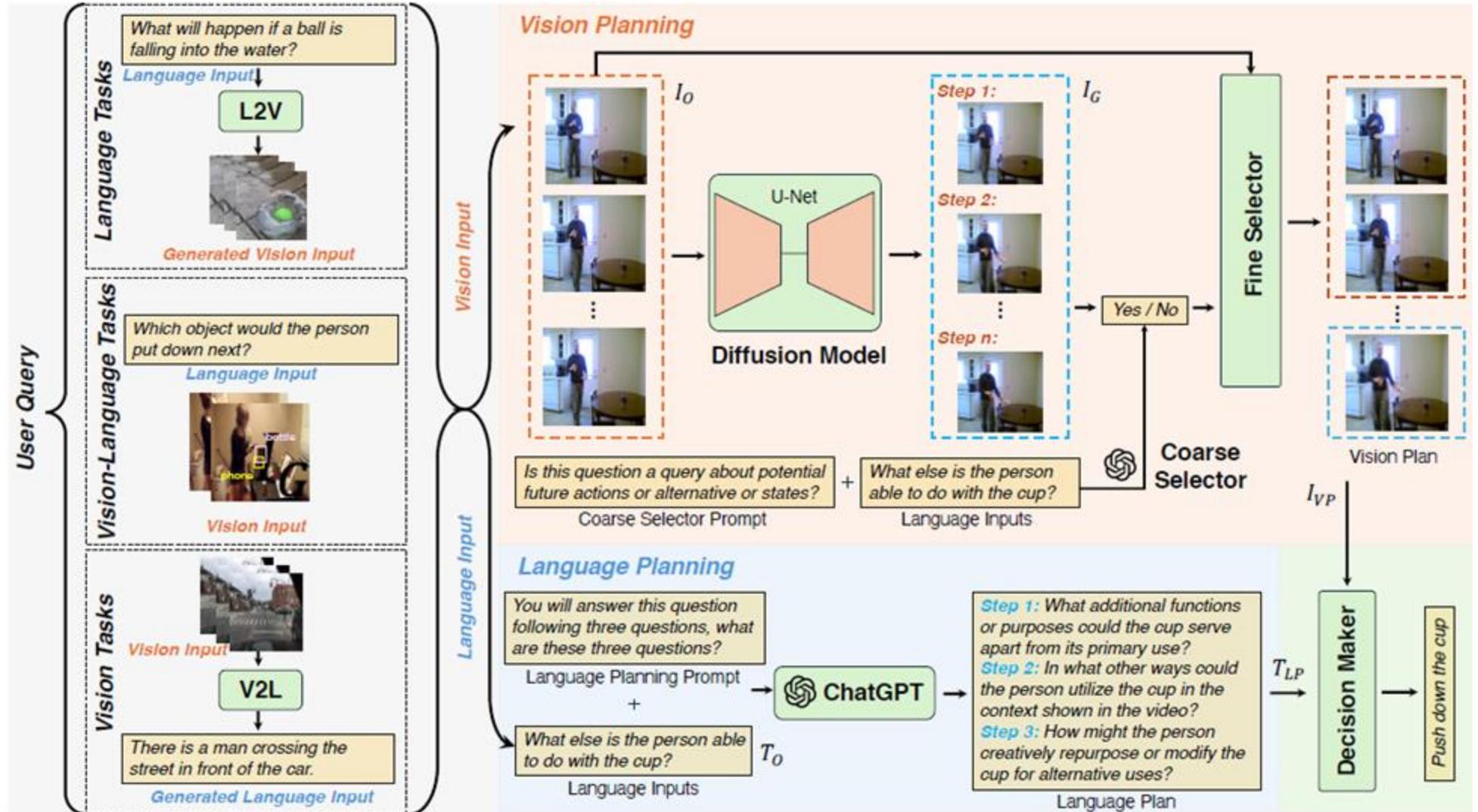
# VLPフレームワークの全体像

## 主要構成要素

- 言語計画 (Language Planning)
- 視覚計画 (Vision Planning)
- 決定メーカー (Decision Maker)

## 処理フロー

- 入力：テキスト、画像、ビデオ
- 言語計画：タスクを小ステップに分解
- 視覚計画：未来フレームを生成
- 決定メーカー：両計画を統合し最終決定



# 言語計画 (Language Planning)

## 1. 概要

- a. ・タスクを小さなステップに分解
- b. ・順序立てた推論プロセスの実現

## 2. 役割

- a. ・複雑な問題の構造化
- b. ・より論理的な思考過程の模倣

## 3. 実装方法

- a. ・ゼロショット Chain-of-Thought 技術の利用
- b. ・例：ChatGPTによる3ステップの質問生成

## 4. 具体例 質問：「その人はカップで他に何ができますか？」

### a. 生成された計画：

- i. ・ステップ1：カップの追加機能は？
- ii. ・ステップ2：現状でのカップの他の使用法は？
- iii. ・ステップ3：カップの創造的な再利用方法は？

# 視覚計画 (Vision Planning)

## 1. 概要

- a. 現在の画像/動画から未来のフレームを生成
- b. 視覚的な推論能力の向上

## 2. 主要構成要素

- a. Vision Planning Generator (VPG): Stable Video Diffusion モデルを使用
- b. Vision Planning Selector (VPS)
  - i. Coarse Selector (CS): ChatGPT (CS)
  - ii. Fine Selector (FS): BLIP-2 (FS) (視覚言語処理タスクのためのモデル)

## 3. 利点

- a. 未来の状態を予測し、より適切な判断が可能
- b. 視覚的な文脈を考慮した推論

## 4. 課題

- a. 生成された未来フレームの品質が重要
- b. 計算コストと推論時間の管理

# 決定メーカー (Decision Maker)

## 決定メーカーの役割

- 言語計画と視覚計画の結果を統合
- 最終的な出力を生成

## 決定メーカーの利点

- 複数の視点からの情報統合
- より堅牢な意思決定プロセス

## LLAVAを用いた実装

- 多段階の対話戦略
  - バニラ回答
  - 言語計画に基づく回答
  - 視覚計画に基づく回答
  - 投票メカニズム

## GPT4-Vを用いた実装

- 単一ステップでの統合
- 言語と視覚情報の同時処理

**GPT4-V Answering**



Original Frames      Generated Frames

Based on the original and generated frames, answer three sub-questions first.

**Step 1:** What is the person doing?  
**Step 2:** What actions has the person performed with the cup so far?  
**Step 3:** What else is the person able to do with the cup?

And then answer the final question:  
**Q:** What else is the person able to do with the cup?

**Answer 1:** The person is standing next to the table, holding a sandwich in his left hand and a glass of orange juice in his right hand.  
**Answer 2:** The cup was held by the person in his hand.  
**Answer 3:** This person can put the cup on the table.  
**Final Answer:** The person can put down the cup.

# 評価実験①：ビデオ質問応答タスク

- 実験概要：
  - データセット：STAR, NExT-QA
  - タスク：ビデオの内容に関する多肢選択式質問に回答

Using Left and Right Brains Together: Towards Vision and Language Planning

Table 1. Results on video question answering.

Model (# Frames)	STAR					NExT-QA			
	Int.	Seq.	Pre.	Fea.	Avg.	Tem.	Cau.	Des.	Avg.
ViperGPT (dense/1fps) (Surís et al., 2023)	-	-	-	-	-	-	-	-	60.0
Flamingo-80B (30) (Alayrac et al., 2022)	-	-	-	-	39.7	-	-	-	-
VFC (32) (Momeni et al., 2023)	-	-	-	-	-	45.4	51.6	64.1	51.5
InternVideo* (8) (Wang et al., 2022)	43.8	43.2	42.3	37.4	41.6	43.4	48.0	65.1	49.1
BLIP-2 <sup>voting</sup> (4) (Li et al., 2023)	41.8	39.7	40.2	39.5	40.3	59.1	61.3	74.9	62.7
BLIP-2 <sup>concat</sup> (4) (Li et al., 2023)	45.5	41.8	41.8	40.0	42.2	59.7	60.8	73.8	62.4
SEVILA (4) (Yu et al., 2023)	48.3	45.0	44.4	40.8	44.6	<b>61.3</b>	<u>61.5</u>	<u>75.6</u>	<u>63.6</u>
LLAVA (4) (Liu et al., 2023)	<u>49.0</u>	<u>47.3</u>	<u>45.5</u>	<u>47.8</u>	<u>47.4</u>	55.7	60.6	74.3	61.1
VLP (4 + 1 (Generated Frame))	<b>52.0</b>	<b>50.1</b>	<b>50.8</b>	<b>49.0</b>	<b>50.5</b>	<u>60.5</u>	<b>63.7</b>	<b>76.7</b>	<b>64.7</b>

主な結果：

- VLPは既存の最先端手法を上回る性能を達成
- STARデータセットで平均50.5%の正解率（従来比+3.1%）
- NExT-QAデータセットで平均64.7%の正解率（従来比+1.1%）

VLPの優位性：

- 言語計画による論理的推論の向上
- 視覚計画による未来予測能力の獲得
- 複雑な質問に対する理解力の向上

# ビデオキャプション生成タスクの評価結果

- データセット: BDD-X (自動運転シナリオ)

- 評価指標:

- BLEU-4 (B)
- CIDEr (C)
- METEOR (M)

- VLPの性能:

- BLEU-4: 35.7 (最高スコア)
- CIDEr: 256.7 (最高スコア)
- METEOR: 31.1 (最高スコア)

- 考察:

- VLPは全ての指標で最高性能を達成
- 特にCIDErスコアで大幅な改善 (+9.2ポイント)
- ビジョンプランニングがキャプション生成の質を向上

Table 2. Results of Video Captioning on the BDD-X dataset. 'B', 'C', and 'M' refer to BLEU-4, CIDEr, and METEOR, respectively.

Method	B	C	M
S2VT (Venugopalan et al., 2015)	30.2	179.8	27.5
S2VT++ (Venugopalan et al., 2015)	27.1	157.0	26.4
SAA (Kim et al., 2018)	31.8	214.8	29.1
WAA (Kim et al., 2018)	32.3	215.8	29.2
ADAPT (Jin et al., 2023)	34.6	247.5	30.6
VLP (Ours)	<b>35.7</b>	<b>256.7</b>	<b>31.1</b>

# 評価実験③：自動運転タスク

- 評価指標：
  - 進路予測：RMSE(度)↓、精度( $A_{\sigma}$ )↑
  - 速度予測：RMSE(m/s)↓、精度( $A_{\sigma}$ )↑
- 主な結果：
  - VPを追加したADAPTモデルが最高性能
  - 進路予測：RMSE 6.2°、 $A_{0.5}$  86.2%
  - 速度予測：RMSE 2.3m/s、 $A_{0.5}$  35.3%
- 考察：
  - 生成された未来フレームが予測精度向上に寄与
  - 特に速度予測で大きな改善 (RMSE 2.5→2.3、 $A_{0.5}$  28.1%→35.3%)

Table 7. Control Signals Prediction Accuracy on BDD-X dataset.

Method	Course						Speed					
	RMSE(degree)↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	$A_{10.0}$ ↑	RMSE(m/s)↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	$A_{10.0}$ ↑
Single	6.3	8.3	84.7	<b>90.5</b>	97.2	98.7	3.4	5.0	25.5	37.8	86.8	98.7
ADAPT	6.4	62.2	85.5	89.9	97.2	<b>98.8</b>	2.5	11.1	28.1	45.3	94.3	99.5
ADAPT + VP	<b>6.2</b>	<b>65.5</b>	<b>86.2</b>	90.3	<b>97.3</b>	<b>98.8</b>	<b>2.3</b>	<b>16.1</b>	<b>35.3</b>	<b>51.8</b>	<b>95.2</b>	<b>99.6</b>

# 事例研究：GPT-4V との組み合わせ（ビジョン関連タスク）

- GPT-4Vを用いたVLPの効果
  - タスク：車の次の行動予測
- バニラ回答vs言語計画vs視覚言語計画の比較
  - バニラ：確実な予測は不可能と回答
  - 言語計画：より詳細な分析を提供するが、依然として予測困難
  - 視覚言語計画：正確な予測を実現
- 視覚計画の効果
  - 生成された未来フレーム：歩行者が道路を横断中
  - GPT-4Vの回答：
    - 車は停止を維持
    - 歩行者が安全に横断し、信号が青に変わった後に進行
- 結論
  - VLPは不確実な状況下でより確実な予測を可能に
  - 視覚と言語の統合が判断精度を向上



**Vanilla Answering**

**[Original Frames]** The images are a sequence of frames from a video. Based on this video, how will the car drive next?

It's *impossible to predict with certainty the car's next move* without more context or additional frames from the video. Here are some possible results.

1. If the traffic light visible in the images is green for the car's direction, and the pedestrian is not in the car's path, the car may proceed through the intersection.
2. If the traffic light is red, or if the pedestrian's path intersects with that of the car, the car should come to a stop.
3. If the car is making a turn and the pedestrian is clear of its path, the car may turn once it's safe to do so, yielding to any other vehicles with the right of way.

Without additional frames showing the traffic lights in more detail or the actions of the pedestrian and other vehicles, it's not possible to provide a definitive answer.

**Language Planning**

**[Original Frames]**

1. What is the Current Context of the Scene? 2. What are the Visible Cues or Indicators from the Car? 3. What are the Likely Actions Based on Road Rules and Driver Behavior? And then provide answers to the final question: How will the car in the video drive next?

**1. Current Context of the Scene:**  
It appears to be an intersection with multiple lanes and traffic signals. There are pedestrian crosswalks, and the scene looks like a typical urban street during daylight hours. Pedestrians are present, suggesting that pedestrian activity is a factor in this context.

**2. Visible Cues or Indicators from the Car:**  
From the reflections and shadows on the car's hood, it seems like the car is stationary as there is no blur or distortion that would indicate movement. The positioning of the car at the stop line suggests it is waiting for a traffic signal or allowing pedestrians to cross.

**3. Likely Actions Based on Road Rules and Driver Behavior:**  
Given the context, if the traffic lights are red, the driver should remain stopped. If the light is green and pedestrians are not in the crosswalk, the driver would likely proceed with caution, following road rules and signals.

For the final question, "How will the car in the video drive next?"—without movement indicators or knowing the current state of the traffic signals, it's *difficult to predict the car's next action definitively*. However, if we assume the traffic light turns green, and there are no pedestrians crossing, the car would likely drive straight ahead or turn if it's in a turning lane, following standard driving procedures. If the light remains red or pedestrians are crossing, the car would stay stationary until it is safe and legal to proceed.

**Vision-Language Planning**

**[Original Frames] + [Generated Frames]**

These images are a sequence of frames from a video. The first two images indicate what is happening, while the last two images are generated for future events. 1. What is the Current Context of the Scene? 2. What are the Visible Cues or Indicators from the Car? 3. What are the Likely Actions Based on Road Rules and Driver Behavior? And then provide answers to the final question: How will the car in the video drive next?

**1. Current Context of the Scene:**  
The scene appears to be an *intersection* in an urban area with a traffic light system. There is a *pedestrian crossing the street* in a crosswalk. **The traffic light visible in the images is red**, suggesting that vehicles in the direction of the viewer must stop.

**2. Visible Cues or Indicators from the Car:**  
The images show a view from inside a car, likely from a dashboard camera. The hood of the car and a part of the windshield are visible, and they are not moving relative to the crosswalk lines, which indicates **the car is stationary**.

**3. Likely Actions Based on Road Rules and Driver Behavior:**  
Since **the traffic light is red**, the car should remain stationary. The driver is expected to **wait for the pedestrian** to finish crossing safely. The driver should wait for the traffic light to turn green before proceeding.

Based on this information and assuming the driver follows standard road rules and safe driving practices, the car in the video will likely **remain stationary until the traffic light turns green**. Once **the pedestrian has safely crossed and the light turns green**, the driver will likely proceed through the intersection.

# 事例研究：GPT-4V との組み合わせ（言語のみのタスク）

## GPT-4Vを用いた言語のみのタスクでのVLP活用例

- タスク：ニューヨークの花火ショーに関するニュース記事作成
- VLPの効果：
  - 言語計画による構造化
  - 生成された視覚情報による詳細な描写
- 主なポイント：
  - より具体的な場所の言及（例：エンパイアステートビル）
  - 花火の色彩豊かな描写
  - 臨場感のある表現の増加
- 結論：VLPにより、より詳細で魅力的な記事が生成可能に

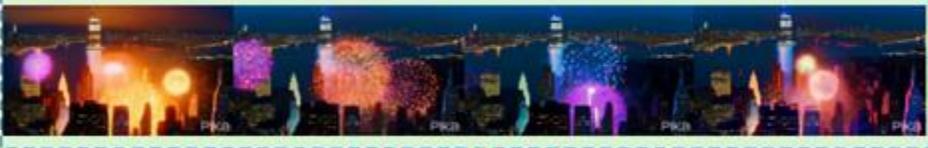
**Vanilla Answering**

You are a journalist writing a news report about the New York's firework show, which illuminate the skyline, evoking a dreamy vibe. Please write the news.

**New York City's Skyline Illuminated by Dazzling Firework Display**  
New York, [Date] – The night sky over New York City was transformed into a dazzling display of color and light as the annual firework show took place, captivating thousands of residents and tourists alike. The event, which is a highlight of the city's cultural calendar, **saw the skyline near the [specific location, e.g., East River, Hudson River, etc.] come alive with an array of spectacular fireworks.** The display started at [time] and lasted for approximately [duration], during which the sky was lit up with an intricate dance of lights, creating a dreamy and almost surreal atmosphere. **Spectators gathered** in large numbers at various vantage points throughout the city, including parks, rooftops, and bridges. Social distancing measures were observed where possible, as the city continues to navigate the post-pandemic era. **The firework show, choreographed by [Name of the Company or Choreographer], featured a series of thematic sequences,** each designed to evoke different emotions and celebrate the city's diverse culture and spirit. **As the final fireworks faded** into the night, the city was left with a sense of awe and joy. Social media platforms were abuzz with photographs and videos, as people shared their experiences of the magical evening.

**Vision-Language Planning**

**[Language Prompt]**  
In the futuristic cityscape of NYC skyline at night, fireworks illuminate the sky with epic realism, captured in highly detailed 8K resolution using a long shot technique, evoking a dreamy vibe.

→ Pika → 

**[Generated Frames]**

You are a journalist writing a news report about the New York's firework show, which illuminate the skyline, evoking a dreamy vibe. Please first describe the outline of the news and then write the news with reference of the pictures.

**Outline for the News Report on New York's Firework Show:**  
1. Introduction. 2. Visual Description. 3. Public Engagement. 4. Technical Aspects. 5. Cultural Impact. 6. Closing Thoughts.

**News Report:**  
**New York's Night Sky Transforms into a Dreamy Palette of Firework Fantasia**  
In the heart of New York City, the night of [Date] unfurled a spectacular show of pyrotechnics that painted the urban skyline with a myriad of colors, evoking a dreamy vibe that enchanted thousands of onlookers. The annual fireworks show, a testament to the city's love for grandeur, illuminated the heavens above **Manhattan**, creating a vibrant backdrop to the iconic silhouettes of the **Empire State Building** and the sprawling bridges that stitch the city's boroughs together. **As twilight melted** into the darker shades of evening, **the first burst of fireworks erupted in a symphony of red and gold near the water's edge,** reflecting on the river's surface like flickering flames. The display escalated into an **orchestra of purple, orange, and blue blooms,** each outshining the last, while the city's towering skyscrapers watched on, bathed in a neon glow. Photographs of the event showcase a dazzling array of fireworks that seemed to compete with the stars, with bursts of **vibrant hues ranging from magenta to fiery oranges,** against the canvas of the night sky. **The grand finale was a masterstroke, a crescendo of light that seemed to set the entire city aflame,** with a dense constellation of sparks leaving spectators in awe.

Figure 4. VLP with GPT4-V on language-only tasks. GPT4-V gives more detailed and vivid descriptions using the generated videos.

# アブレーション研究

Table 3: STARデータセットでのVPとLPのアブレーション研究

- 主な知見：
  - ビジョンプランニング（VP）と言語プランニング（LP）の効果
    - ベースラインと比較して性能が向上
    - VP: 平均2.2%の精度向上
    - LP: 平均3.0%の精度向上
- タスクによる効果の違い
  - ビジョン・言語タスク（STAR）：  
LPがより効果的
  - 純粋なビジョンタスク（BDD-X）：  
VPがより効果的
- 組み合わせの効果
  - VP+LPの組み合わせが最も高い性能を示す

結論：

- VPとLPは単独でも効果があるが、組み合わせることですらなる性能向上が可能。タスクの特性に応じて、VPとLPの重要性が変化する

Table 3. Ablation study of VP and LP on STAR dataset.

Model	Int.	Seq.	Pre.	Fea.	Avg.
LLAVA (Liu et al., 2023)	49.0	47.3	45.5	47.8	47.4
LLAVA+VP (Ours)	51.5	<u>49.9</u>	50.0	47.1	49.6
LLAVA+LP (Ours)	<b>52.3</b>	<b>50.1</b>	<b>51.1</b>	<u>48.2</u>	<u>50.4</u>
LLAVA+LP+VP (Ours)	<u>52.0</u>	<b>50.1</b>	<u>50.8</u>	<b>49.0</b>	<b>50.5</b>

質問タイプ

Int.: Interaction (相互作用)

Seq.: Sequence (順序)

Pre.: Prediction (予測)

Fea.: Feasibility (実現可能性)

Avg.: Average (平均)

# ビジョンプランニングの成功例と失敗例

- 成功例：
  - 原画像のみでは「キャビネットを開ける」と予測
  - 生成された未来フレームにより「紙/ノートを取る」と正しく予測
- 失敗例：
  - 原画像で正しく「箱を置く」と予測
  - 生成された未来フレームにより誤って「カップ/グラス/ボトルを取る」と予測
- 考察：
  - ビジョンプランニングは未来の状態を予測し、より正確な判断を可能にする
  - しかし、生成された未来フレームの品質が結果に大きく影響する
- 重要ポイント：
  - ビデオ生成モデルの精度向上が課題
  - 生成フレームの適切な選択が重要

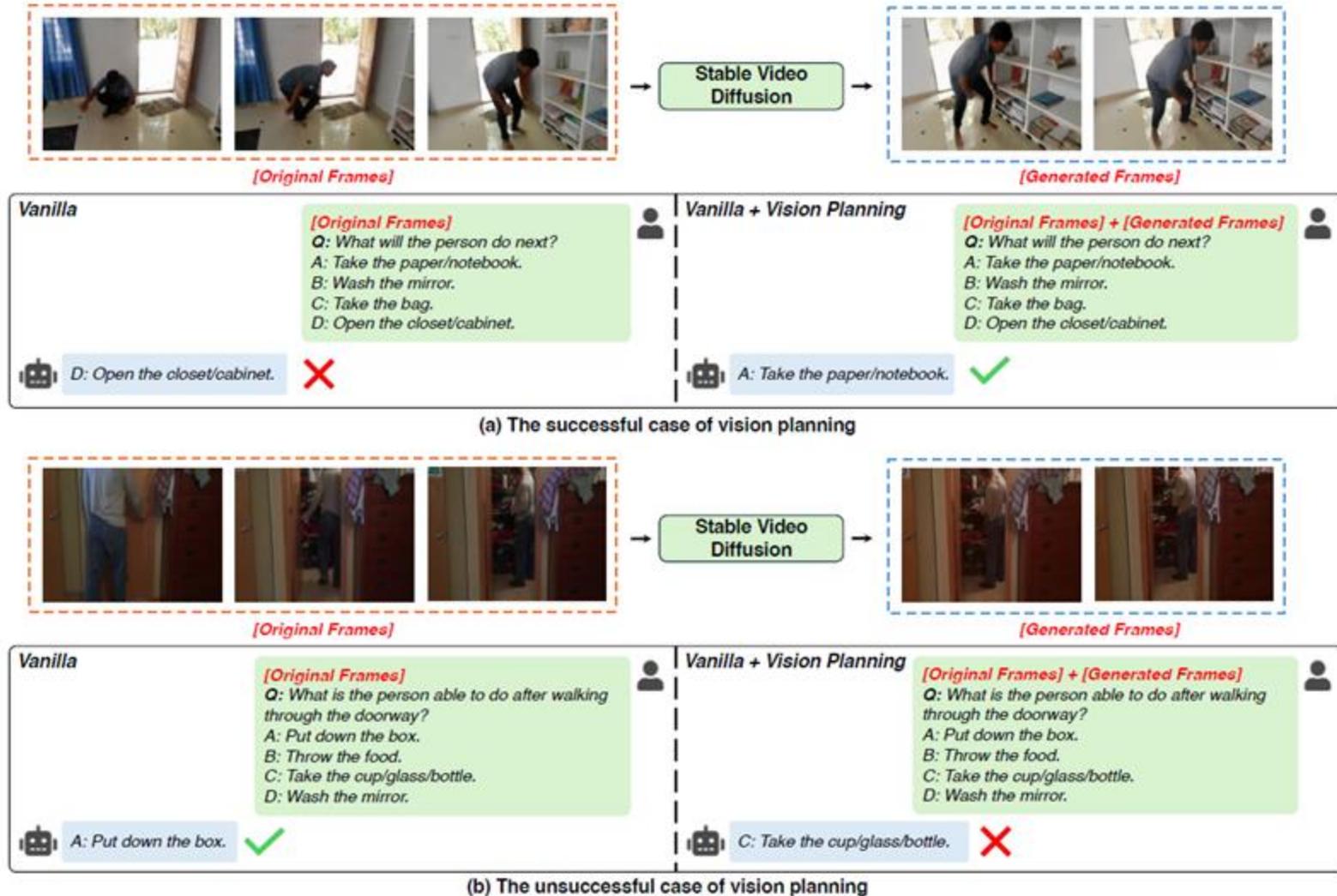


Figure 8. The successful and unsuccessful cases of vision planning.

# 言語プランニングの成功例と失敗例

成功例：

- 質問：人が本を取る前に置いた物は何か？
- バニラの回答：毛布（不正解）
- 言語プランニング後の回答：サンドイッチ（正解）

成功の理由：

- 段階的な質問により、詳細な状況把握
- 物の順序を正確に認識（皿を置いてから本を取る）

失敗例：

- 質問：人はサンドイッチをどうしたか？
- バニラの回答：食べた（正解）
- 言語プランニング後の回答：置いた（不正解）

失敗の原因：

- サブ質問への不適切な回答（予期せずに受け取った、投げ捨てた）
- 誤った情報に基づく最終判断

教訓：

- 言語プランニングの質が結果を左右
- サブ質問の設計と回答の正確性が重要

**Vanilla + Language Planning**

**[Original Frames]**  
Step 1: What objects did the person interact with during the video?  
Step 2: In what order did the person interact with the objects?  
Step 3: Which object did the person put down before they took the book?

In the video, the person interacted with a couch, a plate of food, and a book. They first sat on the couch, then picked up the plate of food, and finally put the plate down before taking the book.

**Vanilla**

**[Original Frames]**  
Q: Which object did the person put down before they took the book?  
A: The picture.  
B: The clothes.  
C: The sandwich.  
D: The blanket.

D: The blanket. ❌

**Vanilla + Language Planning**

**[Original Frames]**  
Q: Based on the above sub questions and corresponding answers, which object did the person put down before they took the book?  
A: The picture.  
B: The clothes.  
C: The sandwich.  
D: The blanket.

C: The sandwich. ✅

(a) The successful case of language planning

**Vanilla + Language Planning**

**[Original Frames]**  
Step 1: What was the person's initial reaction or expression when they received the sandwich?  
Step 2: What did the person do with the sandwich immediately after receiving it?  
Step 3: What was the final outcome or disposition of the sandwich?

1. The person's initial reaction or expression when they received the sandwich was surprise or confusion, as they were not expecting it.  
2. The person immediately started eating the sandwich, taking a bite out of it.  
3. The sandwich was eventually thrown away, indicating that it was not consumed or enjoyed by the person.

**Vanilla**

**[Original Frames]**  
Q: What did the person do with the sandwich?  
A: Tidied up.  
B: Ate.  
C: Put down.  
D: Took.

B: Ate. ✅

**Vanilla + Language Planning**

**[Original Frames]**  
Q: Based on the above sub questions and corresponding answers, What did the person do with the sandwich?  
A: Tidied up.  
B: Ate.  
C: Put down.  
D: Took.

C: Put down. ❌

(b) The unsuccessful case of language planning

# ビデオ生成モデルの品質の重要性

- VLPの性能はビデオ生成モデルの品質に大きく依存
- 異なるビデオ生成モデルの比較結果：

- 主な知見：
  - グラウンドトゥールースフレームが最高性能
  - ドメイン特化型モデル（DMVFN）が汎用モデル（Stable Video Diffusion）より優れる
  - 低解像度生成（MCVD）は性能が劣る

ビデオキャプション生成の評価指標

- B (BLEU-4)
- C (CIDEr)
- M (METEOR)

- 課題：
  - オープンドメインでの高品質ビデオ生成
  - タスク特化型ビデオ生成モデルの開発
- 将来展望：
  - ビデオ生成技術の進歩がVLPの性能向上に直結

Table 10. Ablation Study of video generation model on BDD-X.

Video Generation Method	B	C	M
MCVD - Cityscapes	31.2	195.3	26.8
DMVFN - Cityscapes	35.0	230.1	<b>29.4</b>
DMVFN - Kitti	<b>35.2</b>	<b>234.2</b>	<b>29.4</b>
Stable Video Diffusion	33.9	229.6	28.8
Ground Truth Frames	34.6	247.5	30.6

# 考察と今後の課題

## 1. ビデオ生成モデルの品質改善

- 現状の課題：
  - 生成された未来フレームの品質が不十分
  - 不自然な動きや不明瞭な詳細が存在
- 改善の方向性：
  - より高解像度で自然な動きを生成できるモデルの開発
  - タスク特化型のファインチューニング

## 2. 推論時間の短縮

- 現在のボトルネック：
  - ビデオ生成に60秒以上かかる場合も
- 短縮への取り組み：
  - 拡散ステップの削減（例：50ステップから4ステップへ）
  - 推論時間を10秒程度まで短縮する技術の開発

## 3. その他の課題

- マルチモーダルモデルの更なる統合
- リアルタイム処理への対応
- 多様なドメインへの適用と検証

# まとめと展望

- Vision-Language Planning (VLP) フレームワークの提案
  - 言語計画と視覚計画の統合
  - 人間の脳の左右半球の機能にヒントを得た設計
- 主な成果:
  - ビデオ質問応答、キャプション生成での性能向上
  - 自動運転、ロボット操作タスクでの有効性
  - GPT-4Vとの組み合わせによる効果的な運用
- 今後の展望:
  1. ビデオ生成モデルの品質向上
  2. 推論時間の短縮
  3. より複雑なマルチモーダルタスクへの応用
  4. 実世界のAIシステムへの統合
- VLPの可能性:
  - より人間らしい認知プロセスを持つAIの実現
  - マルチモーダルAIシステムの能力向上

# 山川の考察：VLPフレームワークの脳器官への対応付け

## 入力処理:

- ・ 視覚入力: 後頭葉 (視覚野)
- ・ 言語入力: 側頭葉 (ウェルニッケ野)

## モダリティ変換:

- ・ 頭頂連合野: 異なる感覚モダリティの情報を統合する役割

## 言語計画 (Language Planning):

- ・ 前頭葉 (特に左半球): 言語処理、計画立案
- ・ ブローカ野: 言語生成
- ・ 背外側前頭前野: 実行機能、計画立案

## 視覚計画 (Vision Planning):

- ・ 右半球の頭頂葉: 空間認識、視覚的イメージの操作
- ・ 後頭-頭頂経路: 視覚情報の処理と空間的操作

## 決定メーカー (Decision Maker):

- ・ 前頭前野: 高次の意思決定、統合
- ・ 前帯状皮質: 意思決定、感情処理
- ・ **海馬: 記憶の統合と新しい情報の処理**

## 出力生成:

- ・ 運動野: 行動の計画と実行 (ロボットタスクの場合)
- ・ ブローカ野: 言語出力の生成

本結果は、  
Claude3の知識に基づく簡易的なものであることに注意

# 私達の関連研究：海馬を参照した情報統合

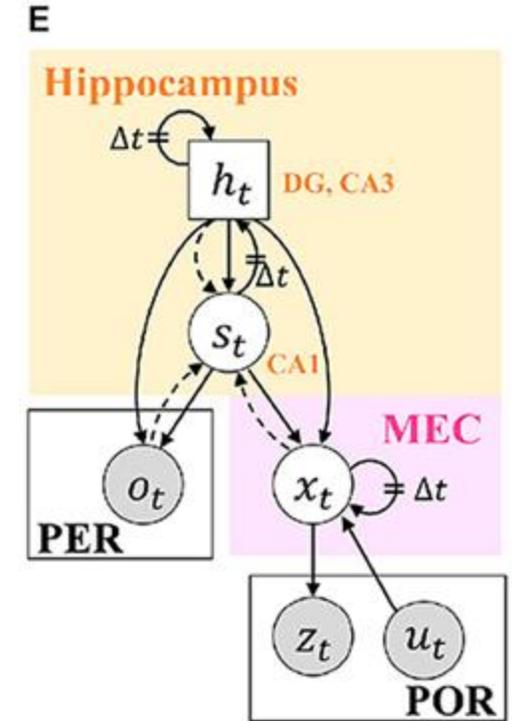
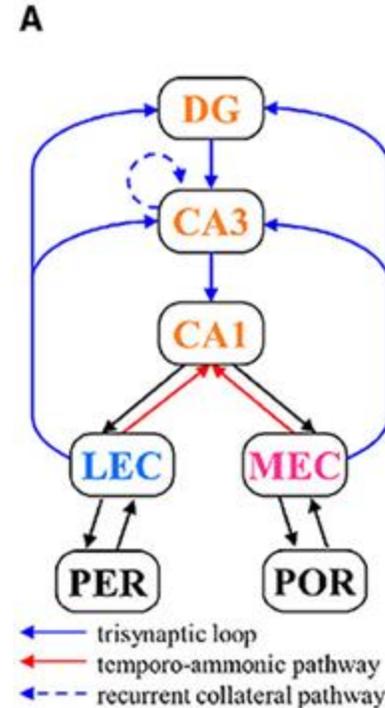
目的:

- 海馬形成の構造を参考にしたロバストな自己位置推定モデルの開発
- Allocentric情報(MEC)とEgocentric情報(LEC)の統合

主要な結果と考察 (左下):

- CA3に対応する潜在変数がスパースな表現を獲得
- ロボットが突然テレポートされる「誘拐ロボット問題」において、高い性能

Nakashima, T., Otake, S., Taniguchi, A., Maeyama, K., El Hafi, L., Taniguchi, T., & Yamakawa, H. (2024). Hippocampal formation-inspired global self-localization: quick recovery from the kidnapped robot problem from an egocentric perspective. *Frontiers in Computational Neuroscience*, 18.





# [3] Brain-Inspired Intelligence

## 研究ミッション

脳参照アーキテクチャ(BRA)をベースにしたモデル開発や分析を通じ、脳の仕組みを解き明かす

### テーマ① 全脳BRA 構築

神経科学知見に基づき、脳全体にわたる  
**BRA**データの構築と評価を行う。  
他方で部分的に計算機能の仮説を**HCD/FRG**  
として構築し、計算モデルを実装する。

#### キーワード例

- 脳参照アーキテクチャ(BRA)
- 脳情報フロー (BIF)
- 仮説的コンポーネント図 (HCD)
- 機能実現グラフ(FRG)
- Structure-constrained Interface Decomposition (SCID) 法

### テーマ② ヒト脳型 AGI

**BRA**を活用し、脳モデルの実装や、脳データの  
解析を行う。モデルを用いた機能不全の分  
析（医療）や脳のように機能や状態（意図・  
欺瞞など）を解釈できる可能性を探求する。

#### キーワード例

- AIアライメント  
(脳に基づく解釈可能性等を含む)
- 脳シミュレーションと脳の分析
- 脳の機能不全のモデル化
- 人と親和性の高いコミュニケーション

### テーマ③ データ 自動生成

**BRA**データの作成/評価に関わるパイプライン  
を構築し、**LLM**を用いて自動化する。  
**2027**年までに最初の全脳**BRA**を構築し、  
その後は自動更新できる状況を目指す。

#### キーワード例

- BRAデータの自動評価と自動作成
- WBA技術ロードマップ
- 大規模言語モデル
- BRA Editorial System (BRAES)
- Bibliographic database for BRA (BDBRA)